Aalto University
School of Science and Technology
Faculty of Electronics, Communication and Automation

Ville Pietiläinen

Approximations for Integration Over the Hyper-parameters in Gaussian Processes

In partial fulfillment of the requirements for the degree of Master of Science, Espoo January 21, 2010.

Supervisor: Jouko Lampinen

Instructor: Aki Vehtari

Aalto-yliopisto
Teknillinen Korkeakoulu
Elektroniikan, Tietoliikenteen ja Automaation Tiedekunta
Bioinformaatioteknologian koulutusohjelma **DIPLOMITYÖN TIIVISTELMÄ**

Tekijä: Ville Pietiläinen
Otsikko: Gaussisten prosessien hyperparametrien yli integroinnin aproksimointi
Päivämäärä: 21. tammikuuta 2010 **Sivumäärä:** 47

Tiedekunta: Elektroniikan, Telekommunikaation ja Automaation tiedekunta
Professuuri: S-114, Laskennallinen tekniikka

Työn valvoja: Prof. Jouko Lampinen
Työn ohjaaja: Dos., TkT Aki Vehtari

Tässä työssä tutkitaan kolmea numeerista menetelmää, jotka approksimoivat integraalia Gaussisten prosessien hyperparametrien posteriorijakauman yli. Tämän integraalin analyyttinen käsittely on usein mahdotonta. Työssä tutkitaan approksimaatioiden ominaisuuksia, ja niiden suorituskykyä verrataan keskenään sekä piste-estimaattia käyttävään menetelmään.

Perinteisesti integraali hyperparametrien posteriorin yli on laskettu käyttäen Markovin ketju Monte Carlo (MCMC) -menetelmiä. MCMC-menetelmät kuitenkin kärsivät Gaussisten prosessien laskennan raskaudesta, sillä Gaussisten prosessien kompleksisuus kasvaa käytettävän datan kasvaessa. Yksi vaihtoehtoinen menetelmä on käyttää piste-estimaattia hyperparametrien posteriorijakauman yli integroimisen sijaan. Tämä on laskennallisesti nopea tapa, mutta se jättää huomioimatta hyperparametreihin liittyvän epävarmuuden.

Tässä työssä esitetyt approksimaatiot pyrkivät ottamaan huomioon hyperparametrien epävarmuuden piste-estimaattia paremmin, mutta kuitenkin pysymään laskennallisesti kevyempinä kuin MCMC-menetelmät. Työn tulokset osoittavat että hyperparametrien posteriorin yli integroiminen on hyödyllistä tietyissä olosuhteissa. Lisäksi näytetään että piste-estimaatti tuottaa integrointimenetelmien kanssa yhtä tarkkoja tuloksia joissain tilanteissa. Käytettävän datan määrä ja mallien käyttötarkoitus vaikuttavat integrointimenetelmien tarpeellisuuteen, ja työssä tarkastellaan näitä olosuhteita.

Avainsanat: Bayesilainen päättely, Gaussiset prosessit, hyperparametrit, integroinnin aproksimointi

Aalto University
School of Science and Technology
Faculty of Electronics, Communications and Automation
Degree Programme of Bioinformation Technology

ABSTRACT OF THE MASTER'S THESIS

Author:	Ville Pietiläinen
Title:	Approximations for Integration over the Hyperparameters in Gaussian Processes
Date:	January 21, 2010 **Number of pages:** 47
Faculty:	Faculty of Electronics, Communications and Automation
Professorship:	S-114, Computational Engineering
Supervisor:	Prof. Jouko Lampinen
Instructor:	Doc., Dr.Tech. Aki Vehtari

This thesis examines three numerical approximations for the analytically intractable integral over the posterior distribution of the hyperparameters in Gaussian processes. The properties of the approximations are studied, and their performance is compared to each other and to a method using a point-estimate.

Traditionally the integral over the posterior of the hyperparameters is computed using Markov chain Monte Carlo (MCMC) -methods. However, MCMC methods suffer from a heavy computational burden of Gaussian processes, because the complexity of Gaussian process models grows with the amount of the data used. An alternative approach has been to use only a point estimate for the hyperparameters instead of integrating over their posterior distribution. This is a computationally attractive approach, but it ignores the uncertainty related to the hyperparameters.

The approximations discussed in this thesis attempt to take the uncertainty in the hyperparameters into consideration better than does a point estimate method, and to be computationally lighter than MCMC methods. The results demonstrate that the integration over the hyperparameters is beneficial in particular conditions. In addition, it is shown that a point estimate method yields equally accurate results with the integration methods in other situations. The amount of the data and the use of the models determine the need for the integration methods and the determining conditions are discussed in this work.

Keywords: Bayesian inference, Gaussian processes, hyperparameters, approximate integration

Foreword

This work was carried out in the Department of Biomedical Engineering and Computational Science at the Helsinki University of Technology.

I would like to thank Professor Jouko Lampinen for the opportunity to carry out this work. I am very grateful to Doc. Aki Vehtari for his excellent guidance during the research. M.Sc. Jarno Vanhatalo also deserves thanks for his invaluable help with numerous issues during the work. I would also like to thank the personnel of the department for a friendly work atmosphere and my dear Laura for love and support at home.

In Espoo, January 21, 2010

Ville Pietiläinen

Contents

List of Figures

List of Tables

Abbreviations and notations

$\mathcal{D}$	Observed data
$\mathbf{e}$	Indicator vector
E_i	Expected number of cases
$E[\cdot]$	Expected value
f_0	Distance parameter of CCD
$\mathbf{f}$	Vector of latent values
$\mathbf{f}_*$	Vector of test latent values
$\hat{\mathbf{f}}$	Maximum a posteriori estimate of $\mathbf{f}$
$\mathbf{I}$	Identity matrix
$k(\cdot,\cdot)$	Covariance function
k_{SE}	Squared exponential covariance function
k_{Matern}	Matérn class covariance function
$k_{v=3/2}$	Matérn class covariance function with $\nu = 3/2$
$KL(p\|\|q)$	Kullback-Leibler divergence from p to q
$\mathbf{K}$	Covariance matrix
$\mathbf{K}_{ij}$	ijth element of $\mathbf{K}$
$\mathbf{K}_{\mathbf{f},\mathbf{f}}$	Covariance matrix of training cases
$\mathbf{K}_{*,\mathbf{f}}$	Covariance matrix between test and training cases
$\mathbf{K}_{\mathbf{f},*}$	Covariance matrix between training and test cases
$\mathbf{K}_{\mathbf{y},\mathbf{y}}$	Covariance matrix of training observations
$\mathbf{K}_{*,*}$	Covariance matrix of test cases
K_ν	Modified Bessel function

l_p	Length scale
$\log$	Natural logarithm
m	Dimensionality of the hyperparameters
M	Model assumptions
N	Size of the training data set
$\mathcal{N}(\cdot, \cdot)$	Normal distribution
$p(\cdot)$	Probability density function
Poisson$(\cdot)$	Poisson distribution
t_i	Target output
^{T}	Matrix transpose
$\mathbf{T}$	Cholesky factorization of the scale matrix
$w(\cdot)$	Importance weight
y_*	Predicted output
$\mathbf{y}$	Vector of observations
$\mathbf{z}$	Standardized variable
$\Gamma(\cdot)$	Gamma function
δ_z	Step size of the grid search
δ_π	Threshold of the grid search
Δ	Integration weight in CCD for the points away from the mode
Δ_k	Integration weight
Δ_0	Integration weight in CCD for the point in the mode
ϵ	Noise
$\boldsymbol{\epsilon}$	Sample from a standard multivariate Gaussian
$\boldsymbol{\theta}$	Vector of (hyper)parameters
$\dot{\boldsymbol{\theta}}$	Markov chain sample of $\boldsymbol{\theta}$
$\hat{\boldsymbol{\theta}}$	Point estimate of $\boldsymbol{\theta}$
$\boldsymbol{\mu}$	Mean vector
ν	Degrees of freedom in Student-t distribution
σ_n^2	Variance of the noise
σ_m^2	Magnitude of the Matérn class covariance function
σ_{SE}^2	Magnitude of the squared exponential covariance function

∇	Gradient
$\|\| \cdot \|\|$	Euclidean distance
CCD	Central composite design
CDF	Cumulative density function
GP	Gaussian process
Hessian	Matrix of second derivatives
IS	Importance sampling
MAP-II	Type II maximum a posteriori estimate
MCMC	Markov chain Monte Carlo
MLPD	Mean log-predictive density
ML-II	Type II maximum likelihood estimate
MSE	Mean squared error
SE	Squared exponential covariance function
Sup	Supremum of f is the smallest number greater or equal to any value of f

Chapter 1

Introduction

Gaussian processes have received some attention in machine learning and statistician communities in recent years. They provide flexible tools for various problems, such as Bayesian regression. Gaussian processes are computationally demanding and the complexity of the models increase with the size of the training data set. Therefore, much late research have concentrated on making the inference computationally more feasible.

Full Bayesian treatment of Gaussian processes requires integration over the posterior distribution of a moderate number of hyperparameters. Even though most calculations in Gaussian processes can be analytically solvable, the integral over the posterior of the hyperparameters often is not. A popular approach is to use numerical integration via Markov chain Monte Carlo (MCMC) -methods. However, due to the computational burden, this approach may be infeasible with large data sets.

An alternative to the MCMC-methods are analytical approximations. The integration over the posterior of the hyperparameters can be approximated using only a single point estimate. This approach is computationally attractive. However, in Bayesian inference, all uncertainty should be taken into consideration, whereas a point estimate does not contain information about the uncertainty.

This research started when a more extensive treatment than a point estimate for the hyperparameters was requested by the reviewers of a research article. The study of the integration methods was beyond the scope of the article, but the importance of the approximations was of interest. Hence a Master's thesis study was carried out.

The aim of this research was to implement three numerical approximations for the integration over the hyperparameters and to study the importance of the approximations to the inference with Gaussian processes. A goal was to examine if the integration enhances the predictive performance of Gaussian processes, and whether the point estimate is sufficient in some situations. The implemented integration methods have already been used in an ongoing project.

This thesis is organized as follows. Chapter 2 gives an introduction to Bayesian inference. Chapter 3 discusses basics of Gaussian processes in supervised learning tasks. The numerical approximations are introduced in chapter 4. Some results comparing the point estimate and the integration methods with real-world and simulated data sets are given in chapter 5. Chapter 6 contains the conclusions of this work.

Chapter 2

Bayesian inference

The aim of the Bayesian inference is to formulate probability distributions of unknown quantities based on observations. All uncertainty is expressed with probability distributions. Thus, probabilities can be assigned to events that actually are not random, but from which there is not enough knowledge.

The outcome of a deterministic system can appear to be random, due to lack of information. For example, the result of a coin toss may be predictable given all the influencing variables. However, without the information about those variables, the result of a coin toss is uncertain. There is the same uncertainty about the result even after the coin has landed, until the outcome is observed. The randomness does not change during the observation, but the uncertainty does.

This viewpoint is very different from the frequentist one, in which probability is defined as the relative frequency of favorable outcomes in an infinite long sequence of random tests. In a frequentist framework it would be difficult to assign probabilities for an event that happens only once. It should also be remembered that the Bayesian methods are applied also to frequently occurring events, not only to unique ones.

2.1 Bayes' rule

First in Bayesian inference an observation model $p(y|\boldsymbol{\theta}, M)$ is chosen, which determines how a future observation y depends on the parameters $\boldsymbol{\theta}$. When a set of observations $\mathcal{D}$

is fixed, $p(\mathcal{D}|\boldsymbol{\theta}, M)$ is the likelihood function of the parameters. A posterior probability distribution of interesting parameters is constructed based on an observed data set $\mathcal{D}$, model assumptions M and the *a priori* assumptions about the parameters $p(\boldsymbol{\theta}|M)$. The posterior distribution is composed with Bayes' rule:

$$p(\boldsymbol{\theta}|\mathcal{D}, M) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)}{p(\mathcal{D}|M)}, \tag{2.1}$$

where

- $p(\mathcal{D}|\boldsymbol{\theta}, M)$ is the likelihood function of the parameters $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta}|M)$ is the prior probability distribution of the model parameters $\boldsymbol{\theta}$. This includes prior knowledge about the values of the parameters, as well as about the structure of the dependencies between the parameters.
- $p(\mathcal{D}|M) = \int p(\mathcal{D}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}$ is the marginal likelihood of the data $\mathcal{D}$, also referred to as the evidence of the model. This is a constant which normalizes posterior to be a proper probability density.

In Bayesian analysis, everything is conditioned to the model assumptions M. Therefore, for notational simplicity, M is often left out.

Even though including prior assumptions seems to bring subjectivity to the inference, all statistical reasoning needs some subjective assumptions, for example, about the model structure. Thus, there is always some subjectivity in statistical inference. The degree of subjectivity can be reduced by using prior distributions that let the likelihood term determine the shape of the posterior.

Bayesian formulation provides also an attractive way of updating the posterior distribution by using the former posterior as a new prior and updating it with new observations. This way the inference can be conducted sequentially and the amount of new information provided by the new observations can be monitored through the changes in the posterior distribution.

2.2 Marginalization

Marginalization means that some parameters are integrated out from their joint distribution:

$$\int p(\theta_1, \theta_2) d\theta_2 = p(\theta_1), \tag{2.2}$$

where $p(\theta_1)$ is the marginal distribution of θ_1. Marginalization is often used in order to acquire the posterior distribution of interesting parameters after constructing the joint posterior of all the variables. This often results in complex integrals that are analytically intractable. Some difficult integrals within Gaussian process framework are discussed in Section 4.

2.3 Prediction

Often the objective of the Bayesian modeling is not only to achieve the posterior distribution of parameters, but to use that estimate to make predictions about future observations. The posterior predictive distribution is constructed by marginalizing the parameters $\boldsymbol{\theta}$ out of the joint probability distribution of the prediction y_* and the parameters $\boldsymbol{\theta}$ given the observations $\mathcal{D}$:

$$p(y_*|\mathcal{D}) = \int p(y_*, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} = \int p(y_*|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}. \tag{2.3}$$

As discussed before, the marginalization to construct the predictive distribution in (2.3) is often analytically intractable.

2.4 Model Comparison

Numerous models can be used to explain a certain data set. Bayesian treatment would be to average inference over possible models weighting each with the corresponding model evidence. However, often one model is chosen due to simpler explainability or computation, and inference is conducted via that one model. Many methods and procedures for choosing the best model have been developed (for example, Vehtari and Lampinen, 2002;

Spiegelhalter et al., 2002).

The models in this thesis are compared using the mean squared error (MSE) and the mean log-predictive density (MLPD) measures. These are defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_i (\hat{y}_i - t_i)^2, \tag{2.4}$$

$$\text{MLPD} = \frac{1}{n}\sum_i \log p(y_{*,i} = t_i | x_{*,i}, \mathcal{D}), \tag{2.5}$$

where t_i are the test outputs (not included in the training data set), n is the number of test outputs, $\hat{y}_i = E[p(y_{*,i}|x_{*,i}, \mathcal{D})]$ are the mean predictions of the model with the corresponding input, and $p(y_*|x_*, \mathcal{D})$ is the posterior predictive distribution of y_* with test input x_*. In this work, separate test and training data are available.

MSE measures the distance between the mean of the predictions and the test outputs. However, it does not consider the uncertainty in the predictions. MLPD uses the density of the posterior predictive distribution as a measure of fit, thus taking into account both the distance from the mean and the uncertainty of prediction.

Two probability distributions can be compared using Kullback-Leibler (KL) divergence, which measures divergence from one distribution to another. It is defined as

$$KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx. \tag{2.6}$$

It should be noted, that KL-divergence is asymmetric, thus $KL(p||q) \neq KL(q||p)$. It is always non-negative and zero only if $p(x) = q(x)$.

2.5 Markov chain Monte Carlo -methods

Integrals in the Bayesian inference are often analytically intractable. Such integrals can be numerically approximated with Markov chain Monte Carlo (MCMC) -methods. In Monte Carlo integration the expectation $E[f(\mathbf{x})] = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$ is approximated by

$$E[f(\mathbf{x})] \approx \frac{1}{n}\sum_{k=1}^{n} f(\mathbf{x}^{(k)}), \tag{2.7}$$

where $\mathbf{x}^{(k)}$ are samples drawn from the distribution $p(\mathbf{x})$.

Markov chain methods are algorithms for drawing samples from the target distribution. They form a Markov chain of samples, in which each sample is conditionally independent of all the other samples given the previous one. For thorough presentation of MCMC methods, see, for example, (Neal, 1993; Gilks, 1996)

Quasirandom number sequence

To estimate integral $\int_a^b f(x)dx$ accurately using Monte Carlo -integration, the samples $\dot{x}$ must cover the range $[a, b]$ sufficiently well. However, even if the samples were drawn from a uniform distribution, they may be clustered, leaving some part of the integration interval without samples, thus possibly failing to estimate the integral well.

The integral over desired space could be evaluated in an evenly spaced grid. The total number of the integration points would be the product of the points used for each dimension. Another solution could be the use of quasirandom number sequence. Such a sequence attempts to cover the desired space as evenly as possible with a given number of points.

Integration over a probability distribution could be enhanced by transforming a set of evenly spaced samples so that there will be more samples in a region of high probability and less samples elsewhere. An illustration is given by figures 4.2(a) and 4.2(c). Figure 4.2(a) has samples in a uniform grid and 4.2(c) has uniform samples transformed to be distributed as a Student-t distribution.

The number of quasirandom numbers can be increased sequentially one by one if desired. This could be advantageous if the accuracy of the integration is examined by increasing the number of points used for integration. However, increasing the number of points in an evenly spaced grid is not as straightforward as with quasirandom numbers. The integrations in this thesis have been conducted using a predetermined number of points.

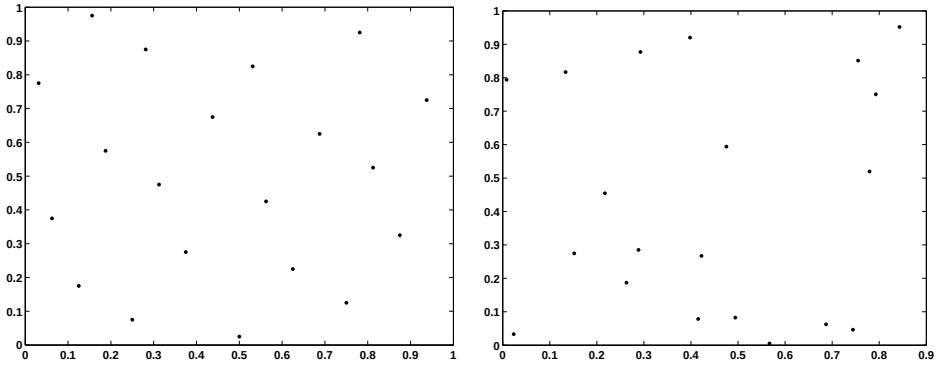
In this thesis, Hammersley quasirandom sequence (Hammersley, 1960) is used. To draw nth of the N k-dimensional samples, let $\pi_2 = 2, \pi_3 = 3, \pi_4 = 5, \ldots \pi_k$ be the first $k-1$ prime numbers. Then n is written in π_i-nary notation, and it is read backwards as π_i-nary decimal. The nth sample $\mathbf{x}_n$ is composed by collecting these decimals y_{ni} using all $i = 2, \ldots, k$ and combining them with $\frac{n}{N} - \frac{1}{2N}$, resulting in $\mathbf{x}_n = \{\frac{n}{N} - \frac{1}{2N}, y_{n2}, y_{n3}, \ldots, y_{nk}\}$.

Table 2.1: Composition of y_{n2}

n	Binary	Reverse decimal	y_{n2}
1	1	0.1	0.5
2	10	0.01	0.25
3	11	0.11	0.75
4	100	0.001	0.125
5	101	0.101	0.625

The steps of composing the y_{nk} is clarified in table 2.1 by using $i = 2$.

This procedure produces points inside a k-dimensional unit hypercube. These samples can be then scaled in order to fill the desired space. Figure 2.1(a) shows 20 points from two dimensional Hammersley quasirandom sequence. They fill the unit box quite evenly, whereas the 20 points from a uniform distribution shown in figure 2.1(b) fill the unit box unevenly leaving large empty spaces with no samples. The even positioning and deterministic composition of quasirandom samples decrease the variance of importance sampling (see section 4.2.3), which is beneficial.



(a) Hammersley quasirandom sequence (b) Uniformly samples sequence

Figure 2.1: 20 samples from two dimensional Hammersley quasirandom (a) and uniformly sampled random sequences (b).

These quasirandom samples can be used to get quasirandom samples from a probability distribution. The cumulative density function (CDF) $P(a) = \int_{-\infty}^{a} p(x)dx$ is a function from $[-\infty, \infty]$ to $[0, 1]$. Hence, samples from desired distribution can be calculated using the inverse CDF with uniformly distributed samples from $[0, 1]$, because if x has CDF $P(x)$, then $y = P(x)$ has a uniform distribution on $[0, 1]$.

Chapter 3

Gaussian processes

A Gaussian process defines a probability distribution of functions. This thesis discusses GP in supervised learning tasks, including regression and spatial analysis in epidemiology. GP can also be used in unsupervised and reinforcement learning tasks, but those are not considered in this thesis.

The aim of a regression task is to learn a mapping from inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ to continuous outputs $y_1, \ldots, y_n$. These observed inputs and outputs compose the training dataset $\mathcal{D}$. The outputs are often assumed to be noisy realizations of an underlying function $f(\mathbf{x})$. GP provides not only one estimate of $f(\mathbf{x})$, but a probability distribution over estimates of $f(\mathbf{x})$.

A Gaussian process provides a prior distribution $p(f)$ over the functions f. A posterior distribution $p(f|\mathcal{D})$ over the functions is calculated using the Bayes' rule (eq. 2.1):

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f)p(f)}{p(\mathcal{D})}. \tag{3.1}$$

An observation y is linked to the the latent variable f through the observation model $p(y|f)$. For example, in regression with additive Gaussian noise, the observation model is

$$y = f + \epsilon, \tag{3.2}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, which can be written as $p(y|f) = \mathcal{N}(y|f, \sigma_n^2)$. In addition to the Gaussian observation model, other models can be used, including Student-t and Poisson-

models.

3.1 Definition of a Gaussian process

Rasmussen and Williams (2006) give the following definition for a Gaussian process:

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

Thus, if a function f is defined by a GP, latent function variables $\mathbf{f} = \{f_1, \ldots, f_n\}$ with corresponding inputs $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are distributed as a multivariate Gaussian:

$$p(\mathbf{f}|\mathbf{X}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}), \tag{3.3}$$

where $\boldsymbol{\mu}$ and $\mathbf{K}$ denote the mean and covariance of the Gaussian distribution, respectively. Because of the consistency properties of Gaussian distribution, any subset of these variables is distributed as a Gaussian, for which the mean and covariance are submatrices of $\boldsymbol{\mu}$ and $\mathbf{K}$.

The entries of the covariance matrix $\mathbf{K}_{ij}$ are defined by a covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$, which will be discussed in Section 3.5. The covariance function has a few hyperparameters, and this thesis presents approximations for the integration over their posterior distribution. Despite of these hyperparameters, GP is referred to as a nonparametric model, meaning that all inference with GP is conditioned on the training data, thus the complexity of the model increases with the size of the training data set.

The distribution of function values $\mathbf{f}$ is always conditioned on the corresponding input variables $\mathbf{X}$. However, in this thesis the distribution of those is not specified, and the inputs are left out of the notation: $p(\mathbf{f}) \hat{=} p(\mathbf{f}|\mathbf{X})$.

3.2 Prediction

A common aim in regression is to predict function value f_* (or observation y_*) in an unobserved test location $\mathbf{x}_*$ given the training data $\mathcal{D}$. In this section, the values of the

hyperparameters are assumed to be given. Their effect will be discussed in Section 3.4. First the joint training and test prior is constructed as

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_{\mathbf{f},\mathbf{f}} & \mathbf{K}_{\mathbf{f},*} \\ \mathbf{K}_{*,\mathbf{f}} & \mathbf{K}_{*,*} \end{bmatrix}\right). \tag{3.4}$$

The posterior distribution of latent variables is constructed using Bayes' rule

$$p(\mathbf{f}_*, \mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}_*, \mathbf{f})}{p(\mathbf{y})}. \tag{3.5}$$

The posterior predictive distribution can be obtained by integrating over the latent variable $\mathbf{f}$ in 3.5:

$$p(\mathbf{f}_*|\mathbf{y}) = \frac{1}{p(\mathbf{y})} \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}_*, \mathbf{f})d\mathbf{f} = \int p(\mathbf{f}_*|\mathbf{f})p(\mathbf{f}|\mathbf{y})d\mathbf{f}. \tag{3.6}$$

Using a Gaussian noise model, the integral in (3.6) can be evaluated analytically. The resulting posterior predictive distribution is

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{y}) &= \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*), \text{where} && (3.7) \\ \boldsymbol{\mu}_* &= \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{y},\mathbf{y}}^{-1}\mathbf{y} && (3.8) \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{*,*} - \mathbf{K}_{*,\mathbf{f}}\mathbf{K}_{\mathbf{y},\mathbf{y}}^{-1}\mathbf{K}_{\mathbf{f},*}, && (3.9) \end{aligned}$$

where $\mathbf{K}_{\mathbf{y},\mathbf{y}} = \mathbf{K}_{\mathbf{f},\mathbf{f}} + \sigma_n^2\mathbf{I}$.

The posterior predictive distribution of the future observations $\mathbf{y}_*$ is constructed by adding the noise $\sigma_n^2\mathbf{I}$ into the covariance matrix in (3.9). The noise parameter σ_n^2 is included into the hyperparameters.

3.3 Non-Gaussian likelihood

If the likelihood in (3.6) is non-Gaussian, the integral over the posterior distribution of the latent variables is analytically intractable. MCMC methods have been used to solve such integrals. However, recently expectation propagation (Minka, 2001) and Laplace

approximations (Williams and Barber, 1998) have been used with success in certain models (Vanhatalo et al., 2009). They both approximate the posterior distribution of latent variables with a Gaussian distribution, thus the predictive distribution $p(\mathbf{f}_*|\mathbf{y})$ becomes analytically tractable. This thesis reviews Laplace approximation in section 3.6 and uses it in section 5.3.

3.4 Effect of hyperparameters

Up to this point, fixed values have been assumed for the hyperparameters that determine the shape of the covariance function. The effect of hyperparameters $\boldsymbol{\theta}$ has to be taken into account in order to perform a full Bayesian inference.

First the posterior distribution of the latent variables and hyperparameters is obtained:

$$p(\mathbf{f}_*, \mathbf{f}, \boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}_*, \mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{3.10}$$

Then, the predictive distribution of $\mathbf{f}_*$ is calculated by integrating latent variables $\mathbf{f}$ and hyperparameters $\boldsymbol{\theta}$ out of the joint distribution:

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*, \mathbf{f}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{f} d\boldsymbol{\theta} = \frac{1}{p(\mathbf{y})} \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}_*, \mathbf{f}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\mathbf{f} d\boldsymbol{\theta}. \tag{3.11}$$

As discussed before, the functional form of the likelihood term determines whether the integral over latent variables is analytically tractable. However, the integral over $\boldsymbol{\theta}$ is usually analytically intractable, and MCMC-methods could be used to perform this integration.

This thesis reviews and examines the properties of three numerical approximations for the integation over the distribution of the hyperparameters (see section 4).

3.5 Covariance functions

This section discusses covariance function $k(\mathbf{x}_i, \mathbf{x}_j)$, which determines the covariance matrix. Two common covariance functions are also reviewed. The form and the hyperparameters of the covariance function determine, for example, how smooth the function

$f(\mathbf{x})$ is.

Any arbitrary function is not a valid covariance function, since the resulting covariance matrix $\mathbf{K}$ must be *symmetric* ($\mathbf{K}^T = \mathbf{K}$) and *positive semidefinite* ($\mathbf{v}^T\mathbf{K}\mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^n$). In addition, the covariance function is *stationary* if it is a function of $\mathbf{x}_i - \mathbf{x}_j$ and *isotropic* if it is a function only of the Euclidean distance of the inputs $||\mathbf{x}_i - \mathbf{x}_j||$. Isotropic covariance function is invariant to shifts and rotations of the input space.

Since the sum of two positive semidefinite matrices is also positive semidefinite, a covariance function can be a sum of two proper covariance functions, $k(\mathbf{x}_i, \mathbf{x}_j) = k_2(\mathbf{x}_i, \mathbf{x}_j) + k_1(\mathbf{x}_i, \mathbf{x}_j)$, where k_1 and k_2 can have distinct properties. A GP with this type of covariance function can have, for example, both fast and slow variations.

The following sections present two popular covariance functions and discuss their properties. For more detailed discussion and other covariance functions, see (Rasmussen and Williams, 2006).

3.5.1 Squared exponential

A popular covariance function in machine learning is squared exponential (SE), which is defined as

$$k_{SE}(\mathbf{x}, \mathbf{x}') = \sigma_{SE}^2 \exp\left(-\sum_{p=1}^{P} \frac{\left(x_p - x_p'\right)^2}{l_p^2} \right), \tag{3.12}$$

where l_p is the length scale and σ_{SE}^2 is the magnitude. These are the hyperparameters of a GP denoted by $\boldsymbol{\theta} = \{l_1, \ldots, l_P, \sigma_{SE}^2\}$. Length scale affects the distance after which the inputs do not significantly correlate. Magnitude scales the overall covariance matrix of a Gaussian process. Squared exponential covariance function is infinitely differentiable, thus a GP using it is very smooth (Rasmussen and Williams, 2006).

As defined above, squared exponential allows dimensions of the input space to have different amount of correlation through their own lengthscales (l_p). If it is reasonable to assume that the properties of the variations of the function are identical in all directions of the input space, each l_p can be replaced with a common length scale l.

Figure 3.1 shows two functions with different length scales. These functions are drawn from the prior distribution $p(f)$, and are not conditioned on any training data. They

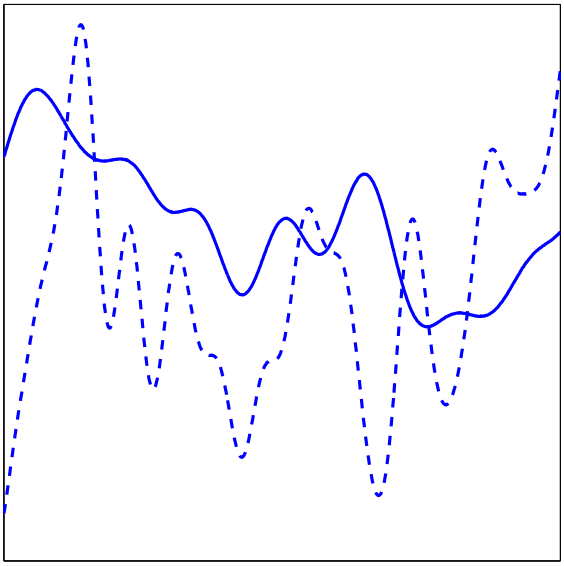


Figure 3.1: Two functions from GP prior with different length scales. Dashed line has length scale of 0.5 and solid line that of 1.

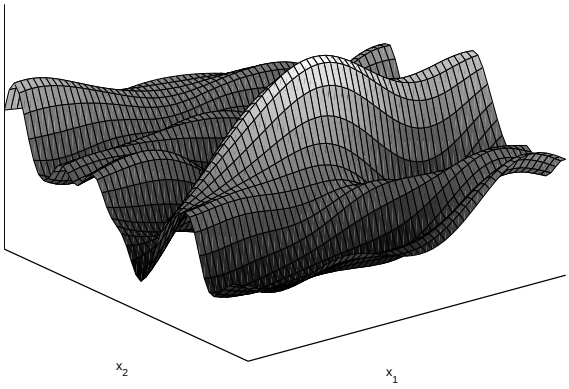


Figure 3.2: A Gaussian process with squared exponential covariance function with different length scales for the two inputs.

demonstrate how the function with shorter length scale varies faster than the one with longer length scale. Figure 3.2 shows a function, which has two inputs. The function is illustrated as a surface, which varies faster in one direction then in the other.

3.5.2 The Matérn class of covariance functions

The Matérn class of covariance functions is popular in geostatistics (Cornford et al., 2002) and is given by

$$k_{Matern}(r) = \sigma_m^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} r\right)^{\nu} K_\nu \left(\sqrt{2\nu} r\right), \tag{3.13}$$

where $r = \sqrt{\sum_{p=1}^{P} \frac{(x_p - x'_p)^2}{l_p^2}}$, l_p is the length scale, σ_m^2 is the magnitude, $\nu > 0$ and K_ν is a modified Bessel function (see, for example, Abramowitz and Stegun, 1972). For $\nu \to \infty$, k_{Matern} becomes a smooth SE-function. For smaller ν, Matérn class covariance functions produce more rough Gaussian processes.

Matérn class covariance function becomes simpler with half-integer ν: $\nu = p + 1/2$, where p is non-negative integer. In this case

$$k_{\nu=p+1/2}(r) = \sigma^2_{\nu=p+1/2} \exp\left(-\frac{\sqrt{2\nu} r}{l}\right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+1)!}{i!(p-1)!} \left(\frac{\sqrt{8\nu} r}{l}\right)^{p-i}, \tag{3.14}$$

and, for example, with $\nu = 3/2$,

$$k_{\nu=3/2}(r) = \sigma^2_{\nu=3/2} \left(1 + \frac{\sqrt{3} r}{l}\right) \exp\left(-\frac{\sqrt{3} r}{l}\right). \tag{3.15}$$

3.6 Normal approximation for the posterior distribution

As discussed in the previous sections, the integral over the posterior distribution of latent variables $\mathbf{f}$ can be analytically intractable. The posterior distribution is often approximated with a Gaussian distribution in order to make the integral in (3.16) analytically tractable. The review of Laplace approximation in this section follows the discussion in (Rasmussen and Williams, 2006).

The posterior predictive distribution can be written as,

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{f}) p(\mathbf{f}|\mathbf{y}) d\mathbf{f}, \tag{3.16}$$

where $p(\mathbf{f}|\mathbf{y}) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f})/p(\mathbf{y})$. If the marginalization is analytically intractable, $p(\mathbf{f}|\mathbf{y})$ is approximated with a Gaussian $q(\mathbf{f})$, in order to make marginalization analytically tractable. Making a second order Taylor expansion of $\log p(\mathbf{f}|\mathbf{y})$ around the mode, a

Gaussian approximation is obtained:

$$q(\mathbf{f}) = N(\mathbf{f}|\hat{\mathbf{f}}, \mathbf{A}^{-1}) \propto \exp\left(-\frac{1}{2}\left(\mathbf{f}-\hat{\mathbf{f}}\right)\mathbf{A}\left(\mathbf{f}-\hat{\mathbf{f}}\right)\right), \tag{3.17}$$

where $\hat{\mathbf{f}} = \text{argmax}_{\mathbf{f}} p(\mathbf{f}|\mathbf{y})$ and $\mathbf{A} = -\nabla\nabla \log p(\mathbf{f}|\mathbf{y})|_{\mathbf{f}=\hat{\mathbf{f}}}$. The approximating Gaussian distribution has its mean set to the mode of the target distribution and the covariance matrix equals to the negative Hessian matrix evaluated in the mode.

The log-posterior of latent variables is written as follows:

$$\begin{aligned}
\log p(\mathbf{f}|\mathbf{y}) &= \log\left\{\frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})}\right\} && (3.18)\\
&= C + \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}) && (3.19)\\
&= C + \log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2}\mathbf{f}^T\mathbf{K}_{\mathbf{f},\mathbf{f}^{-1}}\mathbf{f} - \frac{1}{2}\log|\mathbf{K}_{\mathbf{f},\mathbf{f}}| - \frac{n}{2}\log 2\pi. && (3.20)
\end{aligned}$$

Differentiating (3.18) w.r.t. $\mathbf{f}$ we obtain the first and second derivatives:

$$\begin{aligned}
\nabla \log p(\mathbf{f}|\mathbf{y}) &= \nabla \log p(\mathbf{y}|\mathbf{f}) - \mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1}\mathbf{f} && (3.21)\\
\nabla\nabla \log p(\mathbf{f}|\mathbf{y}) &= \nabla\nabla \log p(\mathbf{y}|\mathbf{f}) - \mathbf{K}_{\mathbf{f},\mathbf{f}}^{-1} = -W - \mathbf{K}^{-1}, && (3.22)
\end{aligned}$$

where $W = -\nabla\nabla p(\mathbf{y}|\mathbf{f})$ (diagonal because observations y_i are independent given latent variables f_i). The mode $\hat{\mathbf{f}}$ can be found, for example, with Newton's method described in (Rasmussen and Williams, 2006, p. 43) using these derivatives.

Chapter 4

Integration over the posterior distribution of the hyperparameters

If the integration over the posterior distribution of hyperparameters in (3.11) is analytically intractable, some approximations are required. A computationally attractive approach is to select only a point estimate for hyperparameters. If this estimate maximizes the marginal posterior density of the hyperparameters, it is referred to as a type II *maximum a posteriori* estimate (MAP-II). This closely relates to type II maximum likelihood (ML-II) estimate (Rasmussen and Williams, 2006), which maximizes the likelihood function of the hyperparameters. The priors used in this work are weakly informative and MAP-II and ML-II estimates should be close to each other.

However, this thesis presents a few numerical approximations for the integration over the posterior distribution of the hyperparameters. MAP-II and the numerical approximations are discussed in the following sections and their performances are compared in chapter 5.

4.1 Type II MAP estimate

The integral over the hyperparameters in (3.11) can be approximated by using a point estimate $\hat{\boldsymbol{\theta}}$ for the hyperparameters:

$$
\begin{aligned}
p(\mathbf{f}_*|\mathbf{y}) &= \int p(\mathbf{f}_*, \mathbf{f}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{f} d\boldsymbol{\theta} && (4.1)\\
&= \frac{1}{p(\mathbf{y})} \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}_*, \mathbf{f}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\mathbf{f} d\boldsymbol{\theta} && (4.2)\\
&= \frac{1}{p(\mathbf{y})} \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}_*, \mathbf{f}|\hat{\boldsymbol{\theta}}) d\mathbf{f}. && (4.3)
\end{aligned}
$$

A good estimate for hyperparameters can be found by maximizing the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Maximizing this equals to the minimization of negative log-posterior cost function:

$$
\begin{aligned}
E &= -\log p(\mathbf{y}|\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) && (4.4)\\
\hat{\boldsymbol{\theta}} &= \text{argmin}_{\boldsymbol{\theta}} E. && (4.5)
\end{aligned}
$$

Corresponding point estimate for the hyperparameters is called MAP-II estimate.

In case of Gaussian observation model, the negative log-posterior cost function becomes

$$
\begin{aligned}
E &= -\frac{1}{2}\mathbf{y}^T \mathbf{K}_{\mathbf{y},\mathbf{y}}^{-1} \mathbf{y} - \frac{1}{2}\log|\mathbf{K}_{\mathbf{y},\mathbf{y}}| - \frac{n}{2}\log 2\pi + \log p(\boldsymbol{\theta}) && (4.6)\\
&= -\frac{1}{2}\mathbf{y}^T \left(\mathbf{K}_{\mathbf{f},\mathbf{f}} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y} - \frac{1}{2}\log|K_{\mathbf{f},\mathbf{f}} + \sigma_n^2 \mathbf{I}| - \frac{n}{2}\log 2\pi + \log p(\boldsymbol{\theta}). && (4.7)
\end{aligned}
$$

The point minimizing (4.6) can be found, for example, by using a gradient based optimizer. However, a problem with the MAP-II estimate is that the uncertainty in the hyperparameters is not considered. Other approximations presented in this work attempt to consider both the location and the shape of the posterior distribution.

If the posterior distribution of the hyperparameters is very narrow, or if the inference with GP is not very sensitive to the choice of hyperparameters, MAP-II estimate can lead to equally good results compared to those of full Bayesian inference.

It should be noted, that the posterior distribution can be multimodal (example given by Rasmussen and Williams, 2006, 116). Gradient based method is guaranteed only to find a local minimum. Therefore, the hyperparameters should either be optimized several times with different initializations, in order to discover several maxima if they exist; or other global optimization method should be used, in order to find the global maximum.

If multiple modes are found, one with the largest posterior probability would be the best estimate if using only a single point estimate. An alternative approach would be to take weighted average of the predictions using the corresponding posterior densities as the weights.

With large data sets, one local optimum is often orders of magnitude more probable than the other optima (Rasmussen and Williams, 2006). Thus, averaging over all modes would have only a small effect to the inference and using only one point estimate would be accurate enough.

4.2 Approximating the integral over the distribution of the hyperparameters

The methods used in this work approximate the integration over the hyperparameters of a GP. These approximations rely on replacing the continuous integration over hyperparameters in (3.11) with a discrete summation using suitable points in the hyperparameter space.

The posterior predictive distribution can be written as

$$
\begin{aligned}
p(\mathbf{f}_*|\mathbf{y}) &= \int p(\mathbf{f}_*|\mathbf{f},\boldsymbol{\theta})p(\mathbf{f}|\mathbf{y},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\mathbf{f}d\boldsymbol{\theta} && (4.8)\\
&\approx \sum_{k=1}^{m}\left[\int p(\mathbf{f}_*|\mathbf{f},\boldsymbol{\theta}_k)p(\mathbf{f}|\mathbf{y},\boldsymbol{\theta}_k)d\mathbf{f}\right]p(\boldsymbol{\theta}_k|\mathbf{y})\Delta_k && (4.9)\\
&= \sum_{k=1}^{m} p(\mathbf{f}_*|\mathbf{y},\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|\mathbf{y})\Delta_k, && (4.10)
\end{aligned}
$$

where the sum is calculated over the values of $\boldsymbol{\theta}_k$ with weights Δ_k. The important idea is to approximate the continuous integral over the posterior distribution of the hyperparameters with a discrete summation using a set of point estimates $\boldsymbol{\theta}_k$ and corresponding area weights Δ_k. In one dimension, this corresponds to estimating the distribution with a histogram, where the values of Δ_k equal to the widths of the bins.

The differences between methods presented in this work arise from the selection of the

integration points $\boldsymbol{\theta}_k$ and the corresponding weights Δ_k. The three approximations are named grid search, central composite design (CCD), and importance sampling (IS) with Student-t proposal distribution:

- Grid search attempts to approximate the integral by a weighted average on a regular grid
- CCD selects a small number of point estimates of hyperparameters symmetrically around the mode of the posterior and approximates the integral by a weighted average of these points
- IS draws quasirandom samples from an importance distribution, evaluates the importance weights for these points and approximates the integral by a weighted average of these points.

These methods will be reviewed in the following subsections.

4.2.1 Grid search

An intuitive approach to approximate a distribution is to explore it in a grid. However, the size of the grid grows exponentially with the dimensions of the distribution. Moreover, the correlations and different scales of the parameters can make a regular grid laid in the main directions of parameters ineffective. Each evaluation of the posterior probability $p(\boldsymbol{\theta}_k|\mathbf{y})$ requires inversion of $\mathbf{K}$, which is a computationally heavy operation. Therefore, it is important to minimize the number of $\boldsymbol{\theta}_k$ while assuring the coverage of the relevant parts of the posterior distribution. Equation (4.10) shows that the predictions are weighted with $p(\boldsymbol{\theta}_k|\mathbf{y})$. Thus, the evaluation of the posterior distribution can be restricted to the area with significant posterior density.

The grid search follows the work of Rue et al. (2009). First the posterior mode $\hat{\boldsymbol{\theta}}$ is located by maximizing the log-posterior distribution $\log\{p(\boldsymbol{\theta}|\mathbf{y})\}$. Then, the shape of the log-posterior is approximated with a Gaussian, the covariance matrix $\boldsymbol{\Sigma}$ of which is the inverse of the negative Hessian at the mode. The Hessian can be evaluated using finite differences. To aid the exploration, a standardized variable $\boldsymbol{z}$ is introduced. Let $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathbf{T}}$ be the eigenvalue decomposition of $\boldsymbol{\Sigma}$. $\boldsymbol{\theta}$ can be defined via $\mathbf{z}$, as follows:

$$\boldsymbol{\theta}(\mathbf{z}) = \boldsymbol{\theta}^* + \mathbf{V}\boldsymbol{\Lambda}^{\mathbf{1/2}}\mathbf{z}. \tag{4.11}$$

The standardized variable $\mathbf{z}$ can be used to explore posterior more effectively, as it takes into account the directions of the correlations, as well as the different scales of the parameters.

After standardization, the posterior is explored in the main directions of $\mathbf{z}$ by taking steps of δ_z from the mode to the negative and positive direction. The exploration into each direction is stopped when the log-posterior drops enough compared to that in the mode:

$$\log\{p(\boldsymbol{\theta}(\mathbf{0})\} - \log\{p(\boldsymbol{\theta}(\mathbf{z})\} > \delta_\pi, \tag{4.12}$$

where δ_π is a chosen threshold. Thus, the points $\boldsymbol{\theta}(\mathbf{z})$ are accepted if their density is significant enough. In addition, all the locations of $\boldsymbol{\theta}(\mathbf{z})$ are explored which can be reached from an accepted location with one step of lenght δ_z in any direction of main axes. If such a location is accepted, adjacent locations are studied, and so on.

Since the points are laid on a regular grid, the integration weights Δ_k in (4.10) are equal. However, their exact values are not of interest since the posterior distribution can be normalized afterwards to be a proper probability distribution.

This algorithm explores points in a grid determined by the location of the mode and the directions of $\mathbf{z}$. All points with a significant posterior density compared to that of the mode are included. The acceptance threshold δ_π can be increased and the step size δ_z decreased, in order to achieve more accurate approximation of the posterior distribution.

A problem with grid search is that the size of the grid grows exponentially with the number of hyperparameters. Even though the amount of evaluation points can be decreased by changing parameters δ_π and δ_z, accurate approximation can be a computationally heavy task.

With a small number of hyperparameters, the grid search provides a fast method for approximating the posterior and integrating over it. An attractive feature of the grid search is that regardless of the symmetric Gaussian approximation, it approximates asymmetric posterior distributions as well as symmetric ones. In addition, it can even cover multiple modes, if the drop in the posterior density does not exceed the threshold δ_π between the modes.

4.2.2 Central composite design

If the dimensionality of the hyperparameters is high, the grid search can become computationally infeasible. Central composite design (CCD) addresses this problem. Using CCD, each direction of $\mathbf{z}$ (reparametrization in Section 4.2.1) are given only two possible levels $\pm f_0$, where f_0 is a parameter controlling distance from the mode.

If the number of dimensions m is high, not all possible combinations of these values are examined but a smaller, computationally more feasible set. The choice of these points is discussed by Sanchez and Sanchez (2005). In addition, the point in the mode and $\pm f_0\sqrt{m}$ on each main axis of $\mathbf{z}$ are included. Thus, all but the central point lay on a sphere with a radius of $f_0\sqrt{m}$. f_0 have to be greater than one in order to get a positive weight for the central point (see equation 4.15). In this work $f_0 = 1.1$.

The advantage of using CCD is that the number of points used remains very low compared to that of grid search. In addition, CCD considers the shape of the posterior distribution and covers a significant area of the posterior distribution of the hyperparameters. However, the integration using CCD can be inaccurate due to the small number of integration points.

Weights of the CCD

Even though the points on the sphere are regularly laid, the central point deviates from those, thus requiring a unique integration weight. To determine the integration weights for the integration points, $p(\boldsymbol{\theta}(\mathbf{z})|\mathbf{y})$ is assumed to be a standard Gaussian after reparametrization. Thus, $E(\boldsymbol{\theta}^T\boldsymbol{\theta}) = m$ and $\int p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1$. These requirements give the integration weights for the points on the sphere with radius $f_0\sqrt{m}$. Due to the symmetrical positioning of the points, the integration weights are equal for the points on the sphere.

$$E(\boldsymbol{\theta}^T\boldsymbol{\theta}) = \sum_k \boldsymbol{\theta}_k^T\boldsymbol{\theta}_k p(\boldsymbol{\theta}_k)\Delta_k = (n_p - 1)mf_0^2(2\pi)^{-m/2}\exp\left(-\frac{mf_0^2}{2}\right)\Delta = m, \quad (4.13)$$

where n_p is the number of design points. From 4.13 it follows that

$$\Delta = \left[(n_p - 1)f_0^2(2\pi)^{-m/2}\exp\left(-\frac{mf_0^2}{2}\right)\right]^{-1}. \quad (4.14)$$

The integral over $p(\boldsymbol{\theta}|\mathbf{y})$ is approximated with $p\left(\boldsymbol{\theta}(\mathbf{z}=\mathbf{0})\right)\Delta_0 + \sum_k p(\boldsymbol{\theta}_k)\Delta_k$. Thus, the weight of the central point is

$$\begin{aligned}\Delta_0 = \frac{1-\sum_k p(\boldsymbol{\theta}_k|\mathbf{y})\Delta}{p(\boldsymbol{\theta}(\mathbf{z}=\mathbf{0})|\mathbf{y})} &= \frac{1-(n_p-1)(2\pi)^{-m/2}\exp\left(-\frac{mf_0^2}{2}\right)\Delta}{(2\pi)^{-m/2}} \\ &= (2\pi)^{m/2}(1-\frac{1}{f_0^2}). \end{aligned} \tag{4.15}$$

The weights can be rescaled as $p(\mathbf{f}_*|\mathbf{y})$ will be normalized in the end. Thus, weights can be written in a simpler form:

$$\Delta_0 = 1 \tag{4.16}$$

$$\Delta = \left[(n_p-1)\exp\left(-\frac{mf_0^2}{2}\right)\left(f_0^2-1\right)\right]^{-1}. \tag{4.17}$$

4.2.3 Importance sampling with Student-t proposal distribution

The integration over posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ can be evaluated by drawing samples $\dot{\boldsymbol{\theta}}$ from $p(\boldsymbol{\theta}|\mathbf{y})$. Then (4.9) can be evaluated as

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \approx \frac{1}{N}\sum_i^N p(\mathbf{f}_*|\dot{\boldsymbol{\theta}}_i,\mathbf{y}) \tag{4.18}$$

where $\dot{\boldsymbol{\theta}}_i$ are the samples drawn from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ (Gelman et al., 2003, p. 342).

The posterior distribution of $\boldsymbol{\theta}$ is not of a form from which samples can be directly drawn. Therefore, (4.18) is evaluated by drawing samples $\dot{\boldsymbol{\theta}}_i$ from another distribution $q(\boldsymbol{\theta})$ and the calculations are corrected with the weight $w(\dot{\boldsymbol{\theta}}_i) = p(\dot{\boldsymbol{\theta}}_i|\mathbf{y})/q(\dot{\boldsymbol{\theta}}_i)$:

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \tag{4.19}$$

$$= \int p(\mathbf{f}_*|\boldsymbol{\theta},\mathbf{y})\left[\frac{p(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})}\right]q(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{4.20}$$

$$\approx \frac{1}{N}\sum_{i=1}^N p(\mathbf{f}_*|\dot{\boldsymbol{\theta}}_i,\mathbf{y})w(\dot{\boldsymbol{\theta}}_i). \tag{4.21}$$

The proposal distribution $q(\boldsymbol{\theta})$ must have nonzero values in the areas where the true posterior has nonzero values. In addition, the values of the approximation should not be significantly lower than those of the true posterior, because the weight of such a sample would be large and it would dominate the summation in (4.21).

It should be noted, that samples having high $q(\dot{\boldsymbol{\theta}})$ but low $p(\dot{\boldsymbol{\theta}}|\mathbf{y})$ do not pose a significant problem, because the weights $w(\dot{\boldsymbol{\theta}})$ for such samples are small, thus neglecting these false samples from the summation in (4.21).

Student-t distribution is chosen to be the importance distribution, from which the samples are drawn. Student-t distribution with a small number of degrees of freedom ν has fat tails and should have sufficient probability density where the target distribution has it. Student-t distribution approaches a Gaussian distribution when the degrees of freedom grows, and desired tail behavior can be selected with suitable ν.

Selecting the importance distribution

The following procedure for the selection of the importance distribution follows the work of Geweke (1989). First the Hessian of the log-posterior distribution is evaluated at the mode $\hat{\boldsymbol{\theta}}$, and the inverse of the negative Hessian is used as the scale matrix of the Student-t distribution. However, this scale matrix may poorly predict the posterior density far from the mode, especially if the posterior is asymmetric. Therefore the scale is adjusted in order to better estimate the posterior.

The posterior is explored in directions determined by the Hessian matrix, in order to find the locations with the largest drop in the importance density compared to the drop in the posterior density. Then, the scale of the importance distribution in corresponding direction is adjusted, in order to match the drops in log-densities. This procedure ensures that the importance distribution does not decline faster than posterior distribution (in explored directions), and it should not underestimate the posterior distribution.

The scale which makes the drops in posterior densities equal is defined as a function δ corresponding to the distance from the mode:

$$f_i(\delta) = \nu^{-1/2}|\delta| \left\{ \left[p(\hat{\boldsymbol{\theta}}) / p(\hat{\boldsymbol{\theta}} + \delta \mathbf{T}\mathbf{e}^{(i)}) \right]^{2/(\nu+k)} - 1 \right\}^{-1/2}, \tag{4.22}$$

where $\mathbf{T}$ is a factorization such that the inverse of $\mathbf{T}\mathbf{T}^T$ is the negative Hessian of log-posterior at the mode, k is the dimensionality of hyperparameter space, and $\mathbf{e}^{(i)}$ is a $k \times 1$ indicator vector, $e_i^{(i)} = ||\mathbf{e}^{(i)}|| = 1$.

The scaling factors q_i and r_i are defined as

$$q_i = \sup_{\delta>0} f_i(\delta) \tag{4.23}$$

$$r_i = \sup_{\delta<0} f_i(\delta). \tag{4.24}$$

In addition let $sgn^+(\alpha) = 1$, if $\alpha \geq 0$ and $sgn^-(\alpha) = 1 - sgn^+(\alpha)$. Now samples $\dot{\boldsymbol{\theta}}$ can be drawn from the split Student-t distribution as follows:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4.25}$$

$$\eta_i = \left[q_i sgn^+(\epsilon_i) + r_i sgn^-(\epsilon_i)\right] \epsilon_i (\zeta/\nu)^{-1/2}, \text{ where } \zeta_i \sim \chi^2(\nu) \tag{4.26}$$

$$\dot{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \mathbf{T}\boldsymbol{\eta}, \tag{4.27}$$

where $\boldsymbol{\epsilon}$ are drawn using quasirandom sequence.

The log-probability density function at $\dot{\boldsymbol{\theta}}$ is (up to an additive constant)

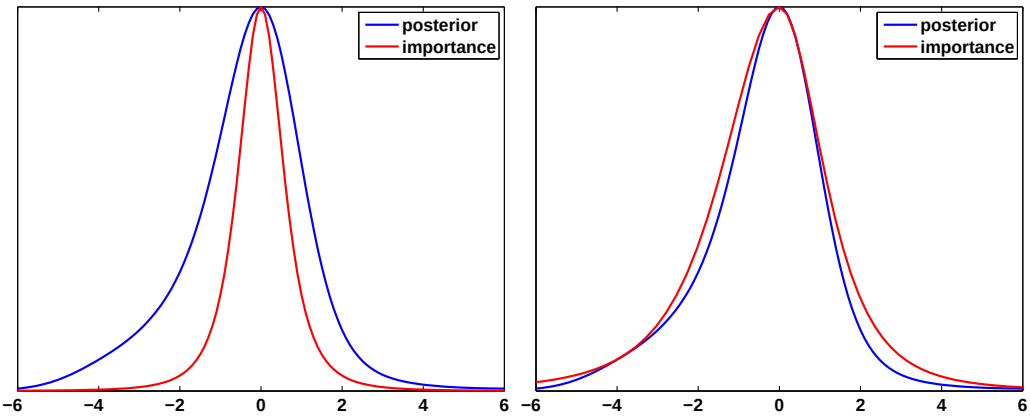
$$q(\dot{\boldsymbol{\theta}}) = -\sum_{i=1}^{k} \left[\log(q_i) sgn^+(\epsilon_i) + \log(r_i) sgn^-(\epsilon)\right] - \left[(\nu + k)/2\right] \log\left(1 + \sum_{i=1}^{k} \frac{\epsilon_i^2}{\zeta_i}\right). \tag{4.28}$$

Due to the rescaling, the distribution of these samples should resemble the posterior distribution better than without the rescaling, making the importance sampling effective.

Figure 4.1(a) illustrates an importance sampling distribution fit to the posterior distribution of the hyperparameters. The symmetric importance distribution poorly fits the asymmetric posterior distribution. Figure 4.1(b) shows that the scaling presented above helps the importance distribution to fit more closely to the posterior distribution.

4.2.4 Summary of the approximation methods

Figure 4.2 shows one example of the posterior distribution of the hyperparameters with the points $\boldsymbol{\theta}_k$ used in approximation with different methods. With the presented methods and parameters, the grid search is able to cover the largest area. It was observed that the

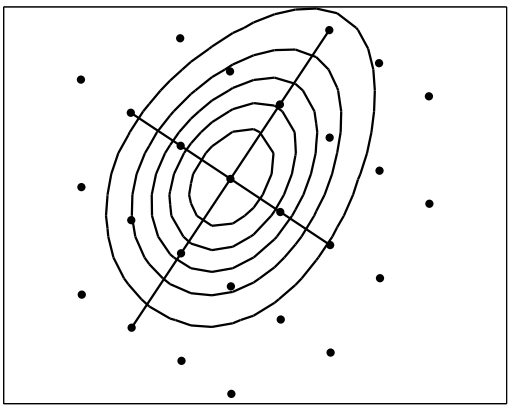


(a) Distributions without scaling (b) Distributions with scaling

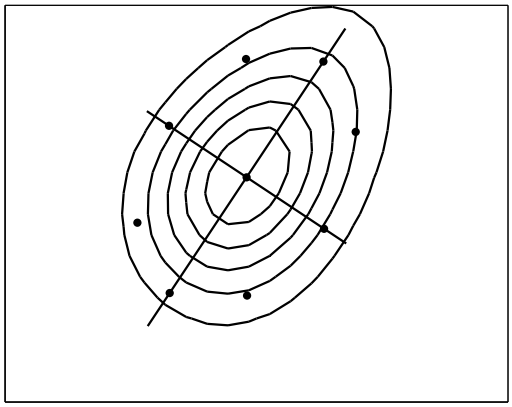
Figure 4.1: An illustration of posterior distribution (blue) and importance distribution (red) with and without scaling.

area covered is more important than the density of the grid. CCD method is able to cover a broad area with only a small number of integration points. However, estimating the shape of the posterior may be difficult with a small number of points, thus the integral might not be as accurate as with the grid search.

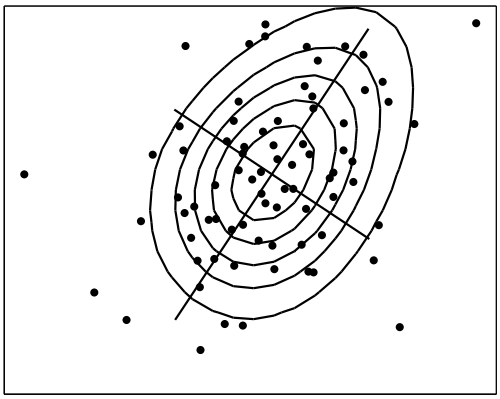
The area covered by these methods is easily varied by changing parameter δ_π for the grid search and f_0 for CCD. However, increasing δ_π makes the exploration computationally more demanding. Increasing the number of samples of IS increases the area covered only little, because the posterior density is low far from the mode. The area can easily be enlarged by using lower degrees of freedom ν.



(a) Grid search



(b) CCD



(c) 80 quasirandom samples from Student-t distribution with $\nu = 10$

Figure 4.2: The posterior distribution of hyperparameters is illustrated with contour lines and integration points $\boldsymbol{\theta}_i$ of different methods are marked with black dots.

Chapter 5

Results

5.1 Regression with a multimodal hyperparameter posterior distribution

This experiment uses a data set introduced by Neal (1997). The observations are noisy realizations from latent function $f(x) = 0.3 + 0.4x + 0.5\sin(2.7x) + 1.1/(1 + x^2)$. Most samples are contamined by noise from zero mean Gaussian with standard deviation 0.1. However, with a probability of 0.05, the standard deviation of the noise was 1 and the corresponding output is regarded as an outlier.

Figure 5.1 shows the latent function and 100 samples (the training data set) drawn from it. The inputs are drawn from a standard Gaussian. A few possible outliers can be seen as some samples deviate significantly from the other samples.

GP inference was conducted using squared exponential covariance function. Student-t distribution with one degree of freedom and a scale of 36 was used as a prior for the hyperparameters. This prior distribution is broad and gives significant probability for both small and large values of hyperparameters. Hence, prior distribution should not significantly restrict the shape of posterior distribution by suppressing areas where likelihood term has non-zero probability. It should be noted that there is a model misspecification, because the fitted model has same observation model for all the observations, whereas the real observations arise from two different observation models. This might increase the uncertainty in the estimation of the hyperparameters.

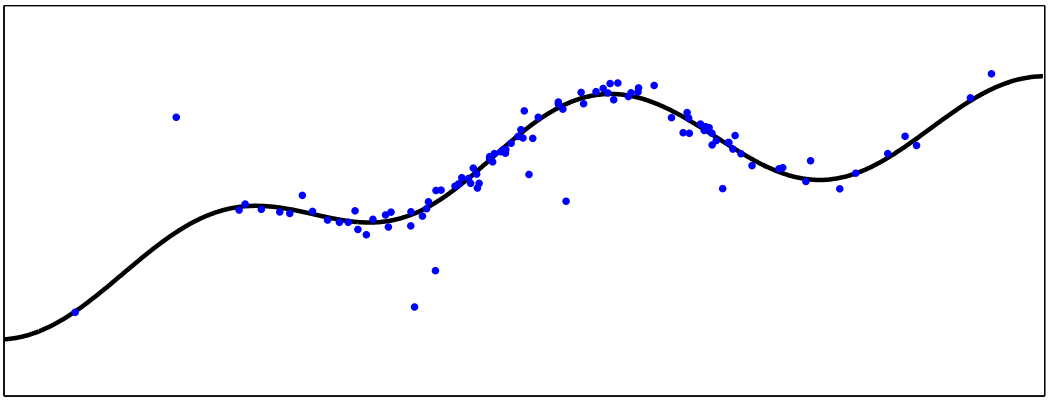


Figure 5.1: Latent function (full line) and 100 samples drawn from it (dots).

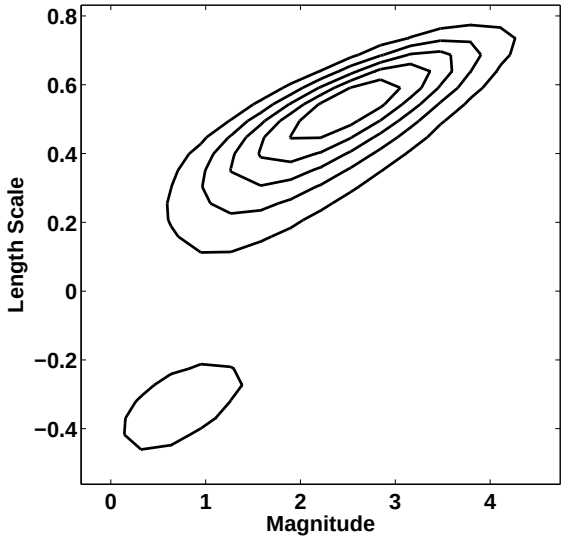


Figure 5.2: Contours of marginal posterior distribution of the length scale and magnitude.

Interestingly, the posterior distribution of the hyperparameters was discovered to contain at least two modes. The more significant mode has a longer length scale than the other, which is reasonable as the latent function $f(x)$ varies quite slowly. However, the presence of outliers suggest that the function might vary rapidly, thus raising another mode the posterior distribution with a noticeably shorter length scale. The marginal distribution of the length scale and magnitude is illustrated in figure 5.2.

The hyperparameter space can be divided in to attraction areas of the modes. An attraction area is defined as hyperparameter values from which the hyperparameters converge to a certain mode during the optimization. Figures 5.3(a)-5.3(c) illustrate the attraction areas for three different optimization algorithms. The irregular shape of the areas should be noticed, as well as that initializing a certain parameter to a high value does not guarantee it to converge to the mode with a high value for that parameter. The red contours illustrate

the marginal posterior distribution of the length scale and magnitude with a fixed value for the noise hyperparameter σ_n^2 (shown in the title). Some projected sample paths of the optimization are also shown with a green line.

As discussed earlier, the inference should be conducted using each discovered mode separately. The approximations in the modes must not overlap significantly. Otherwise overlapping region would be integrated twice.

Both modes were used weighting the predictions with the probability density of the mode in case of the MAP-II estimate. For CCD and importance sampling approaches, the total volumes of normal approximations on the modes were used as weights for the predictions. The weights were close to those estimated from the points of grid search by dividing them to the two modes and calculating corresponding estimates for the volumes, indicating that the approximations did not overlap significantly. Grid search was able to reach both modes with a single initialization if the MAP-II estimate of the hyperparameters converged to the lower mode. Both modes were reachable from the higher mode if the step size δ_z was reduced to 0.5.

Because the two modes were quite close to each other, Student-t distribution with small degrees of freedom tended to give some samples from different mode than it was intended to, thus obtaining samples with extreme weights. Therefore $\nu = 14$ was used, which seemed to prevent the sampling from the wrong mode.

Figure 5.4 shows the means of the predictions using MAP-II estimates from the modes of posterior distribution. The final prediction is a mixture of two Gaussians located around these individual predictions. The green line derives from the short length scale and varies significantly faster than the blue line deriving from the mode with a larger length scale. However, the behaviors of the two functions are very similar in the regions with high density in the training observations, indicating that reasonable variations in the values of the hyperparameters do not matter if there is enough training data present.

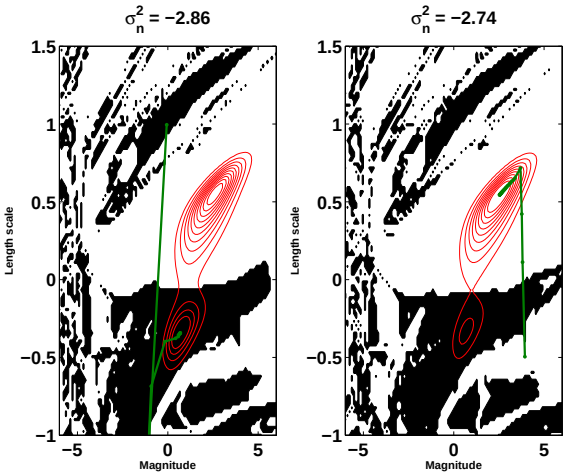
Figure 5.5 shows the samples used in the grid search, CCD and importance sampling approaches. The density of points in grid search does not have significant effect on the predictions; more important is the covered area. An advantage of the grid search is that it covers smoothly a large area of the posterior. In this example, the reparametrization of grid search was done only by scaling the main axes without rotation, in order to get a clearer figure. The lack of rotation did not affect significantly the results. Figure 5.5(a)

shows multiple points around the lower mode because three dimensional points were projected to two dimensions. The projections for the higher mode overlapped. Therefore less points can be seen.

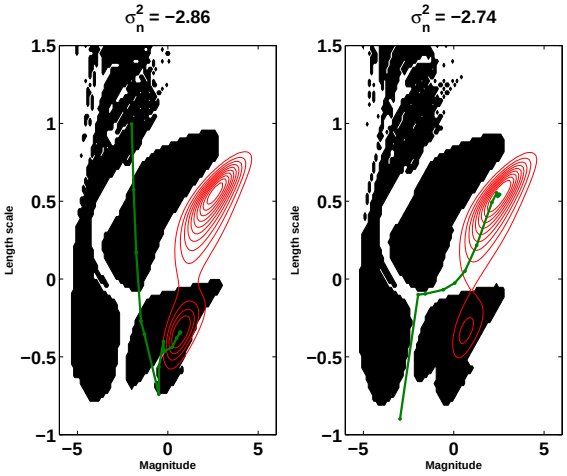
The methods were tested using a set of 1000 samples from the latent function. The test inputs are drawn uniformly from the range of the training inputs. Figure 5.6 summarizes the MLPD and MSE measures for each approach. The high values of MLPD and low values of MSE are preferred.

Integration methods outperformed MAP-II estimate with both measures, clearly indicating that integration was advantageous. However, the differences in MSE are small, demonstrating that the means of the predictions are very similar between the methods. Thus, the difference in MLPD results from differences in variances rather than in means. Integration approximations compensate the mismatches in the means of the predictions with a greater variance, thus increasing the density at the test outputs lying far from the mean of the prediction.

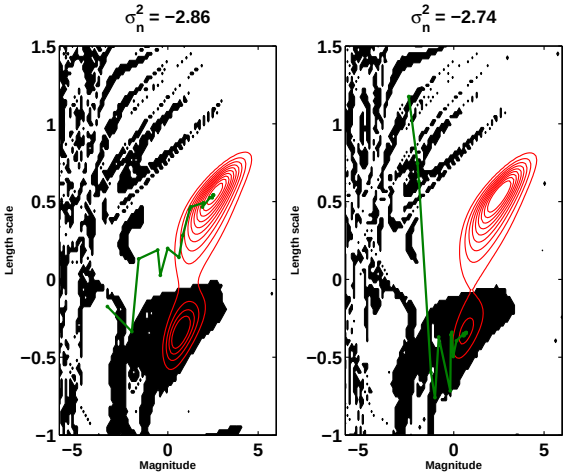
However, if the range of the test inputs is restricted so that the density of the training inputs is high, the differences between MLPD and MSE measures diminish to almost negligible. The differences in behaviors of the approaches are clearest if the distance from a prediction to the training inputs is large. The effect of the hyperparameters diminish when the training data is able to determine the $f(\mathbf{x})$.



(a) Hyperparameters are optimized with the method proposed by Coleman and Li (1996).



(b) Hyperparameters are optimized using quasi-Newton method discussed by Davidon (1991)



(c) The hyperparameters are optimized using scaled conjugate gradient discussed by Bishop (1996)

Figure 5.3: Attraction areas of the two modes. The hyperparameters are optimized using different optimization methods. The scale of the hyperparameters is logarithmic.

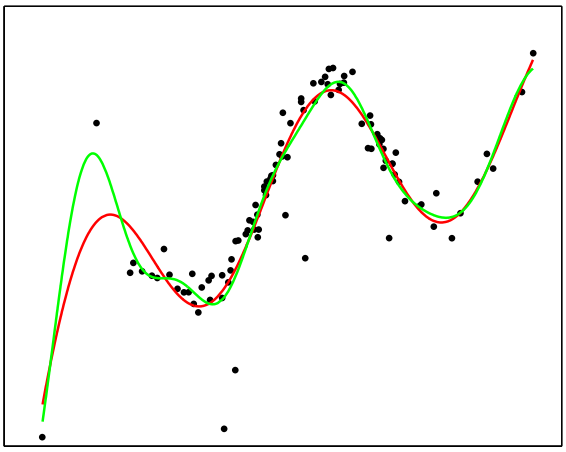
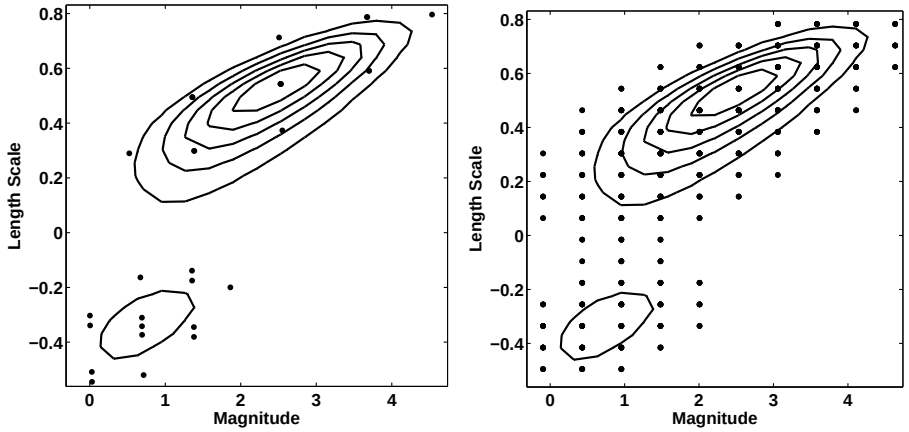
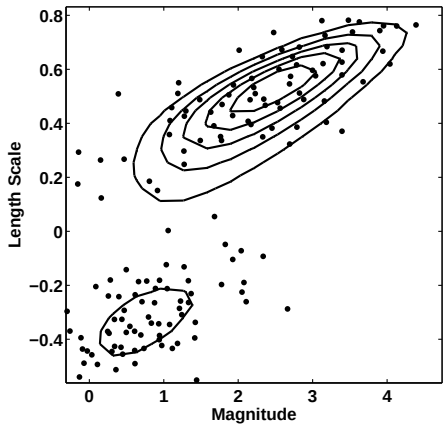


Figure 5.4: The means of predictions with MAP-II estimate from different modes. The red line represents the predictions with higher length scale. Black dots mark the training data



(a) Estimated density and total 30 CCD integration points covering both modes

(b) Estimated density and total 603 grid integration points covering both modes



(c) Estimated density and total 240 IS integration points covering both modes

Figure 5.5: Contours represent the marginal posterior distribution. Hyperparameters used for integration are marked with black dots.

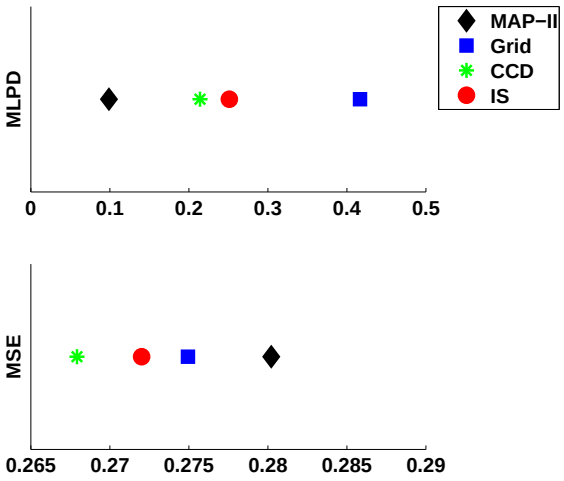


Figure 5.6: MLPD and MSE measures for all the approaches with the Neal data

5.2 Regression with a unimodal hyperparameter posterior distribution

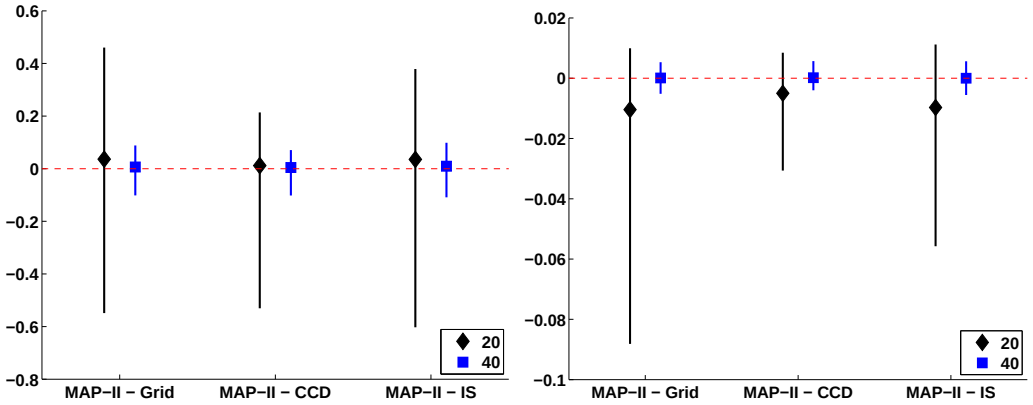
In this test, all the training observations were drawn from the latent function used in section 5.1 with a constant variance for the noise ($\sigma_n^2 = 0.04$). With a sufficient sample size the posterior distribution of the hyperparameters is unimodal. 1000 test inputs were uniformly sampled from the range of the training inputs. Inference was conducted via GP with a squared exponential covariance function. The prior for the hyperparameters was Student-t distribution with scale of 36 and one degree of freedom.

In this section, a correct model is used as all the observations arise from the same model. This should make the estimation of the hyperparameters easier than in the previous section, where there was a model misspecification.

200 sets of 20 and 40 training observations were sampled with fixed training inputs for each set size. Figure 5.7 summarizes the differences between MAP-II and integration methods showing the mean of the difference and the 95% interval. With 20 training inputs, the means of the results favor the point estimate method. However, the difference is small and not significantly in favor of the point estimate, because the distribution is wide.

The distribution of the MLPD measures using MAP-II estimate is slightly wider that those of the integration methods, suggesting that the results of integration methods are more stable. However, the difference is not significant. The distributions of the differences between integration methods are significantly narrower compared to those in figure 5.7.

An interesting phenomenon in this experiment is that the results are even closer to zero and the intervals are considerably narrower when 40 training inputs are used. This suggests that the difference in the predictions of the methods approach zero when the size of the training data increases. Figures 5.8 and 5.9 demonstrate this by showing samples of the predictive distributions. The predictions are close to each other with 20 training inputs, but they are almost identical with 40 training inputs. They also fit the underlying function better when more training data is available.



(a) The differences in MLPD (b) The differences in MSE

Figure 5.7: The means of the differences between MAP-II and integration methods with the 95% confidence intervals.

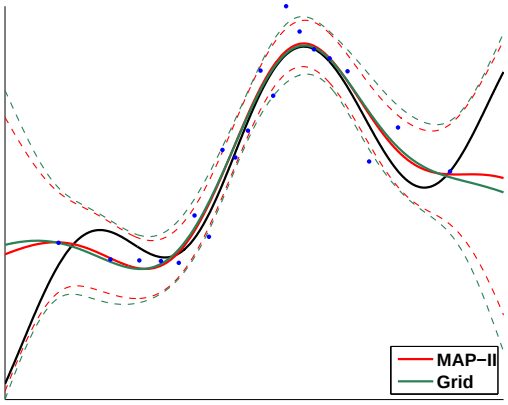


Figure 5.8: The means of the predictions of MAP-II and grid search approaches (solid lines) with the intervals of two standard deviations (dashed lines). The training data of 20 observations (blue dots) and the latent function (black line) are also shown.

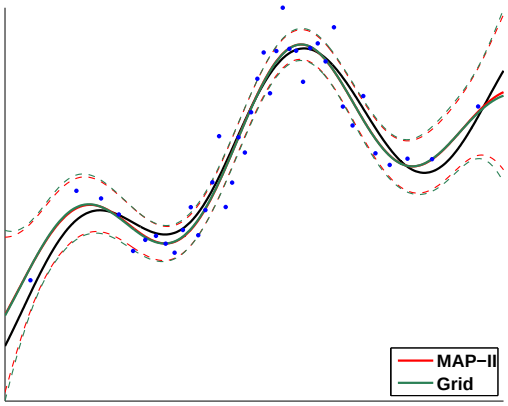


Figure 5.9: The means of the predictions of MAP-II and grid search approaches (solid lines) with the intervals of two standard deviations (dashed lines). The red line is mostly under the green line. The training data of 40 observations (blue dots) and the latent function (black line) are also shown.

5.3 Regression with Poisson observation model

A field of study often using a Poisson observation model is spatial epidemiology. The aim of the spatial epidemiology is to model, for example, the spatial variation in the risk of a certain disease. In this example, simulated data is used, but similar approaches are used in real-world data analysis. Bayesian spatial analysis is discussed, for example, by Best et al. (2005).

First the area of study is divided into regions indexed by i. The area of interest could be, for example, Finland and the regions would be counties or the cells of a grid. The inputs $\mathbf{x}$ would then be the spatial coordinates of each region.

For each area, the expected number of cases E_i is calculated based on, for example, the age, sex, and education structure of the population. If the disease cases are independent of each other, the observed number of the disease cases y_i would be distributed as a binomial distribution. With sufficiently rare disease and a large number of people, the binomial is approximately a Poisson distribution. The mean of the Poisson distribution is the product of the expected number E_i and the local relative risk $\exp(\mu_i)$, which is the variable of interest in a spatial study. The logarithm of the local risks is given a Gaussian process prior, because it is assumed that areas near each other have similar relative risks.

Formally, the model is given as

$$y_i \sim \text{Poisson}(E_i \exp(\mu_i)) \tag{5.1}$$

$$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \tag{5.2}$$

where $\mathbf{K}$ is determined by the covariance function of the Gaussian process. Zero mean assumption for the log-risk $\boldsymbol{\mu}$ is reasonable because the expected number of cases E_i is used.

The inputs relating to the observations were sampled from a standard Gaussian distribution. These correspond to the regions of the spatial study. The log-relative risks μ_i of the regions x_i were drawn from the latent function $f(x) = -1.1 + 0.4x + 0.5\sin(2.7x) + 1.1/(1 + x^2)$. Constant expected number of cases E_i was assumed for each i. Finally the number of cases was drawn from Poisson distribution.

The posterior predictive distribution was conducted via GP with a squared exponential

covariance function. The prior for the hyperparameters was Student-t distribution with scale of 36 and one degrees of freedom. Laplace approximation was used in order to solve the intractable integral over latent variables.

The number of regions and the expected number of occurrences were varied in order to examine their effect on the MAP-II and integration approaches. The number of observations N used was 20, 40, and 80; and the expected number of occurrences E_i was 1, 2, and 4.

350 sets of observations were sampled and the difference between MLPD measures of MAP-II and CCD approaches were calculated for each set. Figure 5.10 summarize the means of the differences and 95% confidence intervals with varying E_i and N. It can be seen, that the confidence interval tends to become narrower as either of the parameters grow. In addition, the difference between methods approach zeros as the values of the parameters grow.

It should be noted that the distributions of the differences have long tails to the negative direction. Majority of the differences concentrate close to zero and the median value is little higher than the mean. Thus, the significance of the differences between methods may be lower than expected. However, it is interesting that the significance of the difference gets lower as the parameters grow.

The distributions of MLPD measures using MAP-II approach are slightly wider to the negative direction than those of integration methods, hence the distributions of the differences have long tails to the negative direction. The results using integration methods are somewhat more stable, and they do not yield as poor results as MAP-II approach in some situations.

Figure 5.11 shows the observations and the predictions with four different combinations of parameters N and E_i. It also demonstrates that the differences between the predictions diminish as the values of the parameters increase.

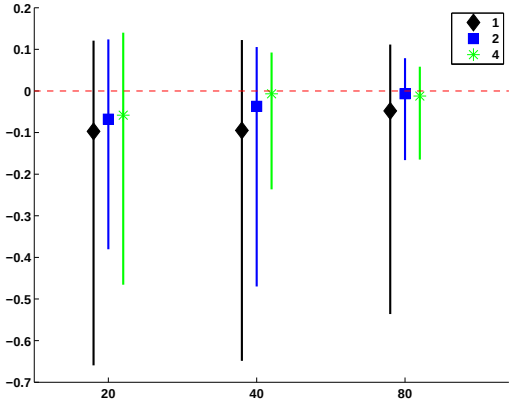


Figure 5.10: The differences in MLPD between MAP-II and CCD approaches in Poisson observation model with parameters $E_i = \{1, 2, 4\}$ and $N = \{20, 40, 80\}$. The means and the 95% confidence intervals are shown.

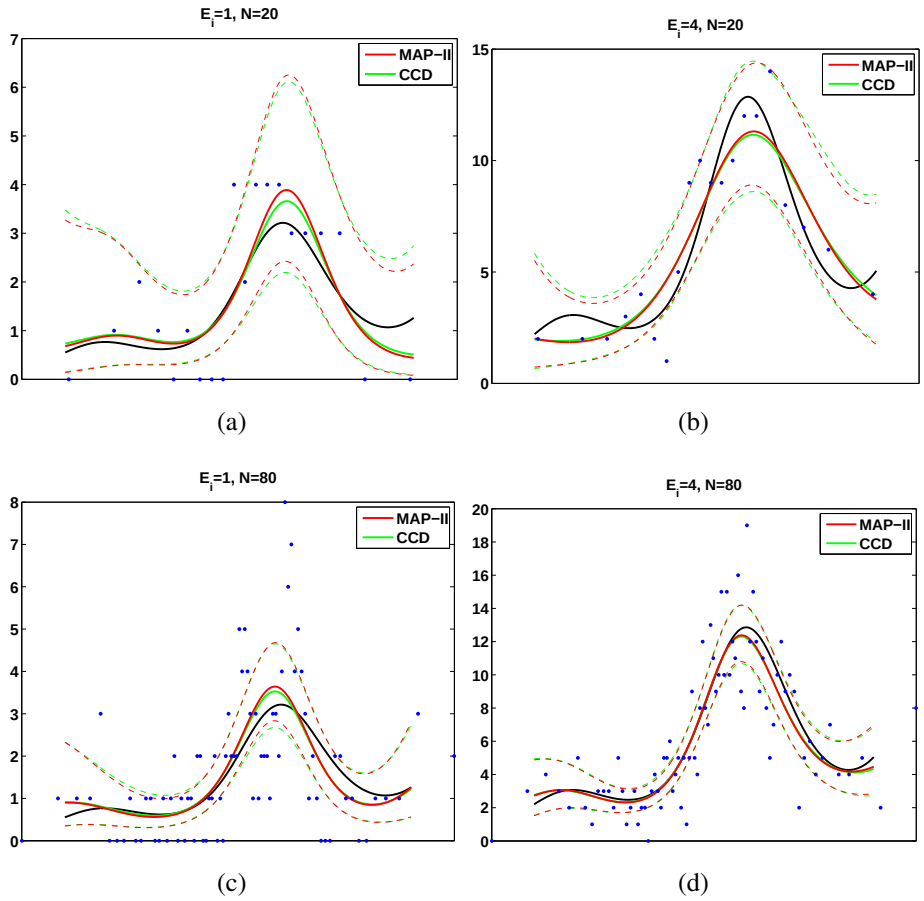


Figure 5.11: The observation and the predictions using MAP-II and CCD methods. The number or observations N and the expected number or occurrences E_i are varied. In figure (c), the green line is under the red one.

5.4 Regression with precipitation data

The data used in this experiment consists of the monthly precipitation measures recorded across the United States in 1995. The data consists of total 5776 stations, locations of which are shown in figure 5.12. 223 of the measurements were removed for testing (plotted with red dots in 5.12) and the rest were used for training.

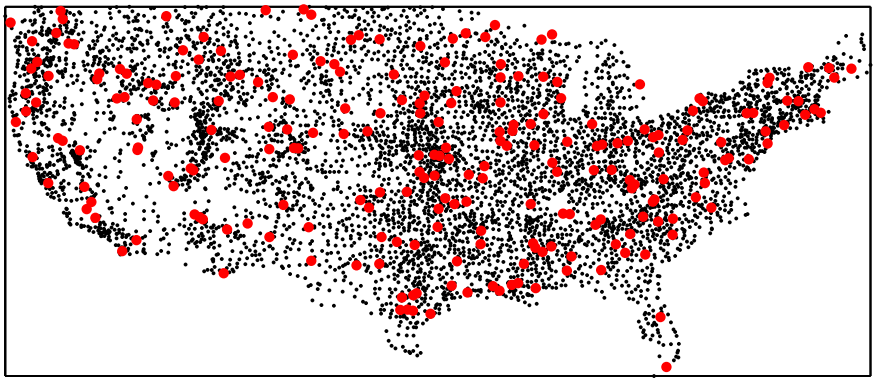


Figure 5.12: Locations of training data (black dots) and test data (red dots).

The aim of this test is to demonstrate that the size of the training data set affects the difference between results using MAP estimate and integration approximation. Subsets of different sizes were sampled from the training data set, and those were used for the training. In order to reduce the effect of the random sampling, results were averaged over 150 subsets.

Vanhatalo and Vehtari (2008) demonstrated that this data set contains both short and long length scale phenomena. Therefore, a GP with two squared exponential covariance functions is used. The prior for the hyperparameters was Student-t with a scale of 36 and one degree of freedom. Vanhatalo and Vehtari (2008) showed that other GP structures explain the data better, the aim of this experiment is to study the difference between MAP-II estimate and integration approximation.

Figure 5.13 shows the estimated precipitation using the whole data set. It suggests that the variation differs in North-South and East-West directions. Therefore, both input dimensions were given individual length scales for both of the covariance functions. Thus, the GP has seven hyperparameters. The high number of hyperparameters would make Grid search computationally demanding. Therefore only CCD approximation was used.

Figure 5.14 summarizes the differences in MSE between MAP-II and CCD methods using

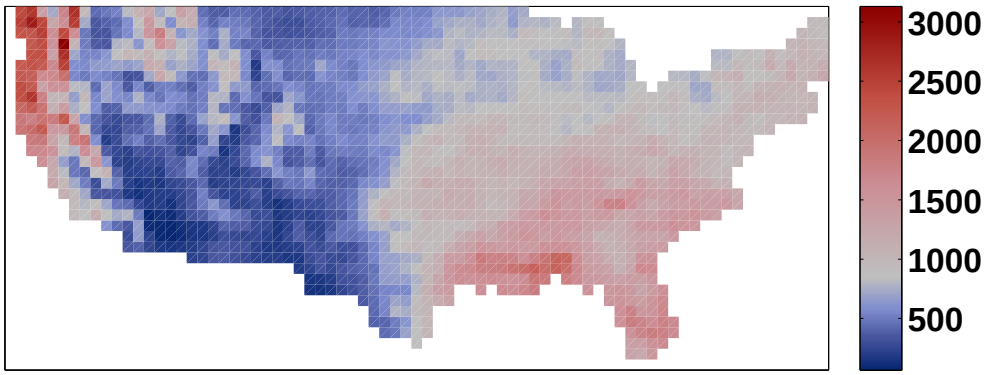


Figure 5.13: Annual precipitation in the US estimated from the full training data set. Reproduced from (Vanhatalo and Vehtari, 2008)

training sets of different sizes. The mean of the differences is positive with all sizes of training data sets, but the difference and the confidence interval diminish as the size grows. This indicates that there is a small difference in the fits with these two methods in favor of integration method. However this vanishes as the size of the training data grows.

Figure 5.15 shows similar comparison for MLPD measures. Negative difference shows that CCD has better fit in predictive distribution. However, the difference between methods also vanishes as the amount of training data increases. Figure 5.16 shows that the predictive distributions become almost identical as the amount of training data increases, because the KL-divergence from predictive distributions of CCD to those of MAP-II goes to zero.

The small size of the training data set makes the distances from the predictions to the training data large. It was shown in previous sections that the difference between the methods was at largest when there is no training samples near the predictions.

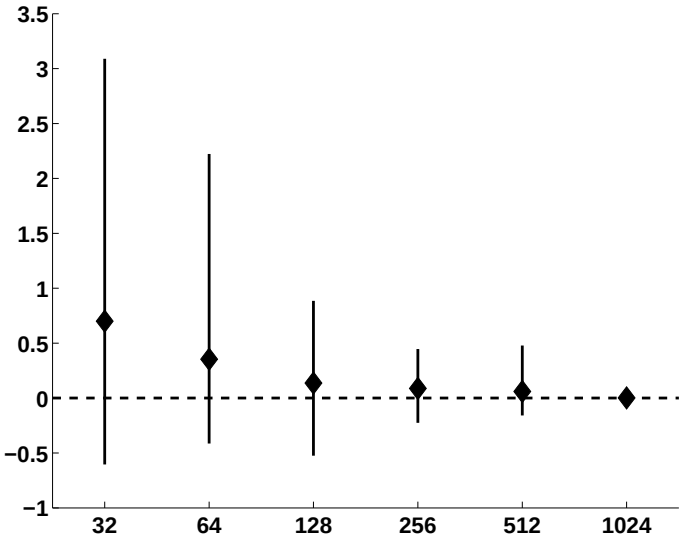


Figure 5.14: The means of the difference in MSE between MAP-II and CCD methods with different sizes of training data sets. Confidence intervals of 95% also shown.

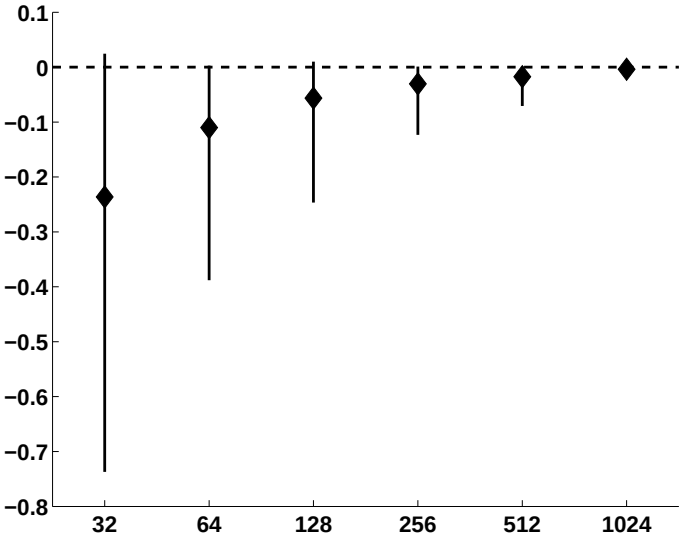


Figure 5.15: The means of the difference in MLPD between MAP-II and CCD methods with different sizes of training data sets. Confidence intervals of 95% also shown.

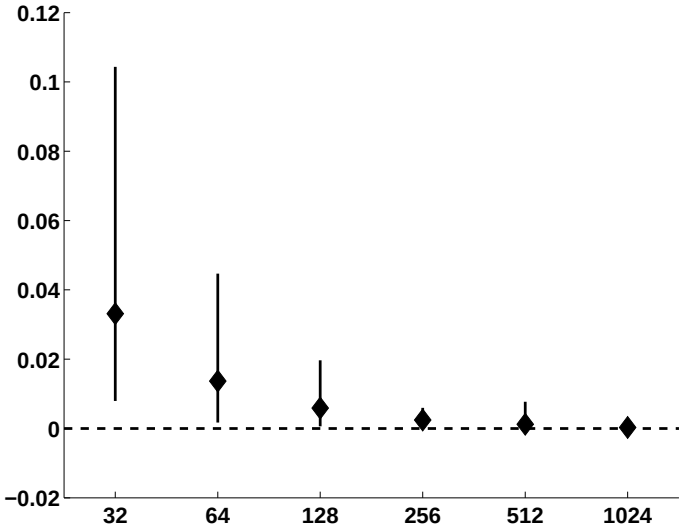


Figure 5.16: Means of KL divergence from predictions of CCD to those of MAP-II method with different sizes of training data sets. Confidence intervals of 95% also shown.

Chapter 6

Conclusion and future work

The aim of this thesis was to study the effect of the integration over the hyperparameters of Gaussian processes. The integrated predictions were compared to those obtained using MAP-II estimate for the hyperparameters. The comparison was conducted via MSE and MLPD measures.

Results demonstrate that it can be advantageous to integrate over the hyperparameters instead of using only a point estimate. Although MAP-II estimate is faster to use than any of the presented methods, all approximations are computationally feasible with a low dimensional hyperparameter space. With many hyperparameters grid search can be too demanding, but CCD provides a computationally attractive approximation for the posterior of the hyperparameters.

However, the difference in both measures between the point estimate and integration approaches diminishes to negligible when the size of the training data set increases. Gaussian processes seem not to be very sensitive to the choice of the hyperparameters, if there is enough data present. However, the data sets used in this thesis were small in order to demonstrate the difference, whereas the real-world data sets are often significantly larger than those used in this thesis.

The difference between the methods was clearest when the distance from prediction to the training samples was large. MAP-II estimate appeared to underestimate the variance in the predictions. Therefore integration approaches should be considered, if the density of the training data is low near the predictions, even if the size of the training data set is large.

The significance of the difference in any measure always depends on the application and the unit of the measure. For example, a benefit of 1 euro might not be a reason to use an integration method, whereas 1 million euros could be one. In addition, if the computation time is very expensive or limited, MAP-II estimate might be a more reasonable choice.

All the methods presented in this work are implemented in Matlab. Only the operations relating to the integration over the hyperparameters were programmed and GPstuff toolbox for Matlab (Vanhatalo et al., 2008) was used for other calculations. These integration methods will be published as a part of GPstuff toolbox in the future.

The integration over a multimodal posterior distribution is not automated in the current implementations of these methods. The integration over two modes in section 5.1 was conducted manually. One possible solution could be to use a mixture of Student-t distribution as a proposal distribution. Then the overlapping of the individual approximations would not be a problem and the integration could be automated. This could be one future improvement to the present methods. Also it might be of interest to study whether the parameters of these approximations (such as, $\delta_z, \delta_\pi, f_0$) have significant effect on the performance of the methods. The optimization of these parameters could be studied.

As a conclusion, the integration approximations were shown to be potentially useful, but the benefit of these methods might not always be significant enough to overcome the computational disadvantages. The choice of using only MAP-II estimate should be validated by comparing predictions of MAP-II method to the ones obtained using integration methods.

Bibliography

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover.

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.

Bishop, C. M. (1996). *Neural Networks for Pattern Recognition*. Oxford University Press, USA, 1 edition.

Coleman, T. F. and Li, Y. (1996). An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6:418–445.

Cornford, D., Nabney, I., and Williams, C. K. I. (2002). Modelling frontal discontinuties in wind fields. *Journal of Nonparametric Statistics*, 14(1-2):45–58.

Davidon, W. C. (1991). Variable metric method for minimization. *SIAM Journal on Optimization*, 1(1):1–17.

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

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339.

Gilks, W. R. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall, London.

Hammersley, J. M. (1960). Monte Carlo methods for solving multivariable problems. *New York Academy Sciences Annals*, 86:844–874.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. *Uncertainty in Artificial Intelligence*, 17:362–369.

Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science.

Neal, R. M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report 9702, Department of Statistics, University of Toronto.

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

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

Sanchez, S. M. and Sanchez, P. J. (2005). Very large fractional factorial and central composite designs. *ACM Trans. Model. Comput. Simul.*, 15(4):362–377.

Spiegelhalter, S. D., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Vanhatalo, J. et al. (2008). *Gaussian process models for Bayesian analysis (for Matlab) V1.1.0*. http://www.lce.hut.fi/research/mm/gpstuff/. Accessed December 21st, 2009.

Vanhatalo, J., Jylänki, P., and Vehtari, A. (2009). Gaussian process regression with Student-t likelihood. In *Advances in Neural Information Processing Systems 23*, pages 1910–1918.

Vanhatalo, J. and Vehtari, A. (2008). Modelling local and global phenomena with sparse Gaussian processes. In *UAI*, pages 571–578.

Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468.

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