Date of acceptance Grade

Instructor
Jarno Vanhatalo

Aspects of reparametrization in Gaussian process regression with the Weibull model.

Ville Tanskanen

Helsinki November 19, 2018

Master's Thesis

UNIVERSITY OF HELSINKI
Department of Mathematics and Statistics

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Mathematics and Statistics

Tekijä — Författare — Author

Ville Tanskanen

Työn nimi — Arbetets titel — Title

Aspects of reparametrization in Gaussian process regression with the Weibull model.

Oppiaine — Läroämne — Subject

Mathematics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's Thesis	November 19, 2018	67 pages

Tiivistelmä — Referat — Abstract

Gaussian processes can be used through Bayesian models so that they are formed through a multi-dimensional Gaussian prior with a special covariance matrix structure and an arbitrary likelihood model. They often include a latent variable structure between the features and the response variable. Bayesian modeling's drawbacks are usually related to the normalizing constants that normalize the product of a prior probability density function and a likelihood function to a proper probability distribution. These integrals are hard or even impossible to calculate analytically and hence some approximations are required.

One popular approximation is the Laplace approximation, which is a Gaussian approximation for the unnormalized log-posterior distribution. Reparametrization of the observation model can lead to changes in properties of the posterior distribution such as shape and convergence. The performance of approximations made for the posterior distribution also change along with the parametrization. The changes are often related to either computational complexity or the predictive performance of the approximation.

This thesis presents the Gaussian processes starting from Bayes' formula and moves quickly towards key concepts in Bayesian modeling such as predictive distributions and hierarchy. An approximation of interest for the posterior distribution, the Laplace approximation, is derived. Traditional optimization algorithm for the Laplace approximation is the Newton method, which is replaced by an algorithm called natural gradient adaptation in this thesis. Then the focus is turned from general introduction of Gaussian processes to more specific treatment of them by choosing the Weibull distribution as an observation model. Two different parametrizations for the Weibull model are studied, one which acts as a baseline and can be thought as traditional parametrization for the model, and another one for which the parameters are orthogonal. The predictive performance of the Laplace approximation is then compared within the two parametrizations in two different kind of data sets. Finally the results show decrease in computation time required for the Laplace approximation but no improvement in the predictive performance for orthogonal parametrization.

Avainsanat — Nyckelord — Keywords

Bayesian modeling, Gaussian process, reparametrization, Weibull distribution, Laplace approximation

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Contents

1 Introduction **1**

2 Gaussian process **3**

2.1 Fitting a Gaussian process 6

2.2 Approximations 8

2.2.1 Laplace approximation 8

2.2.2 MCMC 12

2.3 Predictions 13

2.3.1 Laplace predictions 15

2.3.2 MCMC predictions 16

2.4 Adaptation of hyperparameters 17

2.5 Details of technical implementation 20

3 Gaussian processes in scale and shape varying Weibull models **22**

3.1 Weibull distribution 23

3.2 Reparametrization 24

4 Experiments **27**

4.1 Data 27

4.1.1 Hyperparameters for data 28

4.1.2 Smooth data 29

4.1.3 Step data 29

4.2 Methods to validate and compare the goodness of the posterior approximations 30

4.2.1 Mean and covariance differences 32

4.2.2 Kullback-Leibler divergence from the true conditional posterior to the Gaussian approximations 33

4.2.3 Mapping reparametrized posterior back to the original space . 36

4.2.4 Computing times 36

4.2.5 Convergence of the MCMC chains 37

5 Results **37**

5.1 Smooth data 37

5.2 Step data 44

6 Discussion **47**

7 Conclusions **52**

Appendixes **54**

A Fisher information for Weibull distribution 54

B Predictive distribution given $\boldsymbol{f}$ 58

C Predictive statistics given $\boldsymbol{y}$ 62

D Kullback-Leibler divergence between two multivariate Gaussian distributions 63

1 Introduction

Bayesian modeling through Gaussian processes (GP) is a widely used method in statistics. It provides a excellent, flexible framework for probabilistic and smooth interpolation of a noisy data through latent variable framework. In practice Gaussian processes often lead to difficult inference tasks that are analytically impossible, which is why an approximative inference is a necessity.

Gaussian approximation is a well studied approximation, that is justified with large sample theory, more accurately central limit theorem (Durrett, 2011). The central limit theorem says that the sum of random variables is approaching a Gaussian distribution, regardless of the distribution of the random variables that are being summed, as long as they are independent, have finite variance and similar enough distributions in terms of moments higher than two. The details of different variations of central limit theorem can be found in Billingsley (1995).

Large sample behavior and its theory has interested statisticians for decades in order to justify results and approximations for addressing the real-world problems (de Moivre, 1738). The digitalizing world generates data tirelessly and the data is captured in increasing amounts. The amount of samples is therefore practically approaching the infinity.

The theory of large sample statistics is often about answering the question whether the convergence of a statistic happens or not, and if it does, then where is the statistic converging to. However, the rate of convergence is less often addressed, but in practice it is often as important aspect of statistics as the convergence itself.

Bayesian modeling has been increasingly popular from the 1980's as the computational power increased to the point that scientists could actually start estimating the hard integrals arising from the normalization of the posterior, see Robert & Casella (2011) for a brief history of Markov chain Monte Carlo (MCMC) methods. Bayesian models have gained attention due to their ability to produce probabilistic predictions about the future as well as the ability to include prior information in mathematical form to the inference.

The large sample theory of Bayesian statistics states, that all posterior distributions, formed through a prior distribution and a likelihood function, converge towards a Gaussian distribution, when the observed samples are independently and identically distributed, see for example appendix B from Gelman et al. (2013). This result only applies when the dimension of the posterior remains constant as more data is observed. In Gaussian processes the dimension of the posterior remains constant only in some special situations. Theoretically more suitable discussion for Gaussian processes is presented for example in Rue et al. (2009), as it considers the case where the number of parameters grow with the number of observations. The convergence result of posterior normality holds, even if the number of parameters tend to infinity, if the posterior density is converging to a degenerate Gaussian density. This is indeed the case in many typical Gaussian process models. The mathematical details of posterior and prior distribution along with the definition of observation model

and likelihood function are presented later.

The prior in Gaussian processes is by assumption normally distributed. The posterior distribution in Gaussian processes is analytically normally distributed, only if the observation model (Gelman et al., 2013) is also normally distributed.

When the observation model is not normally distributed, the posterior will not be normal, but will still approach it, as the number of observations grows. The observation model can be reparametrized in multiple different ways. When the parametrization is chosen so that the observation model is as close to a Gaussian distribution as it can be, it is natural to think that also the convergence of posterior distribution towards a Gaussian distribution is faster.

Motivated by the above discussion, I study the effect of two different parametrizations of Weibull distribution observation model in Gaussian process regression to the rate of convergence of the posterior towards the normal distribution. The two parametrizations that are discussed are referred to original and orthogonal parametrizations. Orthogonal, in the sense of the parameter space being equipped with diagonal Fisher information matrix. The hypothesis is that the orthogonal parametrization of the Weibull model will lead to faster convergence towards Gaussian distribution (Hartmann & Vanhatalo, 2018) for which the intuition is given in Cox & Reid (1987).

Sampling of the analytical posterior distribution via Markov chain Monte Carlo methods can be very time consuming, as the dimension of the posterior can be high, or even impossible due to the lack of convergence of the chains. When a closed form approximation of the posterior is used, the computational burden reduces. The accuracy of the closed form approximation changes by the parametrization and the optimal goal is to find a parametrization that reduces the computational burden and is still an accurate approximation.

This thesis relies heavily on the results of probability calculus and probability theory. I assume that the reader has at least the basic knowledge of probability calculus introduced, for example, in Stirzaker (2003). The deeper understanding of the concepts will require knowledge from more theoretical works from for example Jaynes et al. (2003), Billingsley (1995) and Durrett (2011). To understand the proofs, the reader needs to be excellent on matrix algebra (Gentle, 2007) as well as (matrix) calculus (Petersen & Pedersen, 2012). Also elementary knowledge of differential equations (Tenenbaum & Pollard, 2012), and willingness to generalize it to partial differential equations, is very helpful.

