Helsinki University of Technology
Department of Biomedical Engineering and Computational Science Publications
Teknillisen korkeakoulun Lääketieteellisen tekniikan ja laskennallisen tieteen laitoksen julkaisuja
2010 REPORT A23

SPEEDING UP THE INFERENCE IN GAUSSIAN PROCESS MODELS

Doctoral Dissertation

Jarno Vanhatalo

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium E at the Aalto University School of Science and Technology (Espoo, Finland) on the 19th of October, 2010, at 12 noon.

Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Biomedical Engineering and Computational Science

Aalto-yliopiston teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Distribution:
Aalto University
School of Science and Technology
Faculty of Information and Natural Sciences
Department of Biomedical Engineering and Computational Science
P.O.Box 12200
FI-00076 Aalto
FINLAND

Tel. +358 9 470 23172
Fax. +358 9 470 23182

ISBN 978-952-60-3380-8 (print)
ISBN 978-952-60-3381-5 (electronic)
ISSN 1797-3996
URL http://lib.tkk.fi/Diss/2010/isbn9789526033815

Picaset Oy
Helsinki 2010

Aalto University

ABSTRACT OF DOCTORAL DISSERTATION	AALTO UNIVERSITY SCHOOL OF SCIENCE AND TECHNOLOGY P.O. BOX 11000, FI-00076 AALTO http://www.aalto.fi

Author Jarno Vanhatalo

Name of the dissertation

Speeding up the Inference in Gaussian Process Models

Manuscript submitted 15.6.2010	Manuscript revised 9.9.2010

Date of the defence 19.10.2010

☐ Monograph ☒ Article dissertation (summary + original articles)

Faculty	Faculty of Information and Natural Sciences
Department	Department of Biomedical Engineering and Computational Science
Field of research	Computational engineering
Opponent(s)	Prof. Neil Lawrence
Supervisor	Prof. Jouko Lampinen
Instructor	Dr.Tech. Aki Vehtari

Abstract

In this dissertation Gaussian processes are used to define prior distributions over latent functions in hierarchical Bayesian models. Gaussian process is a non-parametric model with which one does not need to fix the functional form of the latent function, but its properties can be defined implicitly. These implicit statements are encoded in the mean and covariance function, which determine, for example, the smoothness and variability of the function. This non-parametric nature of the Gaussian process gives rise to a flexible and diverse class of probabilistic models.

There are two main challenges with using Gaussian processes. Their main complication is the computational time which increases rapidly as a function of a number of data points. Other challenge is the analytically intractable inference, which exacerbates the slow computational time. This dissertation considers methods to alleviate these problems.

The inference problem is attacked with approximative methods. The Laplace approximation and expectation propagation algorithm are utilized to give Gaussian approximation to the conditional posterior distribution of the latent function given the hyperparameters. The integration over hyperparameters is performed using a Monte Carlo, a grid based, or a central composite design integration. Markov chain Monte Carlo methods over all unknown parameters are used as a golden standard to which the other methods are compared. The rapidly increasing computational time is cured with sparse approximations to Gaussian process and compactly supported covariance functions. These are both analyzed in detail and tested in experiments. Practical details on their implementation with the approximative inference techniques are discussed.

The techniques for speeding up the inference are tested in three modeling problems. The problems considered are disease mapping, regression and classification. The disease mapping and regression problems are tackled with standard and robust observation models. The results show that the techniques presented speed up the inference considerably without compromising the accuracy severely.

Keywords sparse Gaussian process, approximate inference, compactly supported covariance function

ISBN (printed) 978-952-60-3380-8	ISSN (printed) 1797-3996
ISBN (pdf) 978-952-60-3381-5	ISSN (pdf)
Language English	Number of pages 43 + 83 app.

Publisher Department of Biomedical Engineering and Computational Science, Aalto University

Print distribution Aalto University, Department of Biomedical Engineering and Computational Science

☒ The dissertation can be read at http://lib.tkk.fi/Diss/2010/isbn9789526033815/

Aalto-yliopisto

VÄITÖSKIRJAN TIIVISTELMÄ	AALTO-YLIOPISTO TEKNILLINEN KORKEAKOULU PL 11000, 00076 AALTO http://www.aalto.fi

Tekijä Jarno Vanhatalo

Väitöskirjan nimi

Gaussisia prosesseja hyödyntävien mallien analyysin nopeuttaminen

Käsikirjoituksen päivämäärä 15.6.2010	Korjatun käsikirjoituksen päivämäärä 9.9.2010
Väitöstilaisuuden ajankohta 19.10.2010	
☐ Monografia	☒ Yhdistelmäväitöskirja (yhteenveto + erillisartikkelit)
Tiedekunta	Informaatio- ja luonnontieteiden tiedekunta
Laitos	Lääketieteellisen tekniikan ja laskennallisen tieteen laitos
Tutkimusala	Laskennallinen tekniikka
Vastaväittäjä(t)	Prof. Neil Lawrence
Työn valvoja	Prof. Jouko Lampinen
Työn ohjaaja	TkT Aki Vehtari

Tiivistelmä

Tässä väitöskirjassa gaussisia prosesseja käytetään määrittämään priorijakaumia latenteille funktioille hierarkisissa bayesilaisissa malleissa. Gaussinen prosessi on epäparametrinen malli, jossa latentin funktion muotoa ei tarvitse kiinnittää vaan sen ominaisuuksia voidaan määritellä epäsuorasti. Ominaisuudet määritellään keskiarvo- ja kovarianssifunktioiden avulla, jotka vaikuttavat muun muassa funktion sileyteen ja vaihtelevuuteen. Gaussisen prosessin epäparametrisuus mahdollistaa joustavien ja monipuolisten todennäköisyyspohjaisten mallien rakentamisen.

Gaussisten prosessien käytössä on kaksi merkittävää ongelmaa. Pääasiallinen ongelma on laskentaan vaadittava aika, joka kasvaa nopeasti aineiston määrän kasvaessa. Toinen ongelma on, ettei malleja pystytä analysoimaan analyyttisesti, mikä hidastaa laskentaa entisestään. Tässä väitöskirjassa tutkitaan menetelmiä näiden ongelmien pienentämiseksi.

Mallien analysoinnin ongelmia ratkotaan aproksimatiivisilla menetelmillä. Laplace aproksimaatiota ja expectation propagation algoritmia käytetään antamaan gaussinen approksimaatio latenttien muuttujien ehdolliselle posteriorille annettuna hyperparametrit. Hyperparametrien yli integrointi suoritetaan Monte Carlo, ruudukko, tai central composite design integrointimenetelmillä. Menetelmiä verrataan Markov ketju Monte Carlo integrointiin kaikkien tuntemattomien muuttujien yli. Laskentaa nopeutetaan gaussisen prosessin harvoilla aproksimaatioilla ja kompaktikantajaisilla kovarianssifunktioilla. Näitä molempia analysoidaan yksityiskohtaisesti ja testataan koeaineistoilla. Myös käytännön yksityikohtia niiden yhdistämisestä aproksimatiivisten analyysimenetelmien kanssa käsitellään.

Analyysin nopeutusmenetelmiä testataan kolmella mallinnusongelmalla: tautikartoitus-, regressio- ja luokitteluongelmalla. Tautikartoitus- ja regressio-ongelmaa lähestytään sekä perinteisillä että robusteilla havaintomalleilla. Tulokset osoittavat, että käsitellyt menetelmät nopeuttavat analyysia huomattavasti huonontamatta tulosten tarkkuutta merkittävästi.

Asiasanat	harva gaussinen prosessi, approksimatiivinen päättely, kompaktikantajainen kovarianssifunktio		
ISBN (painettu)	978-952-60-3380-8	ISSN (painettu)	1797-3996
ISBN (pdf)	978-952-60-3381-5	ISSN (pdf)	
Kieli	Englanti	Sivumäärä	43 + liitteet 83
Julkaisija	Lääketieteellisen tekniikan ja laskennallisen tieteen laitos, Aalto-yliopisto		
Painetun väitöskirjan jakelu	Aalto-yliopisto, Lääketieteellisen tekniikan ja laskennallisen tieteen laitos		
☒ Luettavissa verkossa osoitteessa http://lib.tkk.fi/Diss/2010/isbn9789526033815/			

Foreword

When I first opened the door to professor Jouko Lampinen's office in 2004, I didn't have the smallest hunch where that summer job interview would lead me to. After two summers of struggling with MCMC methods and Bayesian modeling I found myself doing a masters thesis on Gaussian processes. The research group and instructor, Dr.Tech. Aki Vehtari, had remained the same throughout the journey from Metropolis-Hastings to Reversible jump MCMC - and beyond. And then, in the fall 2006, Aki was steering my first steps as a post graduate student.

During these four years my home base, Laboratory of Computational Engineering at Helsinki University of Technology, has gone through a series of transformations. First it merged with biomedical engineers and became part of Department of Biomedical Engineering and Computational Science, and moved to new premises. Later on, my beloved university decided to merge with two other universities from the nearby seaport putting our new shiny deparment as a part of Aalto University. Despite all these transformations, it has provided me a safe and fertile ground for research.

Before I let you, dear reader, take a closer look at my accomplishments, I beg your pardon and spend little time on acknowledgements. First of all, I want to thank my instructor Aki Vehtari for all his support. Whenever there was a problem I could pop in his office, and during the many bursts of frustration his patience proved to be endless. My supervisor Jouko Lampinen receives thanks for excellent working environment and the chance to work with these fascinating problems. Jouko deserves also special mention for his relaxed attitude and ever present figure in laboratory's leisure activities, as well as his unbeatable solid chest in walrus wrestling.

I want to thank also all my collaborators and workmates. I give special thanks to Pasi Jylänki, Jaakko Riihimäki, Jouni Hartikainen, Janne Ojanen, Ville Pietiläinen and Jussi Kumpula for excellent company and interesting conversations. Thank you also for occasionally harassing me with random matters that gave welcomed rest. From the administrative staff I must mention Kaija Virolainen and Laura Pyysalo, who helped with so many practical matters. From outside my research group, I first thank Pia Mäkelä, from National Institute for Health and Welfare, for guiding me in alcohol related matters from nonconsumer point of view. I thank also Professor Zoubin Ghahramani and Dr. Carl Rasmussen for hosting me in the Computational and Biological Learning Lab in Cambridge in 2008.

I want to acknowledge also the funders of this work. The Graduate School in Electronics and Telecommunications and Automation (GETA) provided my basic funding from almost the beginning of my post graduate studies. Finnish Funding Agency for

Technology and Innovation (TEKES) and the Academy of Finland gave important supplement to the graduate school funding through the Finnwell -project and Aki's grant. I want to thank also my scholarship providers, Satakunta regional Fund of Finnish Cultural Foundation, Emil Aaltonen Foundation and the Finnish Foundation for Economic and Technology Sciences – KAUTE, for their generous attitude towards me.

Usually parents are mentioned at this point, and so here also. Not for just being there (since evidently my parents are a necessary condition for the existence of this thesis) but for one particular decision they made in my youth. I'm gratefull to my parents for finanzing my exchange student year during the high school (even after the lengthy negotiations). The English I learned in Texas at the end of 1990's has paid back well during this work. I must mention also my former girlfriend – and present wife – Sari, even though she explicitly forbade that. It is just a must to repeat her words at this point: "I don't want to be acknowledged for doing nothing."Well, this claim is not totally true. Even though her input to the scientific work was infinitesimal, she stole my thoughts from it every time I needed it most.

Jarno Vanhatalo

List of publications and author's research contributions

This dissertation consists of an overview and the following publications:

I Vanhatalo, J. and Vehtari, A. (2007). Sparse log Gaussian processes via MCMC for spatial epidemiology. *JMLR Workshop and Conference Proceedings*, 1:73–89.

II Vanhatalo, J. and Vehtari, A. (2008). Modelling local and global phenomena with sparse Gaussian processes. In McAllester, D. A. and Myllymäki, P., editors, *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 571–578.

III Vanhatalo, J., Jylänki, P., and Vehtari, A. (2009). Gaussian process regression with Student-t likelihood. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1910–1918.

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

V Vanhatalo J., and Vehtari A. (2010). Speeding up the binary Gaussian process classification. In Grünwald P. and Spirtes P., editors, *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 623-631.

VI Vanhatalo, J., Mäkelä, P., and Vehtari, A. (2010). Regional differences in alcohol mortality in Finland in the early 2000s. Technical Report A20, Aalto University, Department of Biomedical Engineering and Computational Science. (Translation of the original Finnish article: Vanhatalo, J., Mäkelä, P., and Vehtari, A. (2010). Alkoholikuolleisuuden alueelliset erot suomessa 2000-luvun alussa. *Yhteiskuntapolitiikka*, 75(3):265–273.)

In all publications, Vanhatalo had the principal responsibility writing and preparing the manuscript. The co-authors contributed to writing by revising the text and suggesting modifications. Vanhatalo had the main responsibility implementing the models and

running the experiments. Vehtari contributed to the initiation of the performed research and in background considerations.

In publication **I**, Vanhatalo and Vehtari made equal contribution to the development of the computational methods. Vanhatalo implemented the methods and ran the experiments.

In publications **II**, and **V** Vanhatalo contributed the main part in all aspects of the articles.

In publications **III**, and **IV**, Vanhatalo had the main responsibility in other matters than implementing and running experiments with models used in comparisons, which were done by Jylänki (in publication **III**) and Pietiläinen (in publication **IV**). In Publication **IV** Pietiläinen contributed also to implementing some of the computational methods related to Gaussian process models under the supervision of Vanhatalo and Vehtari.

In publication **VI**, Vanhatalo ran the experiments and visualized the results. Mäkelä made the main contribution to analysing the results and discussing them from a socio-political point of view.

Sisältö

Chapter 1

Introduction

Mathematical models are a nonseparable part of almost all scientific disciplines, engineering applications and every day life in modern days. Probably the best known example of mathematical models in science are Newton's laws, which describe deterministically the relationship between the forces acting on a body and the motion of that body. In a deterministic approach, observations, for example motion, can be determined exactly from the parameters, force. Opposed to this, probabilistic models deal with uncertain events, where parameters determine probabilities for a number of possible observations. The early works on probability theory concentrated on a pre-data problem where the parameters are known and the aim is to give probability statements about observations before seeing them. In the late 18th century the mathematical machinery was turned to solve the so called inverse probability, where the aim is to infer the unknown parameters from the observations. Bayes (1763) and Laplace (1774) receive independent credit for being the firsts to consider the problem.

Albeit the Bayesian modeling owes its name to reverend Thomas Bayes, the modern day Bayesian theory has come rather far from the early works in the late 18th century. The foundations were laid in the early 20th century (see e.g. Dale, 1999) and current theoretical constructions are well summarized by Bernardo and Smith (2000). The fundamental principle in the Bayesian theory is that all uncertainties are summarized with probabilities, let them relate to parameters or observations. This differs from the classical point of view, where only the observations may be random variables, but is very much aline with modern day mathematical modeling in general. For example, in physics or in engineering, handling uncertainties, let them relate to observations or unknown parameters, is every day business.

Building a Bayesian model, denoted by $\mathcal{M}$, starts with an *observation model* which contains the description of the data generating process, or its approximation. The observation model is denoted by $p(\mathcal{D}|\theta, \mathcal{M})$, where θ stands for the parameters and $\mathcal{D}$ the observations. The observation model quantifies a conditional probability for data given the parameters, and when regarded as a function of parameters it is called *likelihood*. If the parameter values were known, the observation model would contain all the knowledge of the phenomenon and could be used as such. If the observations contain randomness, sometimes called *noise*, one would still be uncertain of the future

observations, but could not reduce this uncertainty since everything that can be known exactly would be encoded in the observation model. Usually, the parameter values are not known exactly but there is only limited knowledge on their possible values. This prior information is formulated mathematically by the *prior probability* $p(\theta|\mathcal{M})$, which reflects our beliefs and knowledge about the parameter values before observing data. Opposed to the aleatory uncertainty encoded in the observation model the epistemic uncertainty present in the prior can be reduced by gathering more information on the phenomenon (for a more illustrative discussion on the differences between these two sources of uncertainty see O'Hagan, 2004)). Bayesian inference is the process of updating our prior knowledge based on new observations – in other words it is the process for reducing the epistemic uncertainty.

The cornerstone of Bayesian inference is the Bayes' theorem which defines the conditional probability of the parameters after observing the data

$$p(\theta|\mathcal{D},\mathcal{M}) = \frac{p(\mathcal{D}|\theta,\mathcal{M})p(\theta|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}. \tag{1.1}$$

This is called the *posterior distribution*. It contains all the information about parameter θ that we are able to extract from the data $\mathcal{D}$ with the model we are using and combines this with the prior information. The normalization constant

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\theta,\mathcal{M})p(\theta|\mathcal{M})d\theta \tag{1.2}$$

is equal to the conditional probability of the data given our model assumptions. It is also called the marginal likelihood for the model. The model, $\mathcal{M}$, stands for all the hypotheses and assumptions that are made about the phenomenon. It embodies the functional forms of the observation model and the prior, which are always tied together, as well as our subjective assumptions used to define these mathematical abstractions. Because everything is conditioned on $\mathcal{M}$, it is a redundant symbol and as such omitted from this on. Usually we are not able to define 'correct' model and most of the time we have only limited ability to encode our prior beliefs in the mathematical formulation. For this reason one should always keep in mind the hidden dependency on $\mathcal{M}$ and verify that the results are sensible and investigate what aspects of reality are not captured by the model. Checking the model is crucial to distinguish useful models from poor ones (see for example Gelman et al., 2004).

The true power of the Bayesian approach comes from the possibility to construct and analyze hierarchical models. In hierarchical models, prior probabilities are appointed also for the parameters of the prior. Let us write the prior as $p(\theta|\eta)$, where η denotes the parameters of the prior distribution, *hyperparameters*. By setting a hyperprior, $p(\eta)$, for the hyperparameters we obtain a hierarchical model structure where the fixed parameter values move further away from the data. This allows more flexible models and leads to vaguer prior, which is beneficial if the modeller is unsure of the specific form of the prior. In theory the hierarchy could be extended as far as one desires but there are practical limits how many levels of hierarchy are reasonable (Goel and Degroot, 1981).

The models considered in this dissertation are hierarchical models where the parameter θ is replaced by a latent function $f(\mathbf{x})$. The observation models are build so

that an individual observation depends on a function value at a certain input location $\mathbf{x}$. The latent function $f(\mathbf{x})$ is given a Gaussian process (GP) prior (Rasmussen and Williams, 2006), whose properties are defined by a mean and covariance function, and the hyperparameters related to them. The hierarchy is continued to the third level by giving a hyperprior for the covariance function parameters. The assumption is that there is a functional description for the studied phenomenon, which we are not aware of, and the observations are noisy realizations of this underlying function. The power of this construction lies in the flexibility and non-parametric form of the GP prior. We can use simple parametric observation models that describe the assumed observation noise. The assumptions about the functional form of the phenomenon are encoded in the GP prior. Since GP is a non-parametric model we do not need to fix the functional form of the latent function, but we can give implicit statements of it. These statements are encoded in the mean and covariance function, which determine, for example, the smoothness and variability of the function.

Despite their attractive theoretical properties, GPs provide practical challenges. Their main complication relates to computational time which increases very rapidly as a function of a number of data points. The other challenge is the analytically intractable inference, which exacerbates the slow computational time. The goal of this dissertation is to study methods to alleviate these problems. The treatment will be divided into two separate methods. First, the inference problem is attacked. When inferring the posterior of the hyperparameters or the latent function (the posterior process) one needs to conduct a series of computational steps each of which scales as $O(n^3)$, where n is the number of data points. The aim is to reduce the number of needed steps. For example, a commonly used method of approximating posterior distributions is via Markov chain Monte Carlo (MCMC), where each Markov step requires evaluation of a (unnormalized) posterior density. In MCMC methods, one needs hundreds of such evaluations, at the minimum. In comparison, giving an analytic approximation to the posterior distribution may require only tens of optimization steps, at maximum. Thus, computational savings are considerable. The second treatment tries to relieve the $O(n^3)$ scaling in computational time. The solution to this is searched from the sparse approximations to Gaussian processes and compactly supported (CS) covariance functions. The computational time with sparse approximations scales as $O(nm^2)$, where $m < n$. How small m can be, depends on the data. The CS covariance functions lead to computations which scale, in general, as $O(n^3)$ but with smaller constant than with traditional covariance functions. The size of the constant depends, again, on data. It turns out that sparse approximations and CS covariance functions are complementary. With data sets where m should be large the constant factor related to compactly supported covariance functions is usually small. And vice versa.

This work is organized as follows. The compendium part of the thesis consists of six chapters that summarize the goals, techniques and discoveries of the research. The published articles are attached at the end of the dissertation. The compendium part is not meant to be an extensive description of the results and scientific contribution in the thesis but to summarize the essential background theory and the main findings. It is in a sense complementary to the publications and hopefully serves as an easy starting point for those less familiar with the research subject. This first chapter has introduced the basics of the Bayesian inference and summarized the goals of the research. Chapter 2

gives a short introduction to Gaussian processes and the general description of models considered in the dissertation. Chapter 3 summarizes the inference methods used in the attached publications. The treatment is at a rather general level at this point and more detailed description is given in the publications **I**-**V**. Chapter 4 is devoted for sparse GPs, which form an essential part of the attached publications **I**, **II** and **IV** – **VI**. Chapter 5 gives a short summary of the publications and the main results in them. Chapter 6 discusses the results.

Chapter 2

Gaussian process models

The probabilistic models considered in this work are build upon the Gaussian process (GP), which is used to define prior distributions over latent functions. The general form of the models can be written as follows:

$$\text{observation model:} \qquad \mathbf{y} \sim \prod_{i=1}^{n} p(y_i|f_i,\gamma) \tag{2.1}$$

$$\text{GP prior:} \qquad f(\mathbf{x})|\theta \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x},\mathbf{x}'\,|\theta)\right) \tag{2.2}$$

$$\text{hyperprior:} \qquad \theta,\gamma \sim p(\theta)p(\gamma). \tag{2.3}$$

Here $\mathbf{y} = [y_1,...,y_n]^{\mathrm{T}}$ is a vector of observations (target values) at (input) locations $\mathbf{X} = \{\mathbf{x}_i = [x_{i,1},...,x_{i,d}]^{\mathrm{T}}\}_{i=1}^{n}$. $f(\mathbf{x})$ is a latent function with value $f_i = f(\mathbf{x}_i)$ at input location $\mathbf{x}_i$. The boldface notation will denote a set of latent variables in a vector $\mathbf{f} = [f_1,...,f_n]^{\mathrm{T}}$. Here, the inputs are real valued vectors $\mathbf{x} \in \Re^d$ but in general other inputs, such as strings or graphs, are possible as well. θ collects the hyperparameters in the covariance function $k(\mathbf{x},\mathbf{x}'\,|\theta)$, and γ collects the hyperparameters in the observation model $p(y_i|f_i,\gamma)$. The notation will be slightly abused since $p(y_i|\cdot)$ is used also for the likelihood, which is the same equation as the observation model but a function of parameters instead of y_i. The mean function is considered zero, $m(\mathbf{x}) \equiv 0$, throughout the work since this simplifies the notation. The zero mean is also used in the attached publications since in methodological works (publications **II**, **III**, and **V**) it follows the standard convention and in the more applied spatial models (publications **I**, **IV** and **VI**) the standardization of the data before inference ensures that zero mean is justified (see for example publication **IV**).

The advantage of using GPs is that we can conduct the inference directly in the function space $f(\mathbf{x}) : \Re^d \rightarrow \Re$. Formal definition for the process is given as (e.g. Rasmussen and Williams, 2006):

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

By definition, any set of function values $\mathbf{f}$, indexed by the input co-ordinates $\mathbf{X}$, have a multivariate Gaussian prior distribution

$$p(\mathbf{f}\,|\mathbf{X},\theta) = \mathrm{N}(\mathbf{f}\,|\mathbf{0},\mathbf{K}_{\mathrm{f,f}}), \tag{2.4}$$

where $\mathbf{K}_{\mathrm{f,f}}$ is the covariance matrix. Notice, that the prior over functions will be denoted by $\mathcal{GP}(\cdot,\cdot)$, whereas the prior over finite set of latent variables will be denoted by $\mathrm{N}(\cdot,\cdot)$. The covariance matrix is constructed from a covariance function, $[\mathbf{K}_{\mathrm{f,f}}]_{i,j} = k(\mathbf{x}_i,\mathbf{x}_j|\theta)$, which characterizes the correlations between different points in the process $\mathrm{E}[f(\mathbf{x}_i), f(\mathbf{x}_j)] = k(\mathbf{x}_i,\mathbf{x}_j|\theta)$ (remember that the prior mean is explicitly set zero in this work). Covariance function encodes the prior assumptions of the latent function, such as the smoothness and scale of the variation, and can be chosen freely as long as the covariance matrices produced are symmetric and positive semi-definite ($\mathbf{v}^{\mathrm{T}}\,\mathbf{K}_{\mathrm{f,f}}\,\mathbf{v} \geq 0, \forall \mathbf{v} \in \Re^n$). An example of a stationary covariance function is the squared exponential

$$k_{\mathrm{se}}(\mathbf{x}_i,\mathbf{x}_j\,|\theta) = \sigma_{\mathrm{se}}^2 \exp\left(-\sum_{k=1}^{d}(x_{i,k}-x_{j,k})^2/l_k^2\right), \tag{2.5}$$

where $\theta = \{\sigma_{\mathrm{se}}^2, l_1, ..., l_d\}$. Here, σ_{se}^2 is the scaling parameter, and l_k is the length-scale, which governs how fast the correlation decreases as the distance increases in the direction k. In this work, the length-scale differs from the standard convention by factor $\sqrt{2}$ (see e.g. Rasmussen and Williams, 2006). See Figure 2.1 for illustration. The squared exponential is only one example of possible covariance functions. Few other functions will be discussed in Chapter 4 and more detailed discussion on common covariance functions and their properties is given by, for example, in (Diggle and Ribeiro, 2007; Finkenstädt et al., 2007; Rasmussen and Williams, 2006).

By definition, the marginal distribution of any subset of latent variables, the function values at fixed input points, can be constructed by simply taking the appropriate submatrix of the covariance and subvector of the mean. Imagine, that we want to predict the values $\tilde{\mathbf{f}}$ at new input locations $\tilde{\mathbf{X}}$. The joint prior for latent variables at observation $\mathbf{X}$ and prediction locations $\tilde{\mathbf{X}}$ is

$$\begin{bmatrix}\mathbf{f}\\ \tilde{\mathbf{f}}\end{bmatrix}|\mathbf{X},\tilde{\mathbf{X}},\theta \sim \mathrm{N}\left(\mathbf{0},\begin{bmatrix}\mathbf{K}_{\mathrm{f,f}} & \mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}}\\ \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}} & \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}}\end{bmatrix}\right), \tag{2.6}$$

where $\mathbf{K}_{\mathrm{f,f}} = k(\mathbf{X},\mathbf{X}|\theta)$, $\mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}} = k(\mathbf{X},\tilde{\mathbf{X}}|\theta)$ and $\mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} = k(\tilde{\mathbf{X}},\tilde{\mathbf{X}}|\theta)$. Here, the covariance function $k(\cdot,\cdot)$ denotes also vector and matrix valued functions $k(\mathbf{x},\mathbf{X}) : \Re^d \times \Re^{d\times n} \rightarrow \Re^{1\times n}$, and $k(\mathbf{X},\mathbf{X}) : \Re^{d\times n} \times \Re^{d\times n} \rightarrow \Re^{n\times n}$. The marginal distribution of $\tilde{\mathbf{f}}$ is $p(\tilde{\mathbf{f}}|\tilde{\mathbf{X}},\theta) = \mathrm{N}(\tilde{\mathbf{f}}|\mathbf{0},\mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}})$ like the marginal distribution of $\mathbf{f}$ given in (2.4). This marginal is also called a prior predictive distribution since it is not conditioned to any observations. The conditional distribution of a set of latent variables given other set of latent variables is Gaussian as well. For example, the distribution of $\tilde{\mathbf{f}}$ given $\mathbf{f}$ is

$$\tilde{\mathbf{f}}|\,\mathbf{f},\mathbf{X},\tilde{\mathbf{X}},\theta \sim \mathrm{N}(\mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\,\mathbf{K}_{\mathrm{f,f}}^{-1}\,\mathbf{f}, \mathbf{K}_{\tilde{\mathrm{f}},\tilde{\mathrm{f}}} - \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}\,\mathbf{K}_{\mathrm{f,f}}^{-1}\,\mathbf{K}_{\mathrm{f},\tilde{\mathrm{f}}}), \tag{2.7}$$

which can be interpreted as the posterior predictive distribution for $\tilde{\mathbf{f}}$ after observing the function values at locations $\mathbf{X}$. The mean and covariance of the conditional distribution are functions of input vector $\tilde{\mathbf{x}}$ and $\mathbf{X}$ plays the role of fixed parameters.

Thus, the above distribution generalizes to a Gaussian process with mean function $m_{\mathrm{p}}(\tilde{\mathbf{x}}) = k(\tilde{\mathbf{x}}, \mathbf{X})\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, \mathbf{f}$ and covariance $k_{\mathrm{p}}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = k(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') - k(\tilde{\mathbf{x}}, \mathbf{X})\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, k(\mathbf{X}, \tilde{\mathbf{x}}')$, which define the posterior distribution of the latent function $f(\tilde{\mathbf{x}})$. The posterior GP is illustrated in Figure 2.2.

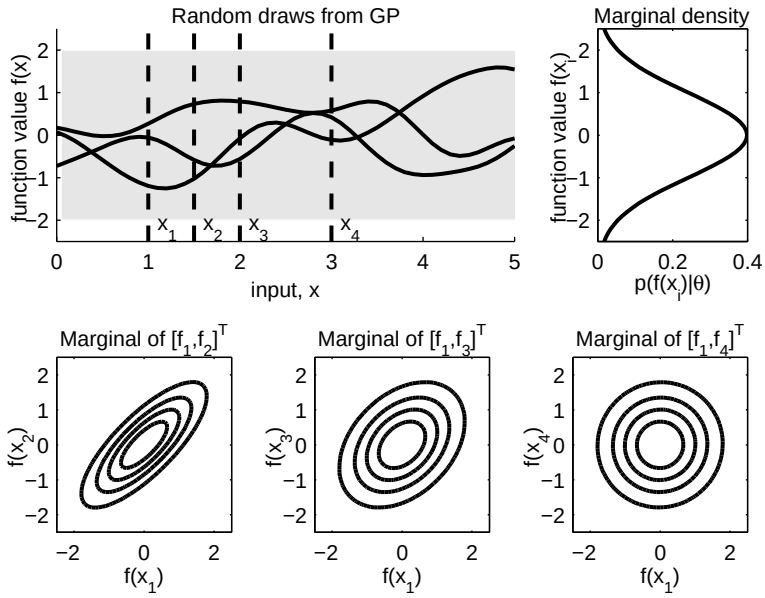


Figure 2.1: An illustration of a Gaussian process. The upper left figure presents three functions drawn randomly from a zero mean GP with squared exponential covariance function. The hyperparameters are $l = 1$ and $\sigma^2 = 1$ and the grey shading represents central 95% probability interval. The upper right subfigure presents the marginal distribution for a single function value. The lower subfigures present three marginal distributions between two function values at distinct input locations shown in the upper left subfigure by dashed line. It can be seen that the correlation between function values $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ is the greater the closer $\mathbf{x}_i$ and $\mathbf{x}_j$ are to each others.

As will be seen, the class of models described by the equations (2.1)-(2.3) is rather rich. Even though the observation model is assumed to be factorizable given the latent variables $f(\mathbf{x}_1), ..., f(\mathbf{x}_n)$, the correlations between the observations are incorporated into the model via the GP prior, and the marginalized observation model $p(\mathbf{y}\,|\gamma, \theta) = \int d\,\mathbf{f}\, p(\mathbf{f}\,|\theta) \prod_{i=1}^{n} p(y_i|f_i, \gamma)$ is no longer factorizable. The models considered here are also rather old since utilizing Gaussian processes is certainly not a new invention. Early examples of their usage can be found, for example, in time series analysis and filtering (Wiener, 1949), and geostatistics (e.g. Matheron, 1973). GPs are still widely and actively used in these fields and usefull overviews are provided by Cressie (1993), Grewal and Andrews (2001), Diggle and Ribeiro (2007), and Gelfand et al. (2010). O'Hagan (1978) was one of the firsts to consider Gaussian processes in

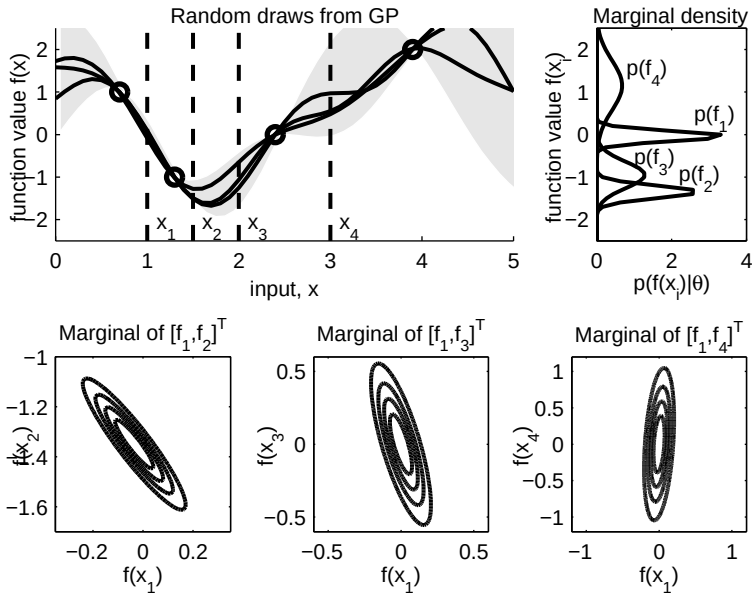


Figure 2.2: A conditional (posterior) GP $p(\tilde{f}|\,\mathbf{f},\theta)$. The observations $\mathbf{f} = [f(0.7) = 1, f(1.3) = -1, f(2.4) = 0, f(3.9) = 2]^{\mathrm{T}}$ are plotted with circles in the upper left subfigure and the prior GP is illustrated in the figure 2.1. When comparing the subfigures to the equivalent ones in Figure 2.1 we can see clear distinction between the marginal and the conditional GP. Here, all the function samples travel through the observations, the mean is no longer zero and the covariance is non-stationary.

a general probabilistic modeling context. He provided a general theory of Gaussian process prediction and utilized it for a number of regression problems. This general regression framework was later rediscovered as an alternative for neural network models (Williams and Rasmussen, 1996; Rasmussen, 1996), and extended for other problems than regression (Neal, 1997; Williams and Barber, 1998). This machine learning perspective is comprehensively summarized by Rasmussen and Williams (2006).

Chapter 3

Inference and prediction

When conducting inference, the interest is in the posterior distributions of the hyperparameters and the latent function, as well as in the predictive distribution of new observations (possibly at new input locations). In an ideal situation, all the desired distributions could be solved analytically, but unfortunately this is not possible in general. This chapter illustrates how the posterior distributions can be approximated. First are discussed methods for evaluating (or approximating) the conditional posterior of latent variables,

$$p(\mathbf{f}\,|\mathcal{D},\theta,\gamma) = \frac{p(\mathbf{y}\,|\,\mathbf{f},\gamma)p(\mathbf{f}\,|\mathbf{X},\theta)}{\int p(\mathbf{y}\,|\,\mathbf{f},\gamma)p(\mathbf{f}\,|\mathbf{X},\theta)d\,\mathbf{f}}, \tag{3.1}$$

where the problem is the integral over $\mathbf{f}$. Above, the training data is denoted by $\mathcal{D} = \{\mathbf{X},\mathbf{y}\}$. Section 3.2 treats the problem of marginalizing over the hyperparameters to obtain the marginal posterior distribution for the latent variables

$$p(\mathbf{f}\,|\mathcal{D}) = \int p(\mathbf{f}\,|\mathcal{D},\theta,\gamma)p(\theta,\gamma|\mathcal{D})d\theta d\gamma. \tag{3.2}$$

The question how to approximate the marginal posterior of the hyperparameters $p(\theta,\gamma|\mathcal{D})$ is left for less attention and the topic is touched only shortly in the section 3.2.

The above considerations generalize straightforwardly to the evaluation of the posterior predictive distribution of latent function, for which we may evaluate first the conditional posterior $p(\tilde{f}|\mathcal{D},\theta,\gamma,\tilde{\mathbf{x}})$ and then marginalize over the hyperparameters to obtain $p(\tilde{f}|\mathcal{D},\tilde{\mathbf{x}})$. The conditional predictive distribution of a new observation can be evaluated for each $\tilde{y}_i$ separately since the observation model is assumed to be factorizable. Thus, we need to be able to evaluate the one dimensional integral

$$p(\tilde{y}_i|\tilde{\mathbf{x}}_i,\mathcal{D},\theta,\gamma) = \int p(\tilde{y}_i|\tilde{f}_i,\gamma)p(\tilde{f}_i|\tilde{\mathbf{x}}_i,\mathcal{D},\theta,\gamma)d\tilde{f}_i, \tag{3.3}$$

after which the marginal posterior predictive distribution $p(\tilde{y}_i|\tilde{\mathbf{x}}_i,\mathcal{D})$ is evaluated analogously to the marginal posteriors $p(\mathbf{f}\,|\mathcal{D})$ and $p(\tilde{f}|\mathcal{D},\tilde{\mathbf{x}})$.

3.1 Conditional posterior of the latent function

3.1.1 Posterior mean and covariance

If the hyperparameters are considered fixed, GP's marginalization and conditionalization properties can be exploited, for example, in prediction. Assume that we have found the conditional posterior distribution $p(\mathbf{f} \,|\mathcal{D}, \theta, \gamma)$, which, in general, is not Gaussian. We can then evaluate the posterior predictive mean simply by using the expression of the conditional mean $\mathrm{E}_{\tilde{f}|\,\mathbf{f},\theta,\gamma}[f(\tilde{\mathbf{x}})] = k(\tilde{\mathbf{x}}, \mathbf{X})\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, \mathbf{f}$ (see equation (2.7) and the text below it) to obtain a parametric posterior mean function

$$m_p(\tilde{\mathbf{x}}|\theta, \gamma) = \int \mathrm{E}_{\tilde{f}|\,\mathbf{f},\theta,\gamma}[f(\tilde{\mathbf{x}})] p(\mathbf{f} \,|\mathcal{D}, \theta, \gamma) d\,\mathbf{f} = k(\tilde{\mathbf{x}}, \mathbf{X})\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, \mathrm{E}_{\mathbf{f}\,|\mathcal{D},\theta,\gamma}[\mathbf{f}]. \quad (3.4)$$

The posterior predictive covariance between any set of latent variables, $\tilde{\mathbf{f}}$, can be evaluated from the relation (see, for example, Gelman et al., 2004, page 23 for justification)

$$\mathrm{Cov}_{\tilde{\mathbf{f}}|\mathcal{D},\theta,\gamma}[\tilde{\mathbf{f}}] = \mathrm{E}_{\mathbf{f}\,|\mathcal{D},\theta,\gamma}\left[\mathrm{Cov}_{\tilde{\mathbf{f}}|\,\mathbf{f}}[\tilde{\mathbf{f}}]\right] + \mathrm{Cov}_{\mathbf{f}\,|\mathcal{D},\theta,\gamma}\left[\mathrm{E}_{\tilde{\mathbf{f}}|\,\mathbf{f}}[\tilde{\mathbf{f}}]\right], \quad (3.5)$$

where the first term simplifies to the conditional covariance in equation (2.7) and the second term can be written as $k(\tilde{\mathbf{x}}, \mathbf{X})\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, \mathrm{Cov}_{\mathbf{f}\,|\mathcal{D},\theta,\gamma}[\mathbf{f}]\, \mathbf{K}_{\mathrm{f,f}}^{-1}\, k(\mathbf{X}, \tilde{\mathbf{x}}')$. Plugging these into the equation and simplifying gives us the posterior covariance function

$$k_p(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = k(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') - k(\tilde{\mathbf{x}}, \mathbf{X}) \left(\mathbf{K}_{\mathrm{f,f}}^{-1} - \mathbf{K}_{\mathrm{f,f}}^{-1}\, \mathrm{Cov}_{\mathbf{f}\,|\mathcal{D},\theta,\gamma}[\mathbf{f}]\, \mathbf{K}_{\mathrm{f,f}}^{-1}\right) k(\mathbf{X}, \tilde{\mathbf{x}}'). \quad (3.6)$$

Even if the exact posterior distribution $p(\tilde{f}|\mathcal{D}, \theta, \gamma)$, or in other words the posterior process, was not analytically solvable we can still evaluate its posterior mean and covariance functions easily, as long as we are able to solve the mean $\mathrm{E}_{\mathbf{f}\,|\mathcal{D},\theta,\gamma}$ and covariance $\mathrm{Cov}_{\mathbf{f}\,|\mathcal{D},\theta,\gamma}[\mathbf{f}]$. Following, for example, Csató and Opper (2002) the conditional posterior mean can be written as

$$\mathrm{E}_{\mathbf{f}\,|\mathcal{D},\theta,\gamma}[\mathbf{f}] = \mathbf{K}_{\mathrm{f,f}}\, \frac{\int d\,\mathbf{f}\, p(\mathbf{f}) \partial p(\mathbf{y} \,|\, \mathbf{f}) / \partial\, \mathbf{f}}{p(\mathcal{D}|\theta, \gamma)}, \quad (3.7)$$

and a similar result can be obtained for the covariance. The problem with the exact formulas is that the integrals in them cannot be computed exactly. The common practice to approximate the posterior distribution $p(\mathbf{f} \,|\mathcal{D}, \theta, \gamma)$ is either with Markov chain Monte Carlo (MCMC) (e.g. Neal, 1997, 1998; Diggle et al., 1998; Kuss and Rasmussen, 2005; Christensen et al., 2006) or by giving an analytic approximation to it (e.g. Williams and Barber, 1998; Gibbs and Mackay, 2000; Minka, 2001; Csató and Opper, 2002; Rue et al., 2009). The analytic approximations considered here assume a Gaussian form in which case it is natural to approximate the posterior predictive distribution with Gaussian as well. In this case the equations (3.4) and (3.6) give its mean and covariance. The Gaussian approximation can be justified if the conditional posterior is unimodal, which it is if the likelihood is log concave, and there is enough data so that the posterior will be close to Gaussian. However, invoking the central limit theorem with GP models is not straightforward since the number of observations may

grow either alongside the latent variables or per latent variable. Examples of the former are traditional GP regression and classification (publications **II**, **III** and **V**) and the latter case is present, for example, in disease mapping models (publications **I**, **IV** and **VI**) where the background population in a grid cell may increase albeit the number of latent variables remains constant. The central limit theorem may apply in the increase alongside latent variables case as well if the effective number of latent variables remains small compared to the number of observations. The goodness of the Gaussian approximation is well discussed, for example, by Rue et al. (2009). A pragmatic justification for using Gaussian approximation is that many times it suffices to approximate well the mean and variance of the latent function. These, on the other hand, fully define Gaussian distribution and one can approximate the integrals over $\tilde{f}_i$ by using the Gaussian form for its conditional posterior.

3.1.2 Gaussian observation model

A special case of an observation model, for which the conditional posterior of the latent variables can be evaluated analytically, is the Gaussian distribution, $y_i \sim \mathrm{N}(f_i, \sigma^2)$, where the parameter γ is replaced by the noise variance σ^2. In this case, both the likelihood and the prior are Gaussian functions of the latent variable, and we are able to analytically integrate over $\mathbf{f}$ to obtain the marginal likelihood of the hyperparameters

$$p(\mathbf{y}\,|\theta, \sigma^2) = \mathrm{N}(\mathbf{y}\,|\mathbf{0}, \mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I}). \tag{3.8}$$

Setting this in the denominator of the equation (3.1), re-arranging the terms and simplifying the expression gives a Gaussian distribution also for the conditional posterior of the latent variables

$$\mathbf{f}\,|\mathcal{D}, \theta, \sigma^2 \sim \mathrm{N}(\mathbf{K}_{\mathrm{f,f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{y}, \mathbf{K}_{\mathrm{f,f}} - \mathbf{K}_{\mathrm{f,f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{K}_{\mathrm{f,f}}). \tag{3.9}$$

Since the conditional posterior of $\mathbf{f}$ is Gaussian, the posterior process, or distribution $p(\tilde{f}|\mathcal{D}, \theta, \sigma^2)$, is also Gaussian. The predictive mean and covariance can be evaluated by placing the mean and covariance from (3.9) in the equations (3.4) and (3.6), after which we obtain the predictive distribution

$$\tilde{f}|\mathcal{D}, \theta, \sigma^2 \sim \mathcal{GP}\left(m_{\mathrm{p}}(\tilde{\mathbf{x}}), k_{\mathrm{p}}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')\right) \tag{3.10}$$

where the mean and covariance are $m_{\mathrm{p}}(\tilde{\mathbf{x}}) = k(\tilde{\mathbf{x}}, \mathbf{X})(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{y}$ and $k_{\mathrm{p}}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = k(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') - k(\tilde{\mathbf{x}}, \mathbf{X})(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}k(\mathbf{X}, \tilde{\mathbf{x}}')$. The predictive distribution for new observations $\tilde{\mathbf{y}}$ can be obtained by integrating $p(\tilde{\mathbf{y}}|\mathcal{D}, \theta, \sigma^2) = \int p(\tilde{\mathbf{y}}|\tilde{\mathbf{f}}, \sigma^2)p(\tilde{\mathbf{f}}|\mathcal{D}, \theta, \sigma^2)d\tilde{\mathbf{f}}$. The result is, again, Gaussian with mean $\mathrm{E}_{\tilde{\mathbf{f}}|\mathcal{D},\theta}[\tilde{\mathbf{f}}]$ and covariance $\mathrm{Cov}_{\tilde{\mathbf{f}}|\mathcal{D},\theta}[\tilde{\mathbf{f}}] + \sigma^2\mathbf{I}$.

3.1.3 Laplace approximation

In the Laplace approximation the mean is approximated by the posterior mode of $\mathbf{f}$ and the covariance by the curvature of the log posterior at the mode. The approximation is constructed from the second order Taylor expansion of $\log p(\mathbf{f}\,|\mathcal{D}, \theta)$ around the mode $\hat{\mathbf{f}}$, which gives a Gaussian approximation to the conditional posterior

$$p(\mathbf{f}\,|\mathcal{D}, \theta, \gamma) \approx q(\mathbf{f}\,|\mathcal{D}, \theta, \gamma) = \mathrm{N}(\mathbf{f}\,|\hat{\mathbf{f}}, \boldsymbol{\Sigma}), \tag{3.11}$$

where $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f}\,|\mathcal{D},\theta,\gamma)$ and $\mathbf{\Sigma}^{-1}$ is the Hessian of the negative log conditional posterior at the mode (Gelman et al., 2004; Rasmussen and Williams, 2006):

$$\mathbf{\Sigma}^{-1} = -\nabla\nabla \log p(\mathbf{f}\,|\mathcal{D},\theta,\gamma)|_{\mathbf{f}=\hat{\mathbf{f}}} = \mathbf{K}_{\mathrm{f,f}}^{-1} + \mathbf{W}. \tag{3.12}$$

Since the likelihood is factorizable here, $\mathbf{W}$ is a diagonal matrix with entries $\mathbf{W}_{ii} = \nabla_{f_i}\nabla_{f_i} \log p(y|f_i,\gamma)|_{f_i=\hat{f}_i}$. The approximation scheme is called the Laplace method following Williams and Barber (1998), but essentially the same approximation is named Gaussian approximation by Rue et al. (2009) in their Integrated nested Laplace approximation (INLA) scheme for Gaussian Markov random field models.

The posterior mean of $f(\tilde{\mathbf{x}})$ can be approximated from the equation (3.4) by replacing the posterior mean $\mathrm{E}_{\mathbf{f}\,|\mathcal{D},\theta}[\mathbf{f}]$ by $\hat{\mathbf{f}}$. The posterior covariance is approximated similarly by using $(\mathbf{K}_{\mathrm{f,f}}^{-1} + \mathbf{W})^{-1}$ in the place of $\mathrm{Cov}_{\mathbf{f}\,|\mathcal{D},\theta}[\mathbf{f}]$. Thus, after some rearrangements and using $\mathbf{K}_{\mathrm{f,f}}^{-1}\,\hat{\mathbf{f}} = \nabla \log p(\mathbf{y}\,|\,\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$, the approximate posterior predictive distribution is

$$\tilde{f}|\mathcal{D},\theta,\sigma^2 \sim \mathcal{GP}\left(m_{\mathrm{p}}(\tilde{\mathbf{x}}), k_{\mathrm{p}}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}')\right), \tag{3.13}$$

where the mean and covariance are $m_{\mathrm{p}}(\tilde{\mathbf{x}}) = k(\tilde{\mathbf{x}},\mathbf{X})\nabla \log p(\mathbf{y}\,|\,\mathbf{f})|_{\mathbf{f}=\hat{\mathbf{f}}}$ and $k_{\mathrm{p}}(\tilde{\mathbf{x}},\tilde{\mathbf{x}}') = k(\tilde{\mathbf{x}},\tilde{\mathbf{x}}') - k(\tilde{\mathbf{x}},\mathbf{X})(\mathbf{K}_{\mathrm{f,f}} + \mathbf{W}^{-1})^{-1}k(\mathbf{X},\tilde{\mathbf{x}}')$ The approximate conditional predictive density of a new observation $\tilde{y}_i$ can now be evaluated, for example, with quadrature integration over each $\tilde{f}_i$ separately

$$p(\tilde{y}_i|\mathcal{D},\theta,\gamma) \approx \int p(\tilde{y}_i|\tilde{f}_i,\gamma)q(\tilde{f}_i|\mathcal{D},\theta,\gamma)d\tilde{f}_i. \tag{3.14}$$

3.1.4 Expectation propagation algorithm

The Expectation propagation (EP) algorithm is a general method for approximating integrals over functions that factor into simple terms (Minka, 2001). The Laplace method constructs a Gaussian approximation at the posterior mode and approximates the posterior covariance via the curvature of the log density at that point. EP, for its part, tries to minimize the Kullback-Leibler divergence from the true posterior to its approximation. EP approximates the conditional posterior with

$$p(\mathbf{f}\,|\mathcal{D},\theta,\gamma) \approx q(\mathbf{f}\,|\mathcal{D},\theta,\gamma) = \frac{1}{Z_{\mathrm{EP}}} p(\mathbf{f}\,|\theta)\prod_{i=1}^{n} t_i(f_i|\tilde{Z}_i,\tilde{\mu}_i,\tilde{\sigma}_i^2), \tag{3.15}$$

where the likelihood terms have been replaced by site functions $t_i(f_i|\tilde{Z}_i,\tilde{\mu}_i,\tilde{\sigma}_i^2) = \tilde{Z}_i\,\mathrm{N}(f_i|\tilde{\mu}_i,\tilde{\sigma}_i^2)$ and the normalizing constant by Z_{EP}.

EP algorithm updates the site parameters $\tilde{Z}_i$, $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ sequentially. At each iteration, first the i'th site is removed from the i'th marginal posterior to obtain a cavity distribution $q_{-i}(f_i) = q(f_i|\mathcal{D},\theta)/t_i(f_i)$. Second step is to find a Gaussian $\hat{q}(f_i)$ to which the Kullback-Leibler divergence from the cavity distribution multiplied by the i'th exact likelihood term is minimized $\hat{q}(f_i) = \arg\min_q \mathrm{KL}\left(q_{-i}(f_i)p(y_i|f_i)||q(f_i)\right)$. This is equivalent to matching the first and second moment between the two distributions (Seeger, 2005). The site terms $\tilde{Z}_i$ are scaling parameters which ensure that also the zeroth moment of the approximate and exact posterior match, that is $Z_{\mathrm{EP}} \approx p(\mathcal{D}|\theta,\gamma)$.

After the moments are solved, the parameters of the local approximation t_i are updated so that the new marginal posterior $q_{-i}(f_i)t_i(f_i)$ matches with the moments of $\hat{q}(f_i)$. For last, the parameters of the approximate posterior (3.15) are updated to give

$$p(\mathbf{f}\,|\mathcal{D},\theta,\gamma) \approx \mathrm{N}(\mathbf{f}\,|\,\mathbf{K}_{\mathrm{f,f}}(\mathbf{K}_{\mathrm{f,f}}+\tilde{\Sigma})^{-1}\tilde{\mu}, \mathbf{K}_{\mathrm{f,f}} - \mathbf{K}_{\mathrm{f,f}}(\mathbf{K}_{\mathrm{f,f}}+\tilde{\Sigma})^{-1}\,\mathbf{K}_{\mathrm{f,f}}), \quad (3.16)$$

where $\tilde{\Sigma} = \mathrm{diag}[\tilde{\sigma}_1^2, ..., \tilde{\sigma}_n^2]$ and $\tilde{\mu} = [\tilde{\mu}_1, ..., \tilde{\mu}_n]^{\mathrm{T}}$. The predictive mean and covariance of $\tilde{f}$ are again obtained from equations (3.4) and (3.6). The predictive distribution of a new observation $\tilde{y}$ is derived analogically to the Laplace approximation.

From the equations (3.16), (3.13) and (3.9) it can be seen that there is a great similarity between the exact solution with the Gaussian obsrvation model and the Laplace and EP approximation. The diagonal matrices $\mathbf{W}^{-1}$ and $\tilde{\mathbf{\Sigma}}$ correspond to the noise variance $\sigma^2\mathbf{I}$ in the Gaussian likelihood and the two approximations can be seen as Gaussian approximation for the likelihood (Nickisch and Rasmussen, 2008).

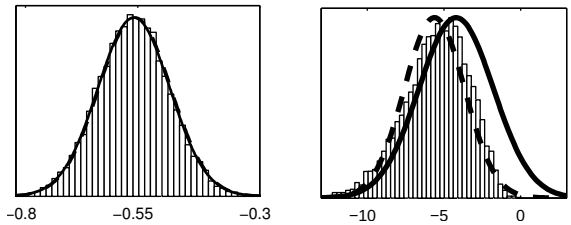
3.1.5 Markov chain Monte Carlo

The accuracy of the two approximations considered this far is limited by the Gaussian form of the approximating function. Another approach, which gives exact solution in the limit of infinite computational time, is to approximate the posterior with Monte Carlo methods (Robert and Casella, 2004). These are based on sampling from $p(\mathbf{f}\,|\mathcal{D},\theta,\ \gamma)$ and using the samples to represent the posterior distribution. In this case, the posterior marginals can be visualized with histograms and posterior statistics approximated with sample means. For example, the posterior expectation of $\mathbf{f}$ is

$$\mathrm{E}_{\mathbf{f}\,|\mathcal{D},\theta,\gamma}[\mathbf{f}] \approx \frac{1}{M}\sum_{i=1}^{M}\mathbf{f}^{(i)}, \quad (3.17)$$

where $\mathbf{f}^{(i)}$ is the i'th sample from the conditional posterior.

The problem with Monte Carlo methods is how to draw samples from arbitrary distributions. The challenge can be overcome with Markov chain Monte Carlo methods (Gilks et al., 1996), where one constructs a Markov chain whose stationary distribution is the posterior distribution $p(\mathbf{f}\,|\mathcal{D},\theta,\ \gamma)$ and uses the Markov chain samples to obtain Monte Carlo estimates. After having the posterior sample of latent variables, we can sample from the posterior predictive distribution of any set of latent variables $\tilde{\mathbf{f}}$ simply by sampling with each $\mathbf{f}^{(i)}$ one $\tilde{\mathbf{f}}^{(i)}$ from $p(\tilde{\mathbf{f}}|\,\mathbf{f}^{(i)},\theta,\gamma)$, which is given in the equation (2.7). Similarly, we can obtain a sample of $\tilde{\mathbf{y}}$ by drawing one $\tilde{\mathbf{y}}^{(i)}$ for each $\tilde{\mathbf{f}}^{(i)}$ from $p(\mathbf{y}\,|\tilde{\mathbf{f}},\theta,\gamma)$. A rather efficient sampling algorithm is hybrid Monte Carlo (HMC) (Duane et al., 1987; Neal, 1996), which utilizes the gradient information of the posterior distribution to direct the sampling to interesting regions. Significant improvement in mixing of the sample chain of the latent variables can be obtained by using the variable transformation proposed by Christensen et al. (2006) (see also publication **I**). The three approximations for the conditional posterior are summarized in Figure 3.1.



(a) Disease mapping (b) classification

Figure 3.1: Illustration of the Laplace approximation (solid line), EP (dashed line) and MCMC (histogram) for the conditional posterior of a latent variable $p(f_i|\mathcal{D},\theta)$ in two applications. On the left, a disease mapping problem with Poisson observation model (publication **IV**) where the Gaussian approximation works well. On the right, a classification problem with probit likelihood (publication **V**) where the posterior is skewed and the Gaussian approximation is clearly a compromise. However, EP approximates the mean and variance better than the Laplace approximation in this case also.

3.2 Marginalization over hyperparameters

3.2.1 Maximum a posterior estimate of hyperparameters

The easiest way to approximate the integral over $p(\theta,\gamma|\mathcal{D})$ is to give the hyperparameters a point estimate such as the maximum a posterior (MAP) estimate

$$\{\hat{\theta},\hat{\gamma}\} = \arg\max_{\theta,\gamma} p(\theta,\gamma|\mathcal{D}) = \arg\max_{\theta,\gamma} \left[\log p(\mathcal{D}|\theta,\gamma) + \log p(\theta,\gamma)\right]. \tag{3.18}$$

In this approximation, the hyperparameter values are given a point mass one at the posterior mode, and, for example, the marginal posterior of latent variables is approximated as $p(\mathbf{f}\,|\mathcal{D}) \approx p(\mathbf{f}\,|\mathcal{D},\hat{\theta},\hat{\gamma})$ (the other posterior marginals come analogously). Alternatively the hyperparameter optimization can be interpreted a model selection over a model family indexed by continuous parameter $\vartheta = [\theta^{\mathrm{T}},\gamma^{\mathrm{T}}]^{\mathrm{T}}$ (Rasmussen and Williams, 2006).

For the MAP estimate one needs to evaluate the log marginal likelihood. In Gaussian case this is straightforward since it has an analytic solution (see equation (3.8)),

$$\log p(\mathcal{D}|\theta,\sigma) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\,\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I}| - \frac{1}{2}\,\mathbf{y}^{\mathrm{T}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\,\mathbf{y}, \tag{3.19}$$

The log marginal likelihood, and thus also the log posterior, is differentiable with respect to the hyperparameters, which allows a gradient based optimization.

If the observation model is not Gaussian the marginal likelihood needs to be approximated. The Laplace approximation to the marginal likelihood is constructed, for example, by writing

$$p(\mathcal{D}|\theta,\gamma) = \int p(\mathbf{y}|\,\mathbf{f},\gamma)p(\mathbf{f}\,|\theta)d\,\mathbf{f} = \int \exp(g(\mathbf{f}))d\,\mathbf{f}, \tag{3.20}$$

and making a second order Taylor expansion of $g(\mathbf{f})$ around $\hat{\mathbf{f}}$. This gives a Gaussian integral over $\mathbf{f}$ multiplied by a constant, and results in the approximation

$$\log p(\mathcal{D}|\theta,\gamma) \approx \log q(\mathcal{D}|\theta,\gamma) \propto \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\hat{\mathbf{f}}^{\mathrm{T}}\,\mathbf{K}_{\mathrm{f,f}}^{-1}\,\hat{\mathbf{f}} - \frac{1}{2}\log|\mathbf{B}|, \tag{3.21}$$

where $|\mathbf{B}| = |I + \mathbf{W}^{1/2}\,\mathbf{K}_{\mathrm{f,f}}\,\mathbf{W}^{1/2}|$. This is the same approximation as the Gaussian approximation by Rue et al. (2009) derived from $p(\mathbf{y},\mathbf{f}\,.\theta,\gamma)/q(\mathbf{f}\,|\mathcal{D},\theta,\gamma)|_{\mathbf{f}=\hat{\mathbf{f}}}$, where the denominator is the Laplace approximation in equation (3.13) (see also Tierney and Kadane, 1986). The gradients of the approximate log marginal likelihood (3.21) can be computed analytically, which enables the use of gradient based optimization with Laplace approximation.

EP's marginal likelihood approximation is its normalization constant

$$Z_{\mathrm{EP}} = \int p(\mathbf{f}\,|X,\theta)\prod_{i=1}^{n}\tilde{Z}_i\,\mathrm{N}(f_i|\tilde{\mu}_i,\tilde{\sigma}_i^2)df_i \tag{3.22}$$

in equation (3.15). This is a Gaussian integral multiplied by constant $\prod_{i=1}^{n}\tilde{Z}_i$, giving

$$\log Z_{\mathrm{EP}} = -\frac{1}{2}\log|K+\tilde{\Sigma}| - \frac{1}{2}\tilde{\mu}^{\mathrm{T}}\left(K+\tilde{\Sigma}\right)^{-1}\tilde{\mu} + C_{\mathrm{EP}}, \tag{3.23}$$

where C_{EP} collects terms that are not explicit functions of θ or γ (there is implicit dependence through the iterative algorithm, though). The parameters C_{EP}, $\tilde{\Sigma}$ and $\tilde{\mu}$ can be considered constants when differentiating the function with respect to the hyperparameters (Seeger, 2005), for which reason the MAP estimate can be found with gradient based optimization methods with EP as well.

The advantage of MAP estimate is that it is relatively easy and fast to evaluate. In the experiments optimization algorithms needed usually at maximum tens of optimization steps to find the mode (see, for example, publication **IV**). The drawback, however, is that it may underestimate the uncertainty in hyperparameters.

3.2.2 Grid integration

The previous section treated methods to evaluate exactly (the Gaussian case) or approximately (Laplace approximation and EP) the marginal posterior $p(\theta,\gamma|\mathcal{D})$ up to the normalization. There the unnormalized posterior was used for optimizing the hyperparameters but it can also be used for exploring the posterior for purposes of numerical integration with a finite sum, such as

$$p(\mathbf{f}\,|\mathcal{D}) \approx \sum_{i=1}^{M} p(\mathbf{f}\,|\mathcal{D},\vartheta_i)p(\vartheta_i|\mathcal{D})\Delta_i. \tag{3.24}$$

Here $\vartheta = [\theta^{\mathrm{T}},\gamma^{\mathrm{T}}]^{\mathrm{T}}$ and Δ_i denotes the area weight appointed to an evaluation point ϑ_i. Thus, the latent variable posterior is a mixture of Gaussians. The other marginal posteriors are approximated similarly with mixture distributions.

In grid integration, the evaluation points are set into a regular grid. The construction of the grid is started from the posterior mode $\hat{\vartheta}$, and continued so that the bulk of the posterior mass is included in the integration. If the grid points are set evenly, the area weights Δ_i are equal. In practice, the construction of the grid is aided by the information about the Hessian of $\log p(\vartheta|\mathcal{D})$ at the mode, which would be the inverse covariance matrix for ϑ if the density were Gaussian. This approximate covariance is used to select the exploration directions and step sizes as illustrated in Figure 3.2(a) and discussed in the publication **IV** and by Rue et al. (2009).

The numerical integration using the grid search is feasible only for a small number of hyperparameters since the number of grid points grows exponentially with the dimension of the hyperparameter space d. For example, the number of the nearest neighbors of the mode increases as $O(3^d)$, which results in 728 grid points already for $d = 6$. If also the second neighbors are included, the number of grid points increases as $O(5^d)$, which results in 15624 grid points for six hyperparameters.

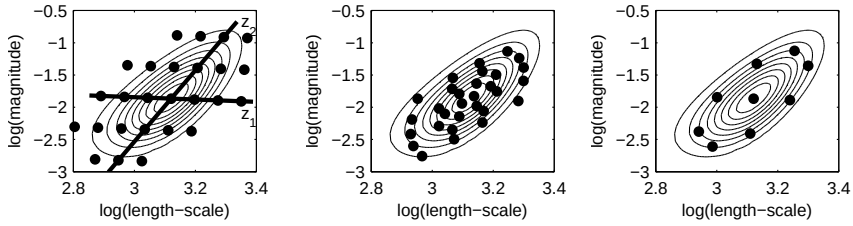
3.2.3 Monte Carlo integration

Monte Carlo integration works better than the grid integration in large hyperparameter spaces since its error decreases with a rate that is independent of the dimension (Robert and Casella, 2004). There are two options to find a Monte Carlo estimate for marginal posteriors, like $p(\mathbf{f}\,|\mathcal{D})$. The first option is to sample just the hyperparameters from their marginal posterior $p(\theta|\mathcal{D})$ or from its approximation given by the Laplace approximation or EP, which is illustrated in Figure 3.2(b). In this case, the posterior marginals are approximated with mixture distributions as in the grid integration but with equal weigths. The other option is to run a full MCMC for all the parameters in the model. That is, we sample both the hyperparameters and the latent variables and estimate the needed posterior statistics by sample estimates or by histograms (Neal, 1997; Diggle et al., 1998). Sampling both, the hyperparameters and latent variables, is usually awfully slow since there is a strong correlation between them. This slows the convergence and mixing of the Markov chain (see publication **I** and **IV**). Sampling from the (approximate) marginal, $p(\theta|\mathcal{D})$, is a much easier task since the parameter space is smaller. Tuning the sampler parameters is also the harder the more parameters are sampled.

Although Monte Carlo integration is more efficient than grid integration, it also has its downside. For most examples, few hundred independent samples are enough for reasonable posterior summaries (Gelman et al., 2004), which seems achievable. The problem, however, is that we are not able to draw independent samples from the posterior. Even with a careful tuning of Markov chain samplers the autocorrelation is usually so large that the required sample size is in thousands, which is a clear disadvantage compared with the MAP estimate, for example.

3.2.4 Central composite design integration

Rue et al. (2009) suggest a central composite design (CCD) for choosing the representative points from the posterior of the hyperparameters when the dimensionality d is moderate or high. In this setting, the integration is considered as a quadratic design



(a) Grid based integration. (b) Monte Carlo integration (c) Central composite design

Figure 3.2: Illustration of the grid based, the Monte Carlo and the central composite design integration over the logarithm of the hyperparameters. The contour shows the posterior density $q(\log(\vartheta)|\mathcal{D})$ and the integration points are marked with dots. The left figure shows also the vectors $\mathbf{z}$ along which the points are searched in the grid integration and central composite desing. The integration is conducted over $q(\log(\vartheta)|\mathcal{D})$ rather than $q(\vartheta|\mathcal{D})$ since the former is closer to Gaussian. (From the publication **IV**.)

problem in a d dimensional space with the aim in finding points that allow for estimating the curvature of the posterior distribution around the mode. The design used by Rue et al. (2009) and in the publication **IV** is the fractional factorial design augmented with a center point and a group of $2d$ star points. In this setting, the design points are all on the surface of a d-dimensional sphere and the star points consist of $2d$ points along each axis. This is illustrated in Figure 3.2(c). The number of the design points grows very moderately and, for example, for $d = 6$ one needs only 45 points. The fractional factorial design is discussed in detail by Sanchez and Sanchez (2005). The CCD integration can be summarized with the equation (3.24) where the weights are evaluated as described in the publication **IV**.

The CCD integration speeds up the integration considerably. The accuracy is between the MAP estimate and the full integration with grid search or Monte Carlo. Rue et al. (2009) and Martino (2007) report good results with this integration scheme, and the results in the publication **IV** are promising as well.

3.3 Summary of the inference methods

The methods treated in this chapter can be arranged in an increasing order of accuracy and computational time. The choice of the method is then a compromise between these two attributes. The inference is the fastest when using MAP estimate for the hyperparameters and Gaussian function for the conditional posterior. With a Gaussian observation model, the Gaussian conditional distribution is exact and the only source of imprecision is the point estimate for the hyperparameters. If the observation model is other than Gaussian, the conditional distribution is an approximation, whose quality depends on, how close to Gaussian the real conditional posterior is, and how well the mean and variance are approximated. The form of the real posterior depends on many things for which reason the Gaussian approximation has to be assessed in-

dependently for every data. Methods for assessing the Gaussian approximation are discussed, for example, by Rue et al. (2009) and in the publication **IV**. Tierney and Kadane (1986) provide asymptotic results for the accuracy of the Laplace approximation and Nickisch and Rasmussen (2008) give extensive comparison between different Gaussian approximations in classification problems. The Laplace approximation is faster than EP but EP approximates better the posterior mean and variance. For example, in classification, this is crucial since the posterior of the latent variables is rather far from normal, as illustrated in Figure 3.1(b) (see also Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008). The publications **III** and **IV** contain examples where Laplace approximation gives, at a practical level, as good results as full MCMC or EP (see also Figure 3.1(a)). At the expense of computational time, the approximation to the marginal posterior of a latent variable could be improved by evaluating correction terms for the EP approximation (Paquet et al., 2009; Cseke and Heskes, 2010) or by improving the Laplace approximation to marginals (Tierney and Kadane, 1986; Rue et al., 2009). These techniques are, however, not on the focus of this work.

A golden standard for the conditional posterior of latent variables can be obtained by an extensive MCMC at the MAP estimate for the hyperparameters. The MAP estimate can be found by utilizing Laplace approximation or EP after which the sampling of the latent variables can be performed efficiently with HMC aided by the variable transformation (Christensen et al., 2006), which was used in the publication **I**. Even if the Laplace approximation and EP lacked in accuracy for the conditional posterior they may approximate the marginal likelihood well. The accuracy of the Laplace approximation depends on the effective number of latent variables and it is usually more accurate for data sets with many observations per input location (Rue et al., 2009). This was the case in disease mapping problems (publications **IV** and **VI**) where Laplace approximation and EP worked practically as well as MCMC. EP has been shown to approximate the marginal likelihood rather accurately in classification problems (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008) as well, where Laplace approximation fails. This suggests that EP's approximation to the marginal likelihood is more reliable. In general, the parameters of the covariance function seem to be weakly identifiable and the predictive performance rather insensitive to the exact values of the length-scale and magnitude. The identifiability of the hyperparameters is well treated, for example, by Diggle et al. (1998), Zhang (2004) and Diggle and Ribeiro (2007) and shortly touched in the publications **IV** and **V**.

When integrating over hyperparameters, the Monte Carlo and grid integration give exact results in the limit of an infinite number of evaluation points. This limit, however, can never be reached, and even finding practically sufficient number of evaluation points may be an overwhelming task. For this reason Rue et al. (2009) proposed to use the grid and CCD integration. Grid integration is very efficient for a small number of hyperparameters (less than 4). CCD tries to incorporate the posterior variance of the hyperparameters into the inference and seems to give good approximation in high dimensions. Since CCD is based on the assumption that the hyperparameter's posterior is (close to) Gaussian, the densities $p(\vartheta_i|\mathcal{D})$ at the points on the circumference should be monitored in order to detect serious discrepancies from this assumption. These densities are identical if the posterior is Gaussian, for which reason great variability on their values indicates that CCD has failed. The posterior of the hyperparameters may be far

from a Gaussian distribution, but for a suitable transformation the approximation may work well. For example, the covariance function parameters should be transformed through logarithm as discussed in the publication **IV**.

There are also other analytic approximations than the Laplace approximation or EP proposed in the literature. Most of these are based on some kind of variational approximation (Gibbs and Mackay, 2000; Csató and Opper, 2002; Tipping and Lawrence, 2005; Kuss, 2006; Opper and Archambeau, 2009). The Laplace approximation and EP were chosen for this work for a few reasons. They both are, in theory, straightforward to implement for any factorizable likelihood (often there are practical problems with the implementation though). They have also been extensively studied and compared to MCMC in classification problems (Kuss and Rasmussen, 2005) where EP is shown to compare well with MCMC whereas Laplace approximation is the number one in computational speed. EP is also shown to outperform the variational approximations in accuracy and to be similar in speed (Nickisch and Rasmussen, 2008). Thus, testing Laplace approximation in other problems was tempting for its speed and EP for its speed and accuracy.

Chapter 4

Sparse Gaussian processes

The evaluation of the inverse and determinant of the covariance matrix scale as $O(n^3)$ in time, and storing the full covariance matrix scales as $O(n^2)$. These rapidly increasing resource needs limit the use of GP models to moderate size data sets. In this chapter, the problem is tackled with sparse approximations to Gaussian process and compactly supported (CS) covariance functions.

4.1 Compactly supported covariance functions

A compactly supported covariance function is a function that gives zero correlation between data points whose distance exceeds a certain threshold leading to a sparse covariance matrix. The challenge in constructing CS covariance functions is to guarantee their positive definiteness. A covariance function with global support can not be cut arbitrarily to obtain a compact support, since the resulting function would not, in general, be positive definite. Sansò and Schuh (1987) provide one of the early implementations of spatial prediction with CS covariance functions. Their functions are build by self-convoluting symmetric kernels with finite support (such as a linear spline). These are, however, special functions for one or two dimensions. Later Wu (1995) introduced radial basis functions with compact support and a generic procedure to construct them. Wendland (1995) developed them further and later, for example, Gaspari and Cohn (1999), Gneiting (1999, 2002), and Buhmann (2001) worked more on the subject.

The CS functions studied in this work are Wendland's piecewise polynomials $k_{pp,q}$ (Wendland, 2005), such as:

$$k_{\text{pp},2} = \frac{\sigma^2_{\text{pp}}}{3}(1-r)_+^{j+2}\left((j^2+4j+3)r^2+(3j+6)r+3\right), \tag{4.1}$$

where $j = \lfloor d/2 \rfloor + 3$ and $r^2 = \sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2/l_k^2$. These functions correspond to processes that are q times mean square differentiable and are positive definite up to input dimension d. Thus, the degree of the polynomial has to be increased alongside the input dimension. The dependence of CS covariance functions to the input dimension is very fundamental. There are no radial compactly supported functions that are positive

definite on every $\Re^d$ but they are always restricted to a finite number of dimensions (see e.g. Wendland, 1995, theorem 9.2).

The key idea with using CS covariance functions is that, roughly speaking, one uses only the nonzero elements of the covariance matrix in the calculations. This may speed up the calculations substantially since in some situations only a fraction of the elements of the covariance matrix are non-zero. In practice, efficient sparse matrix routines are needed (Davis, 2006), which are nowadays a standard utility in many statistical computing packages, such as Matlab or R, or available as an additional package for them. The CS covariance functions have been rather widely studied in the geostatistics applications. The early works concentrated on their theoretical properties and aimed to approximate the known globally supported covariance functions (Gneiting, 2002; Furrer et al., 2006; Moreaux, 2008). There the computational speed-up is obtained using efficient linear solvers for the prediction equation $\tilde{\mathbf{f}} = \mathbf{K}_{\tilde{\mathrm{f}},\mathrm{f}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2)^{-1}\,\mathbf{y}$. The hyperparameters are fitted to either the empirical covariance or the global support covariance function. Kaufman et al. (2008) study the maximum likelihood estimates for tapered covariance functions (those are products of globally supported and CS covariance functions). There, the magnitude can be solved analytically and the length-scale (only one length-scale for all the input dimensions is used) is optimized using a line search in one dimension. The benefits from a sparse covariance matrix have been immediate since the problems collapse into solving sparse linear systems. In case of multiple length-scales line search is not feasible and gradient based optimization becomes more advantageous. However, utilizing the gradient of the log posterior of the hyperparameters needs some extra sparse matrix tools.

The problematic part is the trace in the derivative of the log marginal likelihood, for example

$$\begin{aligned}\frac{\partial}{\partial\theta}\log p(\mathbf{y}|\mathbf{X},\theta) =& \frac{1}{2}\mathbf{y}^{\mathrm{T}}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\frac{\partial(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})}{\partial\theta}(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \\ & - \frac{1}{2}\mathrm{tr}\left((\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})^{-1}\frac{\partial(\mathbf{K}_{\mathrm{f,f}} + \sigma^2\mathbf{I})}{\partial\theta}\right). \qquad (4.2)\end{aligned}$$

If implemented naively, the trace requires the inverse of a sparse covariance matrix, which is not sparse in general. Luckily, Takahashi et al. (1973) introduced an algorithm whereby we can evaluate a sparsified version of the inverse of a sparse matrix. This can be utilized in the gradient evaluations as described in the publication **II**. The same problem was considered by Storkey (1999) who used the covariance matrices of Toeplitz form, which are fast to handle due their banded structure. However, constructing Toeplitz covariance matrices is not possible in two or higher dimensions without approximations. Also the EP algorithm requires special considerations with CS covariance functions. The posterior covariance $\mathbf{K}_{\mathrm{f,f}} - \mathbf{K}_{\mathrm{f,f}}(\mathbf{K}_{\mathrm{f,f}} + \tilde{\Sigma})^{-1}\,\mathbf{K}_{\mathrm{f,f}}$ in equation (3.16) does not remain sparse, for which reason it has to be expressed implicitly during the updates. The EP algoritm is treated in the publications **IV** and **V**.

4.2 FIC and PIC sparse approximations

Snelson and Ghahramani (2006) proposed a sparse pseudo-input Gaussian process

(SPGP), which Quiñonero-Candela and Rasmussen (2005) named later fully independent training conditional (FITC). The original idea in SPGP was that the sparse approximation is used only in the training phase and predictions are conducted using the exact covariance matrix, where the word 'training' comes to the name. If the approximation is used also for the predictions, the word training should drop out leading to FIC. In this case, FIC can be seen as a non-stationary covariance function on its own (Snelson, 2007). The partially independent conditional (PIC) sparse approximation is an extension of FIC (Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2007), and they are both treated here following Quiñonero-Candela and Rasmussen (2005).

The approximations are based on introducing an additional set of latent variables $\mathbf{u} = \{u_i\}_{i=1}^m$, called *inducing variables*. These correspond to a set of input locations $\mathbf{X}_u$, *inducing inputs*. The latent function prior is approximated as

$$p(\mathbf{f}\,|\mathbf{X}) \approx q(\mathbf{f}\,|\mathbf{X}, \mathbf{X}_u) = \int q(\mathbf{f}\,|\mathbf{X}, \mathbf{X}_u, \mathbf{u}) p(\mathbf{u}\,|\mathbf{X}_u) \mathrm{d}\mathbf{u}, \tag{4.3}$$

where $q(\mathbf{f}\,|\mathbf{X}, \mathbf{X}_u, \mathbf{u})$ is the inducing conditional. The above decomposition leads to the exact prior if the true conditional $\mathbf{f}\,|\,\mathbf{u} \sim \mathrm{N}(\mathbf{K}_{\mathrm{f,u}}\mathbf{K}_{\mathrm{u,u}}^{-1}\mathbf{u}, \mathbf{K}_{\mathrm{f,f}} - \mathbf{K}_{\mathrm{f,u}}\mathbf{K}_{\mathrm{u,u}}^{-1}\mathbf{K}_{\mathrm{u,f}})$ is used. However, in FIC framework the latent variables are assumed to be conditionally independent given $\mathbf{u}$, in which case the inducing conditional factorizes, $q(\mathbf{f}\,|\,\mathbf{u}) = \prod q_i(f_i|\,\mathbf{u})$. In PIC latent variables are set into blocks which are conditionally independent of each others, given $\mathbf{u}$, but the latent variables within a block have a multivariate normal distribution with the original covariance. The approximate conditionals of FIC and PIC can be summarized as

$$q(\mathbf{f}\,|\mathbf{X}, \mathbf{X}_u, \mathbf{u}) = \mathrm{N}\left(\mathbf{f}\,|\,\mathbf{K}_{\mathrm{f,u}}\mathbf{K}_{\mathrm{u,u}}^{-1}\mathbf{u}, \mathrm{mask}\left(\mathbf{K}_{\mathrm{f,f}} - \mathbf{K}_{\mathrm{f,u}}\mathbf{K}_{\mathrm{u,u}}^{-1}\mathbf{K}_{\mathrm{u,f}}|\mathbf{M}\right)\right), \tag{4.4}$$

where the function $\mathbf{\Lambda} = \mathrm{mask}\,(\cdot|\mathbf{M})$, with matrix $\mathbf{M}$ of ones and zeros, returns a matrix $\mathbf{\Lambda}$ of size $\mathbf{M}$ and elements $\mathbf{\Lambda}_{ij} = [\cdot]_{ij}$ if $\mathbf{M}_{ij} = 1$ and $\mathbf{\Lambda}_{ij} = 0$ otherwise. An approximation with $\mathbf{M} = \mathbf{I}$ corresponds to FIC and an approximation where $\mathbf{M}$ is block diagonal corresponds to PIC. The inducing inputs are given a zero-mean Gaussian prior $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}_{\mathrm{u,u}})$ so that the approximate prior over latent variables is

$$q(\mathbf{f}\,|\mathbf{X}, \mathbf{X}_u) = \mathrm{N}(\mathbf{f}\,|\mathbf{0}, \mathbf{K}_{\mathrm{f,u}}\mathbf{K}_{\mathrm{u,u}}^{-1}\mathbf{K}_{\mathrm{u,f}} + \mathbf{\Lambda}), \tag{4.5}$$

where the matrix $\mathbf{K}_{\mathrm{f,u}}\mathbf{K}_{\mathrm{u,u}}^{-1}\mathbf{K}_{\mathrm{u,f}}$ is of rank m and $\mathbf{\Lambda}$ is a rank n (block) diagonal matrix. The prior covariance can be seen as a non-stationary covariance function of its own where the inducing inputs $\mathbf{X}_u$ and the matrix $\mathbf{M}$ are free parameters similar to hyperparameters, which can be optimized alongside θ (Snelson and Ghahramani, 2006; Lawrence, 2007) (see also publication **II** and **IV**).

The computational savings are obtained by using the Woodbury-Sherman-Morrison lemma to invert the covariance matrix in (4.5) as

$$(\mathbf{K}_{\mathrm{f,u}}\mathbf{K}_{\mathrm{u,u}}^{-1}\mathbf{K}_{\mathrm{u,f}} + \mathbf{\Lambda})^{-1} = \mathbf{\Lambda}^{-1} - \mathbf{V}\mathbf{V}^{\mathrm{T}}, \tag{4.6}$$

where $\mathbf{V} = \mathbf{\Lambda}^{-1}\mathbf{K}_{\mathrm{f,u}}[\mathrm{chol}(\mathbf{K}_{\mathrm{u,u}} + \mathbf{K}_{\mathrm{u,f}}\mathbf{\Lambda}^{-1}\mathbf{K}_{\mathrm{f,u}})]^{-\mathrm{T}}$ (e.g. Harville, 1997). There is a similar result also for the determinant. With FIC the computational time is dominated

by the matrix multiplications, which need time $O(m^2 n)$. With PIC the cost depends also on the sizes of the blocks in $\mathbf{\Lambda}$. If the blocks were of equal size $b \times b$, the time for inversion of $\mathbf{\Lambda}$ would be $O(n/b \times b^3) = O(nb^2)$. With blocks at most the size of the number of inducing inputs, that is $b = m$, the computational cost in PIC and FIC are similar. Intuitively, PIC approaches FIC in the limit of a block size one and the exact GP in the limit of a block size n. A formal treatment of this is given by Snelson (2007).

4.3 Sparse additive models

In many practical situations, a GP prior with only one covariance function may be too restrictive since such a construction can model effectively only one phenomenon. For example, the latent function may vary rather smoothly across the whole area of interest, but at the same time it can have fast local variations. In this case, a more reasonable model would be

$$f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}), \tag{4.7}$$

where the latent value function is a sum of two functions, of which the other is slowly, and the other fast varying. We can place GP prior for both of the functions g and h with different covariance functions, which results in an additive prior

$$p(\mathbf{f}\,|\mathbf{X}) = \mathrm{N}(\mathbf{f}\,|\mathbf{0}, \mathbf{K}_{\mathrm{g,g}} + \mathbf{K}_{\mathrm{h,h}}). \tag{4.8}$$

An additive model could be formed also using specific covariance functions. For example, rational quadratic covariance function can be seen as a scale mixture of squared exponential covariance functions (Rasmussen and Williams, 2006), and could be useful for data that contain both local and global phenomena. However, using sparse approximations with the rational quadratic would prevent it from modeling local phenomena (see publication **II** and **IV**). The additive model (4.8) suits better for sparse GP formalism since it enables to combine FIC with CS covariance functions.

As discussed in section 4.2, FIC can be interpreted as a realization of a special kind of covariance function. By adding FIC with CS covariance function, for example (4.1), one can construct a sparse additive GP prior

$$\mathbf{f} \mid \mathbf{X}, \mathbf{X}_u, \theta \sim \mathrm{N}(\mathbf{0}, \mathbf{K}_{\mathrm{f,u}}\, \mathbf{K}_{\mathrm{u,u}}^{-1}\, \mathbf{K}_{\mathrm{u,f}} + \hat{\mathbf{\Lambda}}). \tag{4.9}$$

This prior will be referred as CS+FIC. Here, the matrix $\hat{\mathbf{\Lambda}} = \mathbf{\Lambda} + \mathbf{K}_{\mathrm{f,f}}^{\mathrm{CS}}$ is sparse with the same sparsity structure as in $\mathbf{K}_{\mathrm{f,f}}^{\mathrm{CS}}$ and, thus, it is fast to use in computations and cheap to store. CS+FIC can be extended to have more than one component. However, it should be remembered that FIC works well only for long length-scale phenomena and the computational benefits of CS functions are lost if their length-scale gets too large (see publication **IV**). For this reason the CS+FIC should be constructed so that possible long length-scale phenomena are handled with FIC part and the short length-scale phenomena with CS part. The implementation of the CS+FIC model follows closely the implementation of FIC and PIC (for details see publications **II** and **IV**).

4.4 Other means to speed up computation

The sparse approximation is a somewhat misleading term since the matrices involved are not necessarily sparse but low rank. The term originates to the machine learning literature where the early approximations were based on a representative subset of data on which the most time consuming matrix operations were performed (e.g. Smola and Bartlett, 2001; Csató and Opper, 2002; Seeger et al., 2003; Cornford et al., 2005). The difference between FIC and PIC to the earlier approximations is that the locations of the inducing inputs are not limited to a subset of training data but they are considered as free variables that can be optimized alongside the hyperparameters. Later on Titsias (2009) improved the optimization of the inducing inputs and hyperparameters by considering it as maximizing the variational bound between the full and sparse GP. Other recent developments include, among others, the transformation of the inducing inputs into more representative feature space (Lazaro-Gredilla and Figueiras-Vidal, 2009; Qi et al., 2010), and sparse multi-output GPs (Alvarez and Lawrence, 2009; Alvarez et al., 2010).

From the matrix algebra point of view the key step in sparse approximations is to represent the full rank covariance matrix with $\mathbf{V}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{V} \in \Re^{n\times m}$. The (block) diagonal $\mathbf{\Lambda}$ works as a correction term. In spatial statistics similar approaches are called low-rank models (Diggle and Ribeiro, 2007). The low rank models assume that the Gaussian field $S \in \Re^n$ is a linear combination of m basis functions, $S(\mathbf{x}) = \sum_{i=1}^{m} A_i f_i(\mathbf{x})$, where $A = [A_1, ..., A_m]^{\mathrm{T}}$ is a Gaussian random variable in $\Re^m$. The covariance matrix of the field S, at locations $\mathbf{X}$, is then $\mathbf{F}\,\mathrm{Cov}(A)\mathbf{F}^{\mathrm{T}}$, where F is a matrix with elements $\mathbf{F}_{ij} = f_j(\mathbf{x}_i)$. This is a low rank matrix similar to the sparse approximations. The type of an approximation depends on the basis functions used. Familiar examples include spectral representation (Diggle and Ribeiro, 2007; Paciorek, 2007) and splines (Wood, 2003). The difference between sparse approximations and low-rank models is more fundamental than just replacing $\mathrm{Cov}(A)$ by $\mathbf{K}_{\mathrm{u,u}}^{-1}$. The approximation $\mathbf{K}_{\mathrm{f,u}}\mathbf{K}_{\mathrm{u,u}}^{-1}\mathbf{K}_{\mathrm{u,f}}$ is induced from minimizing $\mathrm{KL}(p(\mathbf{f},\mathbf{u})R(\mathbf{y}\,|\,\mathbf{u})||p(\mathbf{f},\mathbf{u})p(\mathbf{y}\,|\,\mathbf{f}))$ over all distributions of the form $p(\mathbf{f},\mathbf{u})R(\mathbf{y}\,|\,\mathbf{u})$, where $R(\mathbf{y}\,|\,\mathbf{u})$ is positive and dependent on $\mathbf{u}$ only (Csató, 2002; Seeger et al., 2003; Titsias, 2009). This justification was used also by Banerjee et al. (2008) in their Gaussian predictive process model. See also (Trecate et al., 1999) for similar considerations. The diagonal term $\mathbf{\Lambda}$, for example, in FIC results from minimising the KL-divergence other way around (Qi et al., 2010).

In spatial statistics the computational cost of GP has commonly been reduced by restricting it to a Gaussian Markov random field (GMRF) (Rue and Held, 2006; Best et al., 2005; Elliot et al., 2001; Rue et al., 2009). GMRF is specified by assuming conditional independence (Markov property) between a set of locations. That is, the full conditional of a latent variable depends only on a small number of other latent variables, called the neighbors. The conditional independence assumption leads to fast computations since the resulting precision matrix is sparse. The advantage of GPs over GMRF models is, however, the flexibility in choosing the covariance function. The GMRFs have also been used to approximate the Gaussian random field (Rue and Tjelmeland, 2002; Rue and Held, 2006).

The above mentioned methods try to speed up the computations given the large co-

variance matrix. Another solution is to take advantage of a possible structure behind the phenomenon. Examples of this are Kalman filtering and smoothing, which are special cases of GP regression, and which scale linearly in a number of time points and cubicly in state variables (e.g. Grewal and Andrews, 2001). If the number of state variables is small, the Kalman smoother gives computationally efficient solution to the GP regression problem with a specific covariance function. Similarly in spatio-temporal models the spatial and temporal components can be separated and the resulting structured covariance matrix handled efficiently (e.g. Finkenstädt et al., 2007). These approaches, however, will not be applicable if we do not know the underlying structure of the phenomena or can not formulate it in a way that leads to an efficient algorithm.

Chapter 5

Summary of the publications

This chapter provides a short summary of the attached publications. The treatment will be kept compact and discussion is omitted at this point. The aim is to give a glance at the problems considered but one should read the publications for a detailed treatment.

5.1 Publication I

Spatial epidemiology concerns both describing and understanding the spatial variation in the disease risk in geographically referenced health data. One of the main classes of spatial epidemiological studies is disease mapping, where the aim is to describe the overall disease distribution on a map and, for example, highlight areas of elevated or lowered mortality or morbidity risk (e.g. Lawson, 2001; Richardson, 2003; Elliot et al., 2001). In this work, a point-referenced health-care data are aggregated into a lattice of grid cells. The mortality in a cell is modeled as a Poisson distribution, whose mean is a product of a standardized expected number of deaths, e_i and a relative risk $\mu = \exp(f(\mathbf{x}))$, where $\mathbf{x}$ is the co-ordinate of the cell. The expected number of deaths is evaluated using age, gender and scholarly degree standardization, and the logarithm of the relative risk is given a Gaussian process prior. The model follows the general approach discussed by Best et al. (2005) and can be summarized as

$$\mathbf{y} \sim \prod_{i=1}^{n} \text{Poisson}(\exp(f_i)e_i) \tag{5.1}$$

$$f(\mathbf{x})|\theta \sim \mathcal{GP}\left(0, k(\mathbf{x}, \mathbf{x}'\,|\theta)\right) \tag{5.2}$$

$$\theta \sim \text{half-Student-}t(\nu, \sigma_t^2) \tag{5.3}$$

Here σ_t^2 is the scale and ν the degrees of freedom of the half-Student-t distribution (Gelman, 2006). The inference is conducted via MCMC by alternating the HMC to sample from the conditional distributions of latent variables given the covariance function parameters and the covariance function parameters given the latent variables. The mixing of latent variable sampling is improved with a transformation taking into account their approximate conditional posterior precision (Christensen et al., 2006).

The main contributions of this publication are the description how to implement the HMC sampling with latent variable transformation for FIC, the first use of FIC with other than Gaussian observation model, and experiments on the suitability of the approximation to spatial modeling. FIC is shown to perform similarly to the full GP whenever the posterior of the length-scale is long enough compared with the spacing of the inducing inputs. The performance of FIC gradually decreases as the number of the inducing inputs is reduced since then the inducing inputs get more sparsely located. FIC is shown to be considerably faster than full GP with a small number of inducing inputs. The results show the potential of FIC in spatial modelling problems and encouraged further studies on the issue.

5.2 Publication II

This publication introduces the CS+FIC model for sparse additive GP regression. The observation model is Gaussian noise and hyperprior half-Student-t distribution, which results in the overall model

$$\mathbf{y}\,|\sigma^2 \sim \mathrm{N}(\mathbf{f}, \sigma^2\mathbf{I}), \tag{5.4}$$

$$f(\mathbf{x})|\theta_h, \theta_g \sim \mathcal{GP}\left(0, k_g(\mathbf{x}, \mathbf{x}'\,|\theta_g) + k_h(\mathbf{x}, \mathbf{x}'\,|\theta_h)\right), \tag{5.5}$$

$$\theta_h, \theta_g, \sigma^2 \sim \text{half-Student-}t(\nu, \sigma_t^2). \tag{5.6}$$

The main contributions of the article are the description how CS covariance functions can effectively be used in GP regression when utilizing the gradient of the log likelihood with respect to the hyperparameters, proposal for the CS+FIC model, and analyzis of the sparse GPs with data sets that contain additive long and short length-scale phenomena. It is shown that CS+FIC outperforms FIC and PIC if data contain short lenght-scale phenomena and is computationally faster than full GP. The computational techniques with CS functions rely, in addition to standard sparse matrix routines, on the sparse inverse algorithm proposed by (Takahashi et al., 1973), which was introduced to GP regression in this work. Key findings about the sparse GPs are that FIC is global by its nature and does not work for short length-scale phenomena. PIC models rather well also short length-scales, but suffers from the discontinuities in its correlation structure. CS+FIC combines the good global properties of FIC and the good local properties of CS covariance functions. The publication demonstrates the usefulness of CS+FIC and encourages studies on its implementation for non-Gaussian observation models.

5.3 Publication III

A commonly used observation model in the GP regression is the Gaussian distribution. This is convenient since the inference is analytically tractable up to the covariance function parameters. However, a known limitation with the Gaussian observation model is its non-robustness, due which outlying observations may significantly reduce the accuracy of the inference. A formal definition of robustness is given, for example, in terms

of an outlier-prone observation model. The observation model is outlier-prone of an order n, if $p(f|y_1,...,y_{n+1}) \rightarrow p(f|y_1,...,y_n)$ as $y_{n+1} \rightarrow \infty$ (O'Hagan, 1979; West, 1984). That is, the effect of a single conflicting observation on the posterior becomes asymptotically negligible as the observation approaches infinity. This contrasts heavily with the Gaussian observation model where each observation influences the posterior no matter how far it is from the others. A well-known robust observation model is the Student-t distribution, which is used also in this work. The model considered is

$$\mathbf{y}\,|\nu,\sigma_t \sim \prod_{i=1}^{n} \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma_t}\left(1+\frac{(y_i-f_i)^2}{\nu\sigma_t^2}\right)^{-(\nu+1)/2}, \tag{5.7}$$

$$f(\mathbf{x})|\theta \sim \mathcal{GP}\left(0, k(\mathbf{x},\mathbf{x}'\,|\theta)\right), \tag{5.8}$$

$$\theta,\nu,\sigma_t \sim \text{half-Student-}t(\nu,\sigma_t^2). \tag{5.9}$$

The idea of robust regression is not new. Outlier rejection was described already by De Finetti (1961) and theoretical results were given by Dawid (1973), and O'Hagan (1979). Student-t observation model with linear regression was studied already by West (1984) and Geweke (1993), and Neal (1997) introduced it for GP regression. Other robust observation models include, for example, mixtures of Gaussians, Laplace distribution and input dependent observation models (Naish-Guzman and Holden, 2008; Kuss, 2006; Goldberg et al., 1998; Stegle et al., 2008).

A common computational approach has been to use the scale-mixture representation of the Student-t distribution, which enables Gibbs sampling (Geweke, 1993; Neal, 1997), and a factorized variational approximation (VB) (Tipping and Lawrence, 2005; Kuss, 2006). The main contributions in this publication are the following. It is shown, how Student-t observation model can be inferred in an efficient and robust way using Laplace approximation. The properties of the model are discussed and results are compared to the earlier approaches which utilize the scale mixture representation. It is shown, that Laplace approximation works quickly and accurately compared with full MCMC and a factorized VB. It is also shown that the Laplace method approximates the posterior covariance somewhat better than the factorized VB, which underestimates the variance because of its factorized construction.

5.4 Publication IV

This work considers the same disease mapping problem as publication **I** (described by the equations (5.1)–(5.3)). The article combines the sparse GPs discussed in chapter 4 with the approximate inference schemes in chapter 3 and provides an extensive description how they should be used to perform fast inference with Gaussian process disease mapping model.

The main contributions of the paper are the following. The FIC, PIC and CS+FIC sparse approximations are analyzed in detail. Their correlation structures and non-stationary properties are analyzed and visualized for two dimensional input space. Also their practical limits, compared with the smallest possible length-scale, are sought. It is shown that, if we want FIC to approximate the full covariance well, the inducing inputs

should be placed in a regular grid and the length-scale should be greater than the distance between adjacent inducing inputs. Also the discontinuous correlation structure of PIC is demonstrated and ways to alleviate this property suggested. The implementation of Laplace approximation and EP for CS+FIC model is explained. The article contains also comparison between the approximate inference methods. The Laplace approximation and EP at $\hat{\theta}$ as well as after the grid, Monte Carlo or CCD integration are compared with the full MCMC solution. It is shown that the Gaussian approximation to $p(\mathbf{f}\,|\mathcal{D},\hat{\theta})$ resembles very closely the true conditional posterior obtained by extensive MCMC. Also the marginal posterior $p(\mathbf{f}\,|\mathcal{D})$ is practically identical to the MCMC solution. The article provides also a short comparison between the sparse GP models and conditional autoregressive (CAR) model implemented with INLA (Rue et al., 2009). Both models give similar results in overall but there are some differences in the latent variable posterior. CAR model gives certain cells very awkward estimates, since the expectation of the latent variables ranges from -20 to 0.7 and the variance from 4×10^{-4} to 3×10^{4}, whereas with GP the expectations are between -0.9 and 0.7, and the variances between 0.004 and 0.23 (here it should be remembered that the relative risk is $\exp(f)$).

In summary, this publications presents practical implementation of GP models for large spatial data. It also demonstrates the advantage of GPs over CAR models with areally sparse spatial data.

5.5 Publication V

The common way to set up Gaussian process classification model is the following. Consider binary observations, $y_i \in \{-1,+1\}, i = 1,...,n$, appointed to inputs $\mathbf{X} = \{\mathbf{x}\}_{i=1}^{n}$. The observations are considered to be drawn from the Bernoulli distribution whose success probability is related to the latent function $f(\mathbf{x}) : \Re^d \rightarrow \Re$ that is mapped to a unit interval by a sigmoid transformation. The usual transformation is the probit function $p(y_i = 1|\mathbf{x}_i) = \Phi(f(\mathbf{x}_i))$, where Φ denotes the cumulative probability function of the standard Normal density. The latent function is then given a zero mean GP prior which leads to the model

$$\mathbf{y} \sim \prod_{i=1}^{n} \Phi(f(\mathbf{x}_i)y_i) \tag{5.10}$$

$$f(\mathbf{x})|\theta \sim \mathcal{GP}\left(0, k(\mathbf{x},\mathbf{x}'\,|\theta)\right), \tag{5.11}$$

$$\theta \sim \text{half-Student-}t(\nu,\sigma_t^2). \tag{5.12}$$

This problem is discussed in detail by Rasmussen and Williams (2006). Nickisch and Rasmussen (2008) provide comprehensive analysis of different approximate inference methods proposed in the literature and show that the best compromise between accuracy and speed is the EP algorithm.

What has not been studied is how to implement EP for CS covariance functions. This topic is touched shortly in the publication **IV** but in this work it is given a detailed description and the needed sparse matrix routines are analyzed in length. Also four of the Wendland's piecewise polynomials (Wendland, 2005) are analyzed from

machine learning point of view. It is shown that they all suffer from an increasing input dimension but may speed up the inference considerably in small dimensional spaces. Since EP algorithm can be applied to other problems than classification as well, the techniques and considerations in the article are not limited to GP classification. The main contributions of the publication are the analysis of the Wendland's functions and the sparse matrix techniques needed for EP to work fast with CS covariance functions. From the results it can be concluded that CS covariance functions, with proper implementation, provide an efficient alternative to full GP and sparse approximations for datasets with low input dimension.

5.6 Publication VI

This last publication differs from the rest in that its main contribution is not in the methodological development but in a real world application. The problem at hand is a disease mapping problem where the aim is to study alcohol related mortality in Finland during the years 2001-2005. In high-income countries, alcohol consumption is the second most important determinant of loss of healthy life years due to illness and death (World Health Organization (WHO), 2009). Alcohol related deaths have an important role in the mortality of working-age population in Finland as well and detailed knowledge on their reasons is valuable. This study helps the health care authorities by pointing out the problematic areas where preventive actions should be taken. The publication is a translation of the original Finnish article (Vanhatalo et al., 2010).

In contrast to the disease mapping model of publications **I** and **IV** the observation model is Negative-Binomial. The Negative-Binomial distribution is a robust version of the Poisson distribution similarly as Student-t distribution can be considered a robustified Gaussian distribution (Gelman et al., 2004). The model considered is

$$\mathbf{y}\,|r \sim \prod_{i=1}^{n} \frac{\Gamma(r+y_i)}{y_i!\Gamma(r)} \left(\frac{r}{r+\mu_i}\right)^r \left(\frac{\mu_i}{r+\mu_i}\right)^{y_i} \tag{5.13}$$

$$f(\mathbf{x})|\theta \sim \mathcal{GP}\left(0, k(\mathbf{x}, \mathbf{x}'\,|\theta_g)\right), \tag{5.14}$$

$$\theta \sim \text{half-Student-}t(\nu, \sigma_t^2), \tag{5.15}$$

where $\mu_i = \exp(f(\mathbf{x}_i))$. The main contributions of the publication are the following. With the techniques described in this dissertation the spatial variation on alcohol related deaths could be studied in 5 kilometers accuracy, which is a big advancement compared with the previous studies at municipality level (e.g. Mäkelä et al., 2001). The publication illustrates the use of the Google Maps and Google Earth mapping programs for analyzing the results. The key findings are that alcohol related deaths are relatively less common in the south-west coast of Ostrobothnia, and more common in Eastern and South-eastern Finland. Research also highlights the difference between population centers and surrounding areas. Risk of dying from alcohol-based illness is generally higher in densely than in sparsely populated areas even after standardizing the incidence rates with the population density. The GP models were compared with the CAR models implemented with INLA and GP was found to work better (detailed comparison is not presented though).

Chapter 6

Discussion

Broadly thinking Gaussian processes have been subject for intensive research for decades and many disciplines have contributed to their study. They have not, necessarily, been called Gaussian processes but, for example, spatial statistics, signal processing and filtering literature have their own naming conventions. The point of view taken in this study comes mainly from the machine learning literature where the subject has flourished to a hot topic since late 1990's. Readers who have waded up to this point may have noticed that the material is largely a bag of tricks collected from various sources. These tricks are not separate, though, but the bag could be labelled as: practical tools for analyzing Gaussian process models. Some of the tools are very old, such as the Laplace approximation based on Taylor expansion, which was used for the first time in Bayesian analysis by Laplace himself (see, for example, the historical notes in Gelman et al., 2004). Some of the tricks are rather new, such as the CCD integration (Rue et al., 2009) or EP algorithm (Minka, 2001), and some are even proposed by the author himself, such as the EP algorithm for CS covariance functions in publications **IV** and **V**.

This brings us to the definition of practical tools. I would say that practical computational tools are such that they produce useful, interpretable results in a sensible time. This does not mean that the results need to be exact, but they have to approximate well enough the essential aspects. Thus, practical tools are always data, model, and problem dependent. For example MCMC methods are many times praised for their asymptotic properties and seemingly easy implementation. Algorithms, such as Metropolis-Hastings, are easy to write for almost any model. The problem, however, is that as data set grows they may not give reliable results in a finite time. With GP models this problem is faced very severely. For example, the disease mapping problems in publications **IV** and **VI** with over 10 000 data points would be impossible to solve with standard office PC's using MCMC. The convergence rate and mixing of the sample chain would be just too slow. Every approximation is, however, a double-edged sword. As we know that the results presented before our eyes are approximations we should never take them as granted. The results should always be examined to learn their limitations and possible faults. A good example of this is the classification problem where neither Laplace approximation nor EP give even close to exact result for the

latent variable posterior (Kuss and Rasmussen, 2005; Nickisch and Rasmussen, 2008). However, if we are interested in the classification performance, EP works almost as well as MCMC.

Stating the general model family as in the equations (2.1) – (2.3) gives a hint, how the models can be programmed effectively. This subject is not treated here but it is worth mentioning that the models can be implemented in a rather modular way with object oriented programming. Every piece of the model, likelihood, GP prior, hyperpriors, covariance functions etc. should be written as an object which returns its essential characteristics in a standard form. The inference methods can then be implemented in a fairly general way for evaluating needed summaries, such as log marginal likelihood or predictive distribution. This kind of implementation is used in MATLAB toolbox GPstuff (`http://www.lce.hut.fi/research/mm/gpstuff/`) which collects all the models and methods discussed in this work.

Handling such a general class of models raises the question of model misspecification. The model misfit related to the observation model is one source of problems. For example, if the data contained outliers or observations with clearly higher variance than what the Gaussian or Poisson observation model predicts, the posterior of the latent function would be highly compromised. For this reason, it is important to be able to test easily robust alternatives for traditional models as well. Such as Student-t or Negative-Binomial distribution. Even though the GP prior is very flexible and few things, namely the mean and covariance, need to be fixed in its construction, it still contains rather heavy assumptions in it. For example, a Gaussian process associated with the squared exponential covariance function is indefinitely mean square differentiable. This is a very strong assumption on the smoothness of the latent function. In fact it is rather peculiar how little attention other covariance functions have gained in machine learning literature. One of the reasons may be that often machine learning problems take place in high dimensional input spaces where data are more likely to lie sparsely. In this case we are not able to infer fast varying phenomena, since the sparse data provide no information on them, and smooth solutions are the ones that get higher posterior probability. The topic is covered in detail by [7] whose results suggest that under certain assumptions on the input vector distribution the nonlinear methods that rely on kernel matrices may be behaving like their linear counterparts. However, as shown in publications **I**, **II**, **IV**, and **VI** the covariance function does influence the results at least in low dimensional problems. Even though compactly supported covariance functions are known to exist, they seem to be rather little used outside spatial statistics. As discussed in the publications **II** and **V**, high dimensional input spaces seem to be problematic for the CS covariance functions, which may set them into disfavor (at least few reviewers of the publications argued this way). However, this is a problem only if GP is used as a black box that is expected to work for any problem without modifications, but not if GPs are tailored for each problem individually. The latter approach is more correct, I would claim, for which reason CS covariance functions should be exploited more whenever the phenomenon is likely to contain short length-scale correlations.

In the spatial statistics literature, Gaussian processes defined by covariance function have earlier been criticized for giving too smooth results and CAR models suggested instead (Best et al., 2005). The results in the publications **IV** and **VI**, however, show that GP may outperform CAR model. Presumably, the reason for differing results

is that the previously used covariance functions have favored too smooth functions and that additive covariance structures are seldom used. GP models with only one covariance function tend to give smooth results, but with several components they can adapt to very fast changes as well. Also, GP's advantage over CAR model comes clearly apparent if the spatial data is areally sparse (see publication **IV**). The sparse GP's discussed here start to be comparable to CAR models also in speed, as well as other sparse methods proposed in the literature (e.g. Cornford et al., 2005; Banerjee et al., 2008).

In statistical literature, inference in the hyperparameters is a natural concern but in machine learning literature they are left in less attention. An indicator of this is the usual approach to maximize the marginal likelihood which implies uniform prior for the hyperparameters. In this work, the hyperparameters are always given an explicit prior. The choice has usually been half Student-t distribution which works as a weakly informative prior (Gelman, 2006). There are a few reasons for explicitly defining a hyperprior. In spatial statistics literature it is well known that the length-scale and magnitude are under identifiable and the proportion σ^2/l is more important to the predictive performance than their individual values (Diggle et al., 1998; Zhang, 2004; Diggle and Ribeiro, 2007). These results are shown for Mátern class of covariance functions but according to the experiments they seem to apply for Wendland's piecewise polynomials as well. With them, the property can be taken advantage of since by giving more weight to short length-scales one favors sparser covariance matrices that are faster in computations. Other advantage is that priors make the inference problem easier by narrowing the posterior and making the hyperparameters more identifiable. This is useful especially for MCMC methods but optimization and other integration approximations gain from the hyperpriors as well. These two reasons are rather practical. More fundamental reason is that in Bayesian statistics leaving prior undefined (meaning uniform prior) is a prior statement as well, and sometimes it may be really awkward (for example, uniform prior works very badly for the parameters of Student-t distribution). Thus, it is better to spend some time thinking what the prior actually says.

What is left out from this dissertation are the improvements to the Gaussian approximations. The shape of the marginal posterior of a latent variable could be estimated more accurately with techniques described by, for example, Rue et al. (2009), Paquet et al. (2009), and Cseke and Heskes (2010) (see also Tierney and Kadane, 1986). Improvements always compromise the computational speed but are essential for more reliable results, for which reason the subject provides an important future research direction. The results here suggest that the inference is not sensitive to the marginalization over the hyperparameters and often the MAP estimate works fine. However, a comprehensive study on the subject is still needed to assess the limits of MAP estimate in GP models. A third future research topic, suggested by the results presented here, is the CS covariance functions. They are not yet extensively studied in spatial statistics nor machine learning literature and, for example, this work has just scratched their surface.

Bibliography

Alvarez, M. and Lawrence, N. D. (2009). Sparse convolved Gaussian processes for multi-output regression. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 57–64.

Alvarez, M. A., Luengo, D., Titsias, M. K., and Lawrence, N. D. (2010). Efficient multioutput Gaussian processes through variational inducing kernels. *JMLR Workshop and conference proceedings*, 9:25–32.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society B*, 70(4):825–848.

Bayes, T. (1763). Essay towards solving a problem in the doctrine of chances. *The Philosophical Transactions*, 53:370–418. Reprinted 1958 in *Biometrica*, 45(3/4):296-315.

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.

Buhmann, M. D. (2001). A new class of radial basis functions with compact support. *Mathematics of Computation*, 70(233):307–318.

Christensen, O. F., Roberts, G. O., and Sköld, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15:1–17.

Cornford, D., Csató, L., and Opper, M. (2005). Sequential, Bayesian geostatistics: A principled method for large data sets. *Geographical Analysis*, 37:183–199.

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

Csató, L. and Opper, M. (2002). Sparse online Gaussian processes. *Neural Computation*, 14(3):641–669.

Csató, L. (2002). *Gaussian Processes - Iterative Sparse Approximations*. PhD thesis, Aston University.

Cseke, B. and Heskes, T. (2010). Improving posterior marginal approximations in latent Gaussian models. *JMLR Workshop and Conference Proceedings*, 9:121–128.

Dale, A. I. (1999). *A History of Inverse Probability: From Thomas Bayes to Karl Pearson (Sources and Studies in the History of Mathematics and Physical Sciences)*. Springer-Verlag.

Davis, T. A. (2006). *Direct Methods for Sparse Linear Systems*. SIAM.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika*, 60(3):664–667.

De Finetti, B. (1961). The Bayesian approach to the rejection of outliers. In *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 199–210. University of California Press.

Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer Science+Business Media, LLC.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 47(3):299–350.

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

Elliot, P., Wakefield, J., Best, N., and Briggs, D., editors (2001). *Spatial Epidemiology Methods and Applications*. Oxford University Press.

Finkenstädt, B., Held, L., and Isham, V. (2007). *Statistical Methods for Spatio-Temporal Systems*. Chapman & Hall/CRC.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Gaspari, G. and Cohn, S. (1999). Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554):723–757.

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P. (2010). *Handbook of Spatial Statistics*. CRC Press.

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. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics*, 8:519–540.

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

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.

Gneiting, T. (1999). Correlation functions for atmospheric data analysis. *Quarterly Journal of the Royal Meteorological Society*, 125:2449–2464.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis*, 83:493–508.

Goel, P. K. and Degroot, M. H. (1981). Information about hyperparamters in hierarchical models. *Journal of the American Statistical Association*, 76(373):140–147.

Goldberg, P. W., Williams, C. K., and Bishop, C. M. (1998). Regression with input-dependent noise: A Gaussian process treatment. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*. MIT Press, Cambridge, MA.

Grewal, M. S. and Andrews, A. P. (2001). *Kalman Filtering: Theory and Practice Using Matlab*. Wiley Interscience, second edition.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555.

Kuss, M. (2006). *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning*. PhD thesis, Technische Universität Darmstadt.

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

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les évènemens. *Mémoires de Mathématique et de Physique, Presentés à l'Académie Royale des Sciences, par divers Savans & lûs dans ses Assemplées, Tome Sixiéme*, pages 621–656. English translation by S. M. Stiegler: Laplace, P. S. (1986). Memoir on the Probability of the Causes of Events. *Statistical Science*, 1(3):364-478.

Lawrence, N. (2007). Learning for larger datasets with the Gaussian process latent variable model. In Meila, M. and Shen, X., editors, *Proceedings of the Eleventh International Workshop on Artificial Intelligence and Statistics*. Omnipress.

Lawson, A. B. (2001). *Statistical Methods in Spatial Epidemology*. John Wiley & Sons, Ltd.

Lazaro-Gredilla, M. and Figueiras-Vidal, A. (2009). Inter-domain Gaussian processes for sparse inference using inducing features. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1087–1095.

Martino, S. (2007). *Approximate Bayesian Inference for Latent Gaussian Models*. PhD thesis, Norwegian University of Science and Technology.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, 5(3):439–468.

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

Moreaux, G. (2008). Compactly supported radial covariance functions. *Journal of Geodecy*, 82(7):431–443.

Mäkelä, P., Ripatti, S., and Valkonen, T. (2001). Alue-erot miesten alkoholikuolleisuudessa. *Suomen Lääkärilehti*, 56(23):2513–2519.

Naish-Guzman, A. and Holden, S. (2008). Robust regression with twinned Gaussian processes. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 1065–1072. MIT Press, Cambridge, MA.

Neal, R. (1998). Regression and classification using Gaussian process priors. In Bernardo, J. M., Berger, J. O., David, A. P., and Smith, A. P. M., editors, *Bayesian Statistics 6*, pages 475–501. Oxford University Press.

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

Neal, R. M. (1997). Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. Technical Report 9702, Dept. of statistics and Dept. of Computer Science, University of Toronto.

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.

O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society. Series B.*, 41(3):358–367.

O'Hagan, A. (2004). Dicing with the unknown. *Significance*, 1:132–133.

Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792.

Paciorek, C. J. (2007). Computational techniques for spatial logistic regression with large datasets. *Computational Statistics and Data Analysis*, 51:3631–3653.

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

Qi, A., Abdel-Gawad, A., and Minka, T. (2010). Sparse-posterior Gaussian processes for general likelihoods. In Grünwald, P. and Spirtes, P., editors, *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 450–457. AUAI Press.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(3):1939–1959.

Rasmussen, C. E. (1996). *Evaluations of Gaussian Processes and Other Methods for Non-linear Regression*. PhD thesis, University of Toronto.

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

Richardson, S. (2003). Spatial models in epidemiological applications. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 237–259. Oxford University Press.

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

Rue, H. and Held, L. (2006). *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 B*, 71(2):1–35.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29:31–49.

Sanchez, S. M. and Sanchez, P. J. (2005). Very large fractional factorials and central composite designs. *ACM Transactions on Modeling and Computer Simulation*, 15:362–377.

Sansò, F. and Schuh, W.-D. (1987). Finite covariance functions. *Journal of Geodesy*, 61(4):331–347.

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

Seeger, M., Williams, C. K. I., and Lawrence, N. (2003). Fast forward selection to speed up sparse Gaussian process regression. In Bishop, C. M. and Frey, B. J., editors, *Ninth International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics.

Smola, A. and Bartlett, P. (2001). Sparse greedy Gaussian process regression. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 619–625. The MIT Press.

Snelson, E. (2007). *Flexible and Efficient Gaussian Process Models for Machine Learning*. PhD thesis, University College London.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian process using pseudo-inputs. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*. The MIT Press.

Snelson, E. and Ghahramani, Z. (2007). Local and global sparse Gaussian process approximations. In Meila, M. and Shen, X., editors, *Artificial Intelligence and Statistics 11*. Omnipress.

Stegle, O., Fallert, S. V., MacKay, D. J. C., and ren Brage, S. (2008). Gaussian process robust regression for noisy heart rate data. *Biomedical Engineering, IEEE Transactions on*, 55(9):2143–2151.

Storkey, A. (1999). *Efficient Covariance Matrix Methods for Bayesian Gaussian Processes and Hopfield Neural Networks*. PhD thesis, University of London.

Takahashi, K., Fagan, J., and Chen, M.-S. (1973). Formation of a sparse bus impedance matrix and its application to short circuit study. In *Power Industry Computer Application Conference Proceedings*. IEEE Power Engineering Society.

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.

Tipping, M. E. and Lawrence, N. D. (2005). Variational inference for Student-t models: Robust bayesian interpolation and generalised component analysis. *Neurocomputing*, 69:123–141.

Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. *JMLR Workshop and Conference Proceedings*, 5:567–574.

Trecate, G. F., Williams, C. K. I., and Opper, M. (1999). Finite-dimensional approximation of Gaussian processes. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 218–224, Cambridge, MA, USA. MIT Press.

Vanhatalo, J., Mäkelä, P., and Vehtari, A. (2010). Alkoholikuolleisuuden alueelliset erot suomessa 2000-luvun alussa. *Yhteiskuntapolitiikka*, 75(3):265–273.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Computational Mathematics*, 4(1):389–396.

Wendland, H. (2005). *Scattered Data Approximation*. Cambridge University Press.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society. Series B.*, 46(3):431–439.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press.

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 Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8*, pages 514–520.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(1):95–114.

World Health Organization (WHO) (2009). Global health risks - mortality and burden of disease attributable to selected major risks. Technical report, World Health organization (WHO).

Wu, Z. (1995). Compactly supported positive definite radial functions. *Advances in Computational Mathematics*, 4(1):283–292.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.