The rest of the thesis is organized as follows. Second chapter introduces the reader to theoretical results of Gaussian processes in general context. Third chapter ties the theory to practice through an example by choosing the Weibull distribution as an observation model for the Bayesian model. Fourth chapter explains the experiments that were done in order to compare the parametrizations. Fifth chapter presents the results from the experiments described in the fourth chapter. Sixth chapter discusses the results and debates their meaning and correctness in a self-critical way. Seventh chapter outlines the thesis' research question and results.

Bayesian details and data for supervised learning The Bayesian modeling and statistics are strongly binded to the Bayes' theorem

$$p(\phi|y) = \frac{p(\phi)p(y|\phi)}{p(y)}, \tag{1}$$

where ϕ and y are random variables with respective density functions $p(\cdot)$. The Equation (1) is the foundation of Bayesian modeling and inference, and it will be specified more carefully for the parameters of Gaussian processes in the upcoming sections. For now, ϕ is a parameter for which there is uncertainty and y is a fixed vector of values observed from the random variable Y. Apart from the theorem, it is also a philosophical perspective where the uncertainty of the parameter is formulated through probabilities. This kind of uncertainty is often defined as epistemic in literature (O'Hagan, 2004). Another class of uncertainty is aleatory uncertainty, which refers to the uncertainty of an event by nature, which y is a realization of, for example result of a set of coin tosses. The goal of Bayesian inference is to reduce one's subjective uncertainty about the parameters by observing data. The uncertainty is subjective, because the prior distribution $p(\phi)$ is always a subjective choice of the modeler. The observation model $p(y|\phi)$ is also subjective description of the stochastic phenomena, but it is often less controversial. Discussion on the existence of these two uncertainties and their relationship is not treated in this thesis and instead, the theory of Bayesian modeling in the context of Gaussian processes is presented and studied.

The data in *supervised machine learning* is often formulated in the feature matrix $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ and in a response vector $\boldsymbol{y} \in \mathbb{R}^N$. The matrix $\boldsymbol{X}$ contains a row for each observation of the features observed in the data. A feature is something that describes one characteristic of that data point. The response variable $\boldsymbol{y}$ usually contains one response value for each data point. However, vector $\boldsymbol{y}$ can also be a response matrix, if more than one characteristic of a data point is the subject of inference or prediction.

The notation of a feature matrix $\boldsymbol{X}$ and a response variable vector $\boldsymbol{y}$ will be used throughout the thesis. This thesis is about supervised learning, which is a subfield of machine learning. For readers not familiar with machine learning, an introductory treatment of it is given by James et al. (2014).

2 Gaussian process

Gaussian process is a powerful statistical model, whose power lies in a covariance function that uses distance measures to address a continuous scale of independence within the observed data. First, the Gaussian process is presented in general context, which means that no assumptions on the observation model has been made. Afterwards a more applied approach will be taken, beginning from Section 3, where the observation model will be defined.

Definition 2.1. A collection of random variables is called *Gaussian process* if any finite linear combination of those random variables follow a Gaussian distribution, i.e. has a multivariate Gaussian distribution.

This definition provides a very general description of the Gaussian process. However, it is equivalent, and usually more practical to provide valid mean and covariance functions for a collection of random variables $\boldsymbol{f}(\boldsymbol{x})$. Validity here means that for any discrete set of locations $\boldsymbol{X}$, the covariance function produces a positive semi definite covariance matrix, and that the mean function remains finite. Note that this definition and the following notation follow closely the book by Rasmussen & Williams (2005). The mean and covariance functions are denoted by

$$\begin{aligned} m(\boldsymbol{x}) &= \mathbb{E}[f(\boldsymbol{x})] \\ k(\boldsymbol{x}, \boldsymbol{x'}) &= cov[f(\boldsymbol{x}), f(\boldsymbol{x'})] = \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))^T] \end{aligned}$$

respectively. Thus the notation

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

means that $f(\boldsymbol{x})$ follows a Gaussian process with a mean function $m(\boldsymbol{x})$ and a covariance function $k(\boldsymbol{x}, \boldsymbol{x'})$. In the literature, the mean function is often taken to be zero to simplify the notation. The mean is also assumed to be zero in this thesis. An example of widely used covariance function (Rasmussen & Williams, 2005) is the squared exponential

$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right), \tag{2}$$

which offers a smoothly varying prior for the latent vector $\boldsymbol{f}$. The variables σ^2 and l act as hyperparameters. The hyperparameters in the case of the squared exponential covariance function control the rate and the magnitude of fluctuation of the latent function. An example of the effect of the the hyperparameters on the prior density can be seen in Figure 1, where 9 plots show draws from a Gaussian process prior with length scale and sigma varying within the Cartesian product $\{0.1, 0.5, 1\} \times \{0.1, 1, 5\}$. More careful treatment of how the hyperparameters can be chosen is presented in Section 2.4.

The covariance function transforms the input locations $\boldsymbol{X}$ into a covariance matrix. A notation $K(\boldsymbol{X}, \boldsymbol{X})$ is used to emphasize the covariance matrix that is formed by calculating the covariance between the rows of $\boldsymbol{X}$ according to a given covariance function to produce an $N \times N$ matrix, where N is the number of rows in $\boldsymbol{X}$. The shorter notation used for this is just K. A notation $K(\boldsymbol{X}, \boldsymbol{X_*})$ is abbreviated as k_*, and it stands for an $N \times N_*$ matrix, where N_* is the row count of the matrix $\boldsymbol{X_*}$. Continuing the above pattern $K(\boldsymbol{X_*}, \boldsymbol{X_*})$, abbreviated as K_{**}, stands for the same thing as $K(\boldsymbol{X}, \boldsymbol{X})$, but it is calculated between the rows of the matrix $\boldsymbol{X_*}$.

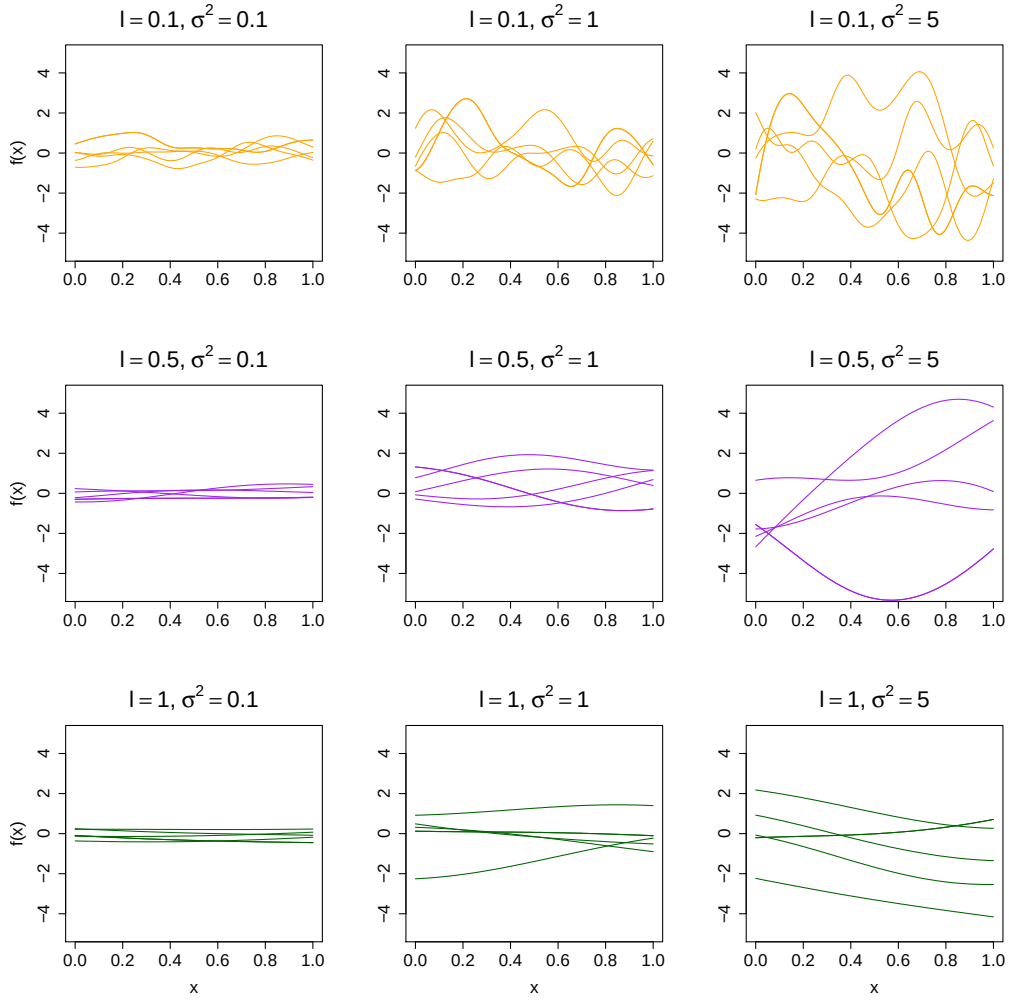


Figure 1: An example of hyperparameters' effect on the latent process. The figure shows a grid of plots where on each plot 5 latent vectors are drawn from a Gaussian process equipped with squared exponential covariance function. Notice how σ^2 controls the magnitude of the variation in the $f(x)$ axis and the length parameter controls the rate of variation in the x axis.

The meaning of $\boldsymbol{X}_*$ is to hold the features of the samples that are in the interest of predictions.

Modeling with Gaussian process is much like any other supervised machine learning problem. First, the model is fitted, after which the conclusions about the phenomena can be drawn or unseen data points can be predicted. Again, the feature matrix is denoted by $\boldsymbol{X}$ and the response variable as $\boldsymbol{y}$, following the same notation introduced in Chapter 1. Often, the observation model depends on more than one parameter. In Gaussian process each of the parameters is assumed to follow a Gaussian process prior distribution. For this reason, a model that has m parameters, also has m Gaussian process priors. The parameters for which the priors are given are denoted by $\boldsymbol{f}(\boldsymbol{X}) = [\boldsymbol{f}_1(\boldsymbol{X}), \ldots, \boldsymbol{f}_m(\boldsymbol{X})]$ and their uncertainty is being reduced by Bayes theorem. For notational reasons the function argument $\boldsymbol{X}$ is left out and the set of m latent vectors $\boldsymbol{f}(\boldsymbol{X})$ will be denoted by $\boldsymbol{f}$. Note that $\boldsymbol{f}$ is formed by stacking together vectors $\boldsymbol{f}_i = [f_i(\boldsymbol{x}_1), \ldots, f_i(\boldsymbol{x}_N)]$ where $i \in \{1, \ldots, m\}$ and again for notational purposes the arguments of the functions are dropped so that the ith vector of $\boldsymbol{f}$ will be denoted as $\boldsymbol{f}_i = [f_{i1}, \ldots, f_{iN}]$. Finally, the vector of all latent vector values spread out in one long vector is denoted by $\tilde{\boldsymbol{f}} = [f_{11}, \ldots, f_{1N}, \ldots, f_{m1}, \ldots, f_{mN}]$.

Each response variable y_i is assumed to be independent from other response variables $\{y_j\}_{j \neq i}$ given the parameters $[f_{1i}, \ldots, f_{mi}]$. See Figure 2 where the variables that are known are filled with gray color, and unknown variables are left white. The bold line represents full dependence between all variables connected by the line and arrows indicate conditional dependence to the direction of the arrow. For example f_{11} depends on $\boldsymbol{x}_1$ but not the other way around, and f_{12} is dependent on every latent vector value $\{f_{1i}\}_{i \neq *}$ and the corresponding $\boldsymbol{x}_2$. Notice that the point of predictive interest is denoted by $\boldsymbol{x}_*$ whose treatment is covered in Section 2.3.

2.1 Fitting a Gaussian process

The terminology between statisticians and professionals coming from machine learning field cross in many cases. Sometimes one meaning has two terms and sometimes they are interchanging the terms and meanings. In this thesis the terminology tries to follow the machine learning literature. The inference in Gaussian processes usually refers to the inference made from the posteriors of the parameters $\boldsymbol{f}$ and hyperparameters $\boldsymbol{\theta}$. This terminology is somewhat common for both of the fields. In statistics inference is sometimes used for two other meanings. Prediction, where the inference is about the future observations, and to describe "the process of discovering the posterior distribution". In this thesis the inference term is separated from prediction and posterior fitting, and the meaning for it is about the conclusions made from some parameter values, once they have been fitted. The inference part is only mentioned here, as a possible application of Gaussian processes, where some important conclusions can be made about the phenomena under research from the posterior distribution of the parameters or the hyperparameters. In this thesis we are mostly interested in predictions.

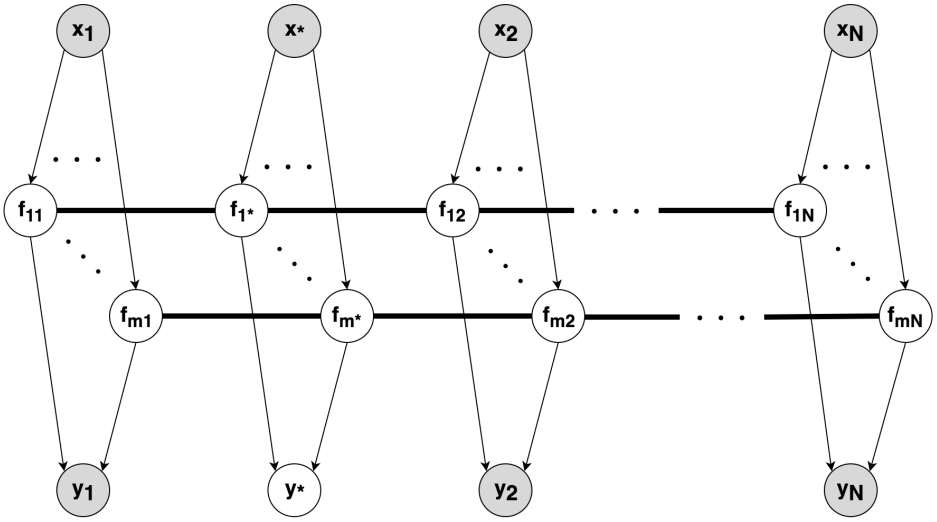


Figure 2: Plate diagram of training set $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ and a point of predictive interest $\boldsymbol{x}_*$ to visualize the dependencies of Gaussian process. Thick black lines denote dependence on both directions whereas arrows denote dependence only on the direction of the arrow.

Fitting a Gaussian process to a dataset is equal to finding the posterior distribution of latent vector $\boldsymbol{f}$ conditioned on the observations $\boldsymbol{y}$ and $\boldsymbol{X}$. The posterior density for the latent vector is given by Bayes formula

$$p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{f}|\boldsymbol{X})p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})}, \tag{3}$$

where $p(\boldsymbol{f}|\boldsymbol{X})$ is the prior information, $p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X})$ is the observation model, often stated as a likelihood when considered as a function of $\boldsymbol{f}$, and $p(\boldsymbol{y}|\boldsymbol{X})$ is the marginal likelihood. Note that hyperparameters $\boldsymbol{\theta}$ are left out from the notation and they could be included by conditioning everything on $\boldsymbol{\theta}$. Hyperparameters and their importance and selection is discussed later on in the chapter 2.4. For now they are just parameters that characterize the covariance (and, if desired, the mean) function.

After the posterior distribution (3) is found it can be used for analysis of the phenomena and prediction of the latent value $\boldsymbol{f}_*$ of the new data points $\boldsymbol{x}_*$. Finding the posterior is not trivial because there are only rare situations when the posterior matches some known distribution. The cause of this is usually the marginal likelihood

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int p(\boldsymbol{f}|\boldsymbol{X})p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X})d\boldsymbol{f}, \tag{4}$$

which is analytically intractable and one needs to rely on different kind of approxi-

mations for the posterior. Two possible approximations are Laplace approximation and MCMC sampling and they will be discussed later on in chapters 2.2.1 and 2.2.2.

Modeling with Gaussian process is all about setting a Gaussian process prior for the latent vector $\boldsymbol{f}$, where the dependence of them is determined by the covariance function. Also assumptions of the conditional independencies between $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{f}$ explained in Figure 2 are required. Rest of the modeling, and mathematically more challenging part, arises from the choice of the observation model. Different kind of data require different observation models. In Rasmussen & Williams (2005) a distinction between classification and regression model is presented. However there are situations when the above distinction is not sufficient, for example semi-bounded regression. The semi-bounded regression term is introduced for a regression problem where the response variable is bounded from one end. In this thesis the response is bounded to the range of $[0, \infty)$. In fact, the optimal approach would be that any observation model could be used. Different likelihoods require different kind of treatment, for example in the Laplace approximation the first and second derivatives of the log-likelihood $\log p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X})$ are required, and they vary for each likelihood function. Thus, the derivatives are to be solved for each desired likelihood separately, either analytically, or by using some numerical differentiation methods. In this thesis, Weibull observation model is used as an example of a slightly more challenging, semi-bounded regression model. Note that the observation model varies depending on the application, and its only requirement is to be able to output a probability or density of $\boldsymbol{y}$ as a function of $\boldsymbol{\theta}$ and $\boldsymbol{X}$.

2.2 Approximations

When the observation model is Gaussian, the posterior of the latent function can be solved analytically as shown in Appendix B, where the treatment is for prediction, but it is equivalent to the posterior when predicting at the observed points $\boldsymbol{X}$. This result will be useful especially in Laplace approximation where the posterior is approximated with a Gaussian distribution.

When the observation model is not Gaussian, the Gaussian process prior will not be conjugate to the likelihood. In those situations the normalizing constant will be analytically intractable. For this reason, approximations are needed.

2.2.1 Laplace approximation

Laplace approximation is a Gaussian approximation for the posterior. It approximates the posterior distribution with the second order Taylor expansion of the

log-posterior $\log(p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}))$ around the maximum a posteriori (MAP) estimate $\hat{\mathbf{f}}$

$$\begin{aligned}\log p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}) &\approx \log p(\hat{\boldsymbol{f}}|\boldsymbol{X}, \boldsymbol{y}) + (\boldsymbol{f} - \hat{\boldsymbol{f}})^T \overbrace{\nabla_{\boldsymbol{f}} \log p(\hat{\boldsymbol{f}}|, \boldsymbol{y})}^{=0} \\ &\quad -\frac{1}{2}(\boldsymbol{f} - \hat{\boldsymbol{f}})^T \overbrace{(-(\nabla\nabla)_{\boldsymbol{f}} \log p(\hat{\boldsymbol{f}}|\boldsymbol{X}, \boldsymbol{y}))}^{:=A}(\boldsymbol{f} - \hat{\boldsymbol{f}}) \\ &= \log p(\hat{\boldsymbol{f}}|\boldsymbol{X}, \boldsymbol{y}) - \frac{1}{2}(\boldsymbol{f} - \hat{\boldsymbol{f}})^T A(\boldsymbol{f} - \hat{\boldsymbol{f}}).\end{aligned} \tag{5}$$

The fact that the derivative of a function is zero at the critical points was used above. From the above equation it follows that

$$\begin{aligned}p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}) &\approx p(\hat{\boldsymbol{f}}|\boldsymbol{X}, \boldsymbol{y}) \exp(-\frac{1}{2}(\boldsymbol{f} - \hat{\boldsymbol{f}})^T A(\boldsymbol{f} - \hat{\boldsymbol{f}})) \\ &\propto \exp(-\frac{1}{2}(\boldsymbol{f} - \hat{\boldsymbol{f}})^T A(\boldsymbol{f} - \hat{\boldsymbol{f}})),\end{aligned}$$

from which it is easy to recognize the kernel of Gaussian distribution $\mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}}, A^{-1})$. Since the kernel defines a distribution uniquely it can be concluded, that the posterior is approximately normally distributed with mean $\hat{\mathbf{f}} \coloneqq \arg\max_{\mathbf{f}} \mathbf{p}(\mathbf{f}|\mathbf{y}, \mathbf{X})$ and covariance of $A^{-1} \coloneqq -\left((\nabla\nabla)_{\boldsymbol{f}} \log(p(\hat{\mathbf{f}}|\mathbf{X}, \mathbf{y})\right)^{-1}$. The Laplace approximation is then denoted as

$$\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X} \dot{\sim} \mathcal{N}(\boldsymbol{f}|\hat{\boldsymbol{f}}, A^{-1}), \tag{6}$$

where the symbol $\dot{\sim}$ translates to "is approximately distributed as".

When searching for the parameters $\hat{\boldsymbol{f}}$ and A for the Gaussian distribution given by Laplace approximation, the normalizing constant $p(\boldsymbol{y}|\boldsymbol{X})$ from the posterior distribution $p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X})$ can be forgotten, as it will not change the result of the maximization. Taking the logarithm of the unnormalized posterior will not change the location of the maximum but it will simplify the mathematical operations, which is why it is used. The unnormalized log-posterior is denoted by

$$\begin{aligned}\Psi(\boldsymbol{f}) &\coloneqq \log p(\boldsymbol{y}|\boldsymbol{f}) + \log p(\boldsymbol{f}|\boldsymbol{X}) \\ &= \log p(\boldsymbol{y}|\boldsymbol{f}) - \frac{n}{2}\log(2\pi) - \frac{1}{2}\log|K| - \frac{1}{2}\boldsymbol{f}^T K^{-1}\boldsymbol{f}.\end{aligned} \tag{7}$$

For finding $\hat{\boldsymbol{f}}$, the gradient is often required, which is achieved by differentiation

$$\nabla_{\boldsymbol{f}} \Psi(\boldsymbol{f}) = \nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\boldsymbol{f}) - K^{-1}\boldsymbol{f}. \tag{8}$$

The second derivative of the unnormalized posterior

$$\nabla\nabla_{\boldsymbol{f}} \Psi(\boldsymbol{f}) = \nabla\nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\boldsymbol{f}) - K^{-1} = -W - K^{-1} \tag{9}$$

is required for the approximative covariance as well as for some optimization algorithms such as Newton method presented below. Notice that the above calculation defined $W \coloneqq -\nabla\nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\boldsymbol{f})$ which is also known as the negative Hessian matrix of the likelihood function.

In majority of the Bayesian modeling approaches it is essential to find the maximum of the posterior distribution. The Laplace approximation requires the MAP estimate of the posterior distribution as well. The desire to find the MAP estimate relates to the fact that unimodal distributions' probability masses are centered around their MAP estimates. The optimization of a function, in this case the posterior, is usually approached with gradient based methods, such as gradient ascent, Newton method and conjugate gradient (Fletcher & Reeves, 1964), when the function is assumed to be smooth. The standard implementation of finding the MAP estimate $\hat{\boldsymbol{f}}$ of the latent vector, given in Rasmussen & Williams (2005), uses the Newton method. Newton method is like gradient ascent, but with an optimal step size determined by the inverse of second derivative of negative log posterior density and thus connecting the information of second derivatives between the parameters.

The update rule for the Newton iteration is

$$\begin{aligned}
\boldsymbol{f}^{new} &= \boldsymbol{f} - (\nabla\nabla_{\boldsymbol{f}}\Psi(\boldsymbol{f}))^{-1}\nabla_{\boldsymbol{f}}\Psi(\boldsymbol{f}) \\
&= \boldsymbol{f} + (K^{-1} + W)^{-1}(\nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\boldsymbol{f}) - K^{-1}\boldsymbol{f}) \\
&= (K^{-1} + W)^{-1}(K^{-1} + W)\boldsymbol{f} + (K^{-1} + W)^{-1}(\nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\boldsymbol{f}) - K^{-1}\boldsymbol{f}) \\
&= (K^{-1} + W)^{-1}((K^{-1} + W)\boldsymbol{f} + \nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\boldsymbol{f}) - K^{-1}\boldsymbol{f}) \\
&= (K^{-1} + W)^{-1}(W\boldsymbol{f} + \nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\boldsymbol{f})). \qquad (10)
\end{aligned}$$

By knowing the step for one iteration, it is possible to start with an initial guess of $\hat{\boldsymbol{f}}_0$ and iterate the above updating scheme until convergence. The covariance matrix A^{-1} for Laplace approximation ends up in its final form $A^{-1} = (K^{-1} + W(\hat{\boldsymbol{f}}))^{-1}$ by (9), where $W(\hat{\boldsymbol{f}}) = -\nabla\nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\hat{\boldsymbol{f}})$ emphasizes that the Hessian is evaluated at the $\hat{\boldsymbol{f}}$.

Notice that the Laplace approximation is not always sufficient. When the posterior is not unimodal the normal approximation performs poorly, as the example illustrates in Figure 3. It illustrates a simple example where Laplace approximation is used to approximate a mixture of two Gaussian distributions. Gaussian processes sometimes face multimodal posterior distributions, and if possible the multimodality should always be examined by some methods, for example MCMC convergence. In the example it is trivial to search for the maximum of the posterior $p(x)$, corresponding to $\hat{\boldsymbol{f}}$ in the above discussion, but in higher dimensions it might not be the case as every dimension increases the search space exponentially. This burden, often stated as the curse of dimensionality is the reason why smarter optimization methods than the grid search are required. A gradient based method called Newton method was introduced above. Overall maximizing or minimizing a function over a high dimensional search space is a well known task in machine learning that ties it closely to the field of optimization.

Note that a normal distribution might not approximate the posterior distribution

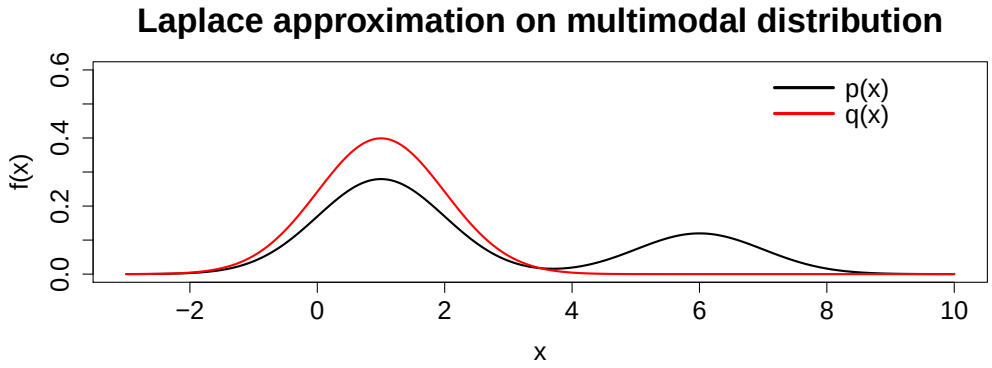


Figure 3: An example where Laplace approximation is not recommended. Here the distribution $p(x)$ that is being approximated is a multimodal mixture of two Gaussian distributions. The approximation $q(x)$ misses the probability mass of the second Gaussian distribution completely and thus the approximation will yield to overly confident approximation of the first peak.

very well, even if it was unimodal. Mismatch in skewness, kurtosis and higher moments can cause inaccurate predictions. It has been proposed, for example by Kass & Slate (1994) and MacKay (1998), that reparametrization of the observation model can make the model more accurate. More precise treatment of reparametrization will be given in Chapter 3.2.

Natural gradient adaptation The Newton method assumes that the space that is being explored is euclidean so that the inner product is defined as the standard dot product. When the space that is being explored is not euclidean the inverse of the second derivative of the log posterior (9) might not be the most effective direction to move, as the Newton method suggests. In those cases optimizing with the Newton method can lead to undesired result, slow convergence or computationally impossible situations.

To overcome the difficulties with the standard Newton method, a method called natural gradient adaptation is used. The natural gradient adaptation is a method that does not rely on the euclidean approximation of the parameter space, but encompasses information about the curvature of it through a matrix called Riemannian metric tensor. In statistical models, the Riemannian metric tensor is typically defined by the Fisher information (S.-I. Amari, 1998). While this is a nice result, it does not completely define the curvature in the space where the posterior density lies, as it does not contain the information about the curvature of the space of the prior.

Thus, the Riemannian metric tensor that encompasses the posterior density space's curvature is the expectation of the second derivative of the unnormalized log poste-

rior density

$$\begin{aligned}\mathbb{E}_Y\left[\nabla\nabla_{\boldsymbol{f}}\Psi(\boldsymbol{f})\right] &= -\mathbb{E}_Y\left[W\right] - \mathbb{E}_Y\left[K^{-1}\right] \\ &= -\mathbb{E}_Y\left[-\nabla\nabla_{\boldsymbol{f}}\log p(\boldsymbol{y}|\boldsymbol{f})\right] - K^{-1} \\ &= -\mathcal{I}(\boldsymbol{f}) - K^{-1}, \end{aligned} \tag{11}$$

where $\mathcal{I}(\boldsymbol{f})$ is called the Fisher information. Thus, the only difference to the standard Newton method is that we replace the negative Hessian matrix W in (10) with its expectation, that is, the Fisher information matrix $\mathcal{I}(\boldsymbol{f})$. The Fisher information matrix's elements are defined in (36), and an alternate form of them is also presented in Appendix A. Now the matrix $\mathcal{I}(\boldsymbol{f}) + K^{-1}$ which will be inverted, is positive definite by definition and will remove the computational problems related to the negative Hessian and slow convergence. Some motivation for using natural gradient can be found from S. Amari & Douglas (1998).

In practice, this transition requires some mathematical treatment as the elements of the Fisher information matrix are needed to be calculated. In the case of Weibull distribution these elements are derived in appendix A.

2.2.2 MCMC

Another, and more traditional way to approximate the posterior distribution is Markov Chain Monte Carlo (MCMC) methods. They are methods that do not assume anything about the modality of the distribution, and hence can be useful in multimodal cases, unlike the Laplace approximation described above.

MCMC methods are iterative and produce a chain of samples, whose sample frequencies approximate the posterior distribution as the number of samples drawn grows. Often there are multiple chains of samples generated to reduce the problem where a single chain converges to a local distribution, such as one of the modes in a multimodal distribution. Thus doing random restarts can help the sampling to avoid local distributions. Multiple chains also give better information about the convergence of the sampling overall, as then it is possible to compare the generated samples between the chains. Longer chains produce better approximations. A key assumption, called Markov property, is that the next element in the chain is only dependent on the current element. In this thesis the theory of MCMC methods is not presented and more comprehensive treatment of it can be found in Robert & Casella (2005). Stan software by Carpenter et al. (2017) is used for MCMC sampling in this thesis. The MCMC samples will be compared with the Laplace approximation and thus measure the difference between the true posterior predictive distribution, approximated by the Stan, and the posterior predictive distribution produced by the Laplace approximation.

Technical details Similar to this thesis' research question, the rate of convergence in MCMC sampling is an important factor. MCMC sampling produces sampled chains of posterior distribution, and each sample is generated only using the

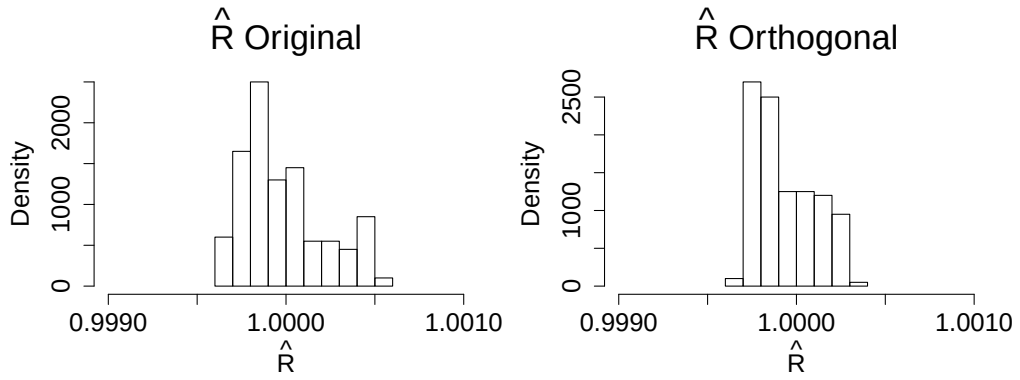


Figure 4: Histograms of $\hat{R}$ values from MCMC sampling with original and orthogonal parametrization and $N = 100$. Notice that the scale in the x-axis is very small so that the values are practically 1, which implies that the MCMC sampling has converged to the true posterior. In the literature a value of 1.1 is used as the upper limit of $\hat{R}$ to declare convergence in the chain (Carpenter et al., 2017)

previous sample. The convergence of the chain answers the question if the chain is actually producing samples from the true distribution or is it just sampling a local part of it. The convergence is monitored through $\hat{R}$-value (Gelman & Rubin, 1992), which compares the variance within and between MCMC chains started from different locations, and that can be used to detect if the chain have not been converged. To increase the rate of convergence and the sampling speed, a standard normal distributed random variable $\boldsymbol{z} = L^{-1}\boldsymbol{f}$ is introduced, and Stan is guided to sample its posterior instead of $\boldsymbol{f}$'s posterior. Notice that the L stands for Cholesky decomposition's lower triangular matrix for the prior covariance matrix K for $\boldsymbol{f}$. This trick in the implementation, described as Matt trick in the Stan manual, gave a tremendous boost in the computation as well as increased the convergence of the chains so that majority of the chains converged.

An example of one posterior sampling's convergence analysis with $N = 100$ is visible in Figure 4, where the two histograms of $\hat{R}$ values, one for both parametrization, are presented. The parametrization refers to the parametrization of the observation model of Gaussian process, mentioned in Chapter 1, and not to the parametrization of the Matt trick introduced in previous paragraph. Notice that as the histograms are very near to one, it means that the chains have converged, and there should be no question whether or not the samples present the true distribution in this case. Also a large amount of training samples is provided to convince the reader of convergence, as the convergence is less likely to happen in higher dimensional posteriors.

2.3 Predictions

Suppose that there is a test point $\boldsymbol{x}_*$ and that a probabilistic prediction for the corresponding y_* is desired. Notice that once the corresponding distribution for the

predictive latent vector $\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*$ is known, it can be used to produce a distribution for y_* by using the properties of the observation model. The posterior predictive distribution combines the posterior information with the posterior predictive distribution given $\boldsymbol{f}$ (see Appendix B) in the following way

$$\begin{aligned} p(\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) &= \int p(\boldsymbol{f}_*, \boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) d\boldsymbol{f} \\ &= \int p(\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{f}, \boldsymbol{x}_*) p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) d\boldsymbol{f} \\ &= \int p(\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{f}, \boldsymbol{x}_*) p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}) d\boldsymbol{f}, \end{aligned} \tag{12}$$

where the marginalization property of probability were used as well as the independencies $\boldsymbol{f}_* \perp\!\!\!\perp \boldsymbol{y}|\boldsymbol{f}$ and $\boldsymbol{f} \perp\!\!\!\perp \boldsymbol{x}_*$ assumed in Figure 2. Now that the distribution of f_* is known, it is straightforward to produce point estimates such as $\mathbb{E}\left[y_*|\hat{\boldsymbol{f}}_*\right]$, where $\hat{\boldsymbol{f}}_*$ is the maximum a posterior estimate (MAP) of the latent predictive process. Predictions given by $\mathbb{E}\left[y_*|\hat{\boldsymbol{f}}_*\right]$ are called the MAP predictions. Conditioning on the MAP estimate of $\boldsymbol{f}$ is less often used than marginalizing the latent predictive process $\boldsymbol{f}_*$ out from the joint distribution

$$\begin{aligned} p(y_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) &= \int p(y_*, \boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) d\boldsymbol{f}_* \\ &= \int p(y_*|\boldsymbol{f}_*) p(\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) d\boldsymbol{f}_*, \end{aligned} \tag{13}$$

and calculating some key statistic such as expectation from it.

Equation (13) is a continuous version of weighted average of different predictive distribution for $y_*|\boldsymbol{f}_*$ weighted by the probabilities of the latent predictive distribution $\boldsymbol{f}_*$. Often the posterior predictive distribution for latent process (12) can be analytically infeasible due to a mismatch between prior and likelihood in terms of conjugancy, thus in that case, the (13) also becomes analytically infeasible from its dependence on (12).

In most applications aiming for the full posterior predictive distribution is an overkill. Instead, the response variable $\boldsymbol{y}$'s posterior predictive mean and variance are usually sufficient for the analysis. Especially the variance gives more information about the uncertainty around the estimate. Posterior predictive mean for latent vector $\boldsymbol{f}_*$ given $\boldsymbol{y}$ and $\boldsymbol{X}$ is

$$\mathbb{E}\left[\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*\right] = k_*^T K^{-1} \mathbb{E}\left[\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}\right]. \tag{14}$$

The posterior predictive variance for latent vector $\boldsymbol{f}_*$ given $\boldsymbol{y}$ is

$$Var\left[\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*\right] = k_{**} - k_*^T \left(K^{-1} - K^{-1} Var_{\boldsymbol{f}}\left[\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y}\right] K^{-1}\right) k_*, \tag{15}$$

see Appendix C for the derivation of the above Equations (14) and (15). The previous results (14) and (15) are especially useful when predicting using Laplace approximation introduced in chapter 2.2.1, where they will take a more convenient form as the posterior mean $\mathbb{E}[\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}]$ and variance $Var[\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}]$ can be analytically approximated. A way to predict the response variable y_*, is to derive the posterior predictive mean and covariance for it and marginalizing out the latent process. The posterior predictive mean for the response variable

$$\mathbb{E}[y_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{x_*}] = \mathbb{E}[\mathbb{E}[y_*|\boldsymbol{f_*}]] \tag{16}$$

is achieved through iterated expectation and the independence $y_* \perp\!\!\!\perp \boldsymbol{x_*}, \boldsymbol{X}, \boldsymbol{y}|\boldsymbol{f_*}$. The posterior predictive variance for the response variable

$$Var[y_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{x_*}] = \mathbb{E}[Var[y_*|\boldsymbol{f_*}]] + Var[\mathbb{E}[y_*|\boldsymbol{f_*}]] \tag{17}$$

is achieved through the law of total variance. Equations (16) and (17) suggest conditioning on the random variable $\boldsymbol{f_*}$, whose distribution is often unknown. To overcome this problem, one can either rely on approximating the posterior distribution analytically with Laplace approximation or rely on MCMC approximation. From either one of the approximations of the posterior it is possible to produce approximations of posterior predictive distribution. Another possibility for predicting the y_* values is to use (14) and (15) to compute the expectation of y_* given a particular value of $\boldsymbol{f_*}$ and address the uncertainty with the knowledge about the variance of $\boldsymbol{f_*}$. Popular choices for the particular value of $\boldsymbol{f_*}$ are the mean, mode and median of the posterior predictive distribution. Anyhow, many terms in the above computation are analytically infeasible, and thus most of the times some approximations are required.

2.3.1 Laplace predictions

Predictions with Laplace approximated posterior were implemented similarly as in Algorithm 3.2 in the Rasmussen & Williams (2005). The difference to the Rasmussen and Williams is that the standard Newton method were not used in the optimization when finding the MAP estimate $\hat{\boldsymbol{f}}$, but instead the natural gradient adaptation. Under the Laplace approximation the latent predictive mean (14) is approximated by

$$\begin{aligned}
\mathbb{E}[\boldsymbol{f_*}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{x_*}] &= k_*^T K^{-1}\mathbb{E}[\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}] \\
&\approx k_*^T K^{-1}\hat{\boldsymbol{f}} \\
&= k_*^T \nabla \log p(\boldsymbol{y}|\hat{\boldsymbol{f}}) \\
&\coloneqq \mu_*.
\end{aligned} \tag{18}$$

Here we assume that $\hat{\boldsymbol{f}}$ is a local maximizer for the unnormalized posterior $\Psi(\boldsymbol{f})$, thus the following equation holds

$$\begin{aligned} \nabla_{\boldsymbol{f}}\Psi(\hat{\boldsymbol{f}}) &= 0 \\ \iff K^{-1}\hat{\boldsymbol{f}} - \nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\hat{\boldsymbol{f}}) &= 0 \\ \iff K^{-1}\hat{\boldsymbol{f}} &= \nabla_{\boldsymbol{f}} \log p(\boldsymbol{y}|\hat{\boldsymbol{f}}). \end{aligned} \tag{19}$$

The difference with variance again, is that the W matrix is not diagonal, and the Algorithm 3.2 from Rasmussen & Williams (2005) cannot be implemented as such. Notice also that the natural gradient adaptation is used instead of the Newton method, and thus the Hessian matrix is replaced with a Fisher information matrix, that is, replacing W with $\mathcal{I}(\boldsymbol{f})$ as stated in (11). Notice that in the Fisher information is only used when searching for the MAP estimate. In predictive approximations the negative Hessian at the MAP estimate, denoted as $\hat{W} := W(\hat{\boldsymbol{f}})$, is used. Relying on the Gaussian approximation of (15) and the result given in (9) the predictive latent variance becomes

$$\begin{aligned} &Var\left[\boldsymbol{f_*}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{x_*}\right] = k_{**} - k_*^T\left(K^{-1} - K^{-1}Var_{\boldsymbol{f}}\left[\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}\right]K^{-1}\right)k_* \\ &\approx k_{**} - k_*^T\left(K^{-1} - K^{-1}(\hat{W}+K^{-1})^{-1}K^{-1}\right)k_* \\ &= k_{**} - k_*^TK^{-1}k_* + k_*^TK^{-1}(\hat{W}+K^{-1})^{-1}K^{-1}k_* \\ &= k_{**} - k_*^TK^{-1}k_* + k_*^TK^{-1}(K - K(\hat{W}^{-1}+K)^{-1}K)K^{-1}k_* \\ &= k_{**} - k_*^TK^{-1}k_* + k_*^TK^{-1}KK^{-1}k_* - k_*^TK^{-1}K(\hat{W}^{-1}+K)^{-1}KK^{-1}k_* \\ &= k_{**} - k_*^T(\hat{W}^{-1}+K)^{-1}k_* \\ &:= \Sigma_*, \end{aligned} \tag{20}$$

where the matrix inversion lemma, also known as Sherman-Morrison-Woodbury formula (Higham, 2002a), was used on the fourth equality. With the mean and variance known, a set of samples can be drawn from the approximative multivariate Gaussian distribution $\boldsymbol{f_*} \sim \mathcal{N}(\mu_*, \Sigma_*)$. Once the samples are available, the predictions of the response y_* are produced by approximating the equations (16) and (17) with sample mean and sample covariance.

2.3.2 MCMC predictions

Predicting with MCMC follows a similar routine as the Laplace approximation prediction, but instead of Gaussian distribution, it gets its samples from MCMC sampling. After acquiring the posterior samples we can produce posterior predictive samples from them using (87) and again use the posterior predictive samples to approximate (16) and (17) to produce probabilistic predictions of the response variable y_*. In order to produce posterior predictive samples from posterior samples $\{\boldsymbol{f}_i\}_{i=1}^S$ we need to derive S predictive means and one covariance from (87). Notice that only one covariance matrix is required, as it does not depend on the sampled $\boldsymbol{f}$ values.

After deriving the means and the covariance matrix, it is possible to sample $\{\boldsymbol{f}_{*i}\}_{i=1}^{S}$ by setting

$$\boldsymbol{f}_{*i} = \mu(\boldsymbol{f}_i) + Lz, \tag{21}$$

where $\mu(\boldsymbol{f}_i)$ is the mean related to ith posterior sample, L is the Cholesky decomposition of the posterior predictive covariance and $z \sim N(0, I)$. This approach is computationally efficient and approaches the posterior predictive distribution as the number of samples sampled for both, posterior and posterior predictive distributions, go to infinity.

2.4 Adaptation of hyperparameters

A common problem for many machine learning models is that they do not only depend on the parameters, but also on some hyperparameters, which are difficult to choose. They often define the frame within which the model can vary. For example hyperparameters can be the number of layers and nodes in each layer in a neural network, or they can be the number of trees in a random forest. Hyperparameters in this thesis, and especially in a Gaussian process model equipped with squared exponential covariance function, are the variables l and σ^2 used in (2). Other covariance functions could be used, and thus the hyperparameters and number of them could change. In statistical modeling it is possible to infer the hyperparameters from the data. Inferring them from the data in Bayesian modeling begins by conditioning the posterior distribution (3) with the hyperparameters $\boldsymbol{\theta}$

$$p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})}. \tag{22}$$

From the above equation the density function to optimize with respect to the hyperparameters is the marginal likelihood

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \int p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X}, \boldsymbol{\theta})d\boldsymbol{f}, \tag{23}$$

which is the probability of the observed response variable given only the hyperparameters and the observed features $\boldsymbol{X}$. When $\boldsymbol{y}$ is fixed, the marginal likelihood is a function of hyperparameters $\boldsymbol{\theta}$ and maximizing it with respect to the hyperparameters leads to the set of hyperparameters that have most likely generated the observed data. When the observation model is Gaussian, the marginal likelihood can be calculated analytically. However, in most cases the marginal likelihood is intractable due to non-Gaussian observation model, and approximations are the only way to proceed, as the high dimensional integral of (23) is often analytically intractable. Fortunately, Laplace approximation is again a way to approximate the hard density function

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{\theta})p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\theta})d\boldsymbol{f} = \int \exp\left(\Psi\left(\boldsymbol{f}\right)\right) d\boldsymbol{f}, \tag{24}$$

for which the Taylor expansion at the $\hat{\boldsymbol{f}}$ is used again, similar to the treatment given for (5). The Taylor expansion yields to approximation

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) &\approx \int \exp\left(\Psi(\hat{\boldsymbol{f}}) + (\boldsymbol{f}-\hat{\boldsymbol{f}})^T \overbrace{\nabla_{\boldsymbol{f}}\Psi(\hat{\boldsymbol{f}})}^{=0} \right. \\ &\left. -\frac{1}{2}(\boldsymbol{f}-\hat{\boldsymbol{f}})^T \left(-(\nabla\nabla)_{\boldsymbol{f}}\Psi(\hat{\boldsymbol{f}})\right)(\boldsymbol{f}-\hat{\boldsymbol{f}}) \right) d\boldsymbol{f} \\ &= \exp(\Psi(\hat{\boldsymbol{f}})) \int \exp\left(-\frac{1}{2}(\boldsymbol{f}-\hat{\boldsymbol{f}})^T \overbrace{\left(-(\nabla\nabla)_{\boldsymbol{f}}\Psi(\hat{\boldsymbol{f}})\right)}^{:=A} (\boldsymbol{f}-\hat{\boldsymbol{f}}) \right) d\boldsymbol{f}. \end{aligned} \tag{25}$$

Thus the approximative log marginal likelihood $\log q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})$ can be written as

$$\begin{aligned} &\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) \approx \log q(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) \\ &= \Psi(\hat{\boldsymbol{f}}) + \log\left(\int \exp\left(-\frac{1}{2}(\boldsymbol{f}-\hat{\boldsymbol{f}})^T A (\boldsymbol{f}-\hat{\boldsymbol{f}}) \right) d\boldsymbol{f} \right) \\ &= \log p(\hat{\boldsymbol{f}}|\boldsymbol{X},\boldsymbol{\theta}) + \log p(\boldsymbol{y}|\hat{\boldsymbol{f}},\boldsymbol{\theta}) + \frac{1}{2}\log\det(2\pi A^{-1}) \\ &= -\frac{1}{2}\log\det(2\pi K) - \frac{1}{2}\hat{\boldsymbol{f}}^T K^{-1}\hat{\boldsymbol{f}} + \log p(\boldsymbol{y}|\hat{\boldsymbol{f}},\boldsymbol{\theta}) + \frac{1}{2}\log\det(2\pi A^{-1}) \\ &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\det(K) - \frac{1}{2}\hat{\boldsymbol{f}}^T K^{-1}\hat{\boldsymbol{f}} + \log p(\boldsymbol{y}|\hat{\boldsymbol{f}},\boldsymbol{\theta}) + \frac{1}{2}\log 2\pi - \frac{1}{2}\log\det(A) \\ &= -\frac{1}{2}\hat{\boldsymbol{f}}^T K^{-1}\hat{\boldsymbol{f}} + \log p(\boldsymbol{y}|\hat{\boldsymbol{f}},\boldsymbol{\theta}) - \frac{1}{2}\log\left(\det(K)\det(W(\hat{\boldsymbol{f}}) + K^{-1})\right), \end{aligned} \tag{26}$$

which leads to the possibility of adapting the hyperparameters that maximize the marginal likelihood.

Optimizing the marginal likelihood is a reasonable approach to determine the hyperparameters as it produces very natural way of describing the data by the hyperparameters that most likely have generated it. However, with this approach one winds up in the familiar discussion between frequentist and Bayesian statisticians. Choosing the hyperparameters that maximize the marginal likelihood is called the maximum likelihood estimate (MLE) for hyperparameters. Bayesians usually argue, that there is some prior information, that is left out when choosing the parameters by MLE. For example, in the squared exponential covariance function (2) some knowledge about the hyperparameters $\boldsymbol{\theta} = (\sigma, l)$ is known from the values of $\boldsymbol{X}$ and $\boldsymbol{y}$ and their relationship. For example, if the values of x vary in the range of $[0, 1]$, then the length scale l should not be larger than 1, unless there is really strong linear trend in $\boldsymbol{y}$. If the $\boldsymbol{y}$ varies a lot within this short period of features, then the

prior knowledge of σ^2 is that it should be quite large depending on the magnitude of variation.

The above discussion is steering the adaptation of hyperparameters towards a Bayesian approach, often called hierarchical modeling, where the hyperparameters are given a suitable prior distribution too. In hierarchical model, the depth of the hierarchy is decided by the modeler, however, at some point he needs to set the top layer. The problem here is that one could, for example, set some prior probability distribution for the hyperparameters, which would again depend on some parameters, hyperhyperparameters. Then for these parameters it would again be possible to set another prior distribution. This infinite loop is then not terminated until the modeler sets some fixed parameters for the top level. In Gaussian process context we can extend the hierarchy to cover the hyperparameters and even the models, that is, different covariance functions. If the hierarchy is extended to cover only the hyperparameters so that the model is assumed to be fixed, one should aim for the hyperparameters $\hat{\boldsymbol{\theta}}$ that maximize the posterior distribution of the hyperparameters

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}), \tag{27}$$

where

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{28}$$

The independence $\boldsymbol{\theta} \perp\!\!\!\perp \boldsymbol{X}$ is assumed, and the prior knowledge is coded into the prior distribution $p(\boldsymbol{\theta})$.

Optimizing the hyperparameters can be done in both ways, with marginal likelihood or with hyperposterior, and they are in fact both being used in the literature. Marginal likelihood approach is favored by Rasmussen & Williams (2005), whereas hierarchical approach is used by Rue et al. (2009) and Vanhatalo et al. (2013). In this thesis the optimization is done by maximizing the hyperposterior (28).

Maximum likelihood and maximum a posteriori estimates are both point estimates, which carry problems in certain cases. Neither of the estimates accounts the uncertainty around the maximum, and thus can be too strict, as the parameter values close to the maximum are also plausible candidates to have generated the data. In special situations it is possible that ML, MAP or other point estimates lead to undesired behavior. For example in Figure 5, the maximum density of the MAP estimate drops very fast. All of the probability mass in the left is left out of the considerations, which then yields the model to ignore the possibility that the data was generated differently, by the parameters located on the left in the plot. The problematics of point estimates is brought up as it can become very suboptimal depending on the focus of one's approach. In the applications of Gaussian processes, the multimodal distributions are rare when using some known likelihood function (Vanhatalo et al., 2009). In local experiments the Weibull likelihood model that is used in this thesis, was always log-concave, and when combined with log-concave Gaussian prior, it is likely that the log-posterior that is being approximated is also unimodal.

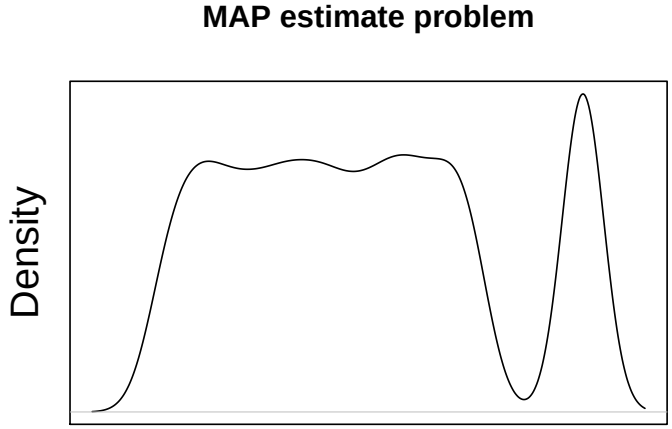


Figure 5: Problem with the MAP estimates. A single point of high probability density value is chosen, but the other possibilities on the left are left out. This decision is strict as it does not consider other parameter values.

If one would like to be really coherent with the Bayesian hierarchical model, he would integrate over the posterior of the hyperparameters to address the uncertainty of them. In mathematical notation, one would then like to achieve the posterior

$$\begin{aligned} p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}) &= \int p(\boldsymbol{f}, \boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) d\boldsymbol{\theta} \\ &= \int p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) d\boldsymbol{\theta}. \end{aligned} \tag{29}$$

In this thesis the hyperparameters are considered to be fixed, but it should be kept in mind that the most coherent way of modeling is to address the uncertainty in the hyperparameters and integrate over the posterior of them. However, every step increases the computational burden, and so tradeoffs between coherency and computation time are often needed.

2.5 Details of technical implementation

Inverting Equation (10) is computationally instable for the Weibull observation model, and it is implemented assuming that the W is diagonal with non-negative elements in the case of classification in Rasmussen & Williams (2005). The natural gradient adaptation is used, and thus, the W in Rasmussen & Williams (2005) and (10) is replaced with the Fisher information matrix $\mathcal{I} = \mathcal{I}(\boldsymbol{f})$. When the W matrix is

replaced with the Fisher information matrix, then the Newton method becomes natural gradient adaptation as described in Section 2.2.1. The goal of the reparametrization was to make the $\mathcal{I}$ matrix diagonal to ease the calculation. However, with the original parametrization of the Weibull observation model, the $\mathcal{I}$ matrix will still contain two off diagonal bands, as the probability of a single observation y_i depends on two latent processes so that $\frac{\partial^2}{\partial\alpha_i\partial\alpha_{i+N}}\log p(\boldsymbol{y}|\boldsymbol{\alpha}) = \sum_{i=1}^{N}\log\frac{\partial^2}{\partial\alpha_i\partial\alpha_{i+N}}p(y_i|\boldsymbol{\alpha}) \neq 0$. Notice that $\boldsymbol{\alpha}$ here stands for the long vector of latent vectors' elements ran through and activation function, that is, $\boldsymbol{\alpha} = [\alpha(f_{11}),\ldots,\alpha(f_{1N}),\ldots,\alpha(f_{21}),\ldots,\alpha(f_{2N})] \coloneqq [\alpha_1,\ldots,\alpha_{2\times N}]$. Notice that the difference between $2N$ and $2\times N$ is significant as the first one indexes a pair of latent vectors with two indexes, first denoting the index of the latent vector and the second one indicating the element of the corresponding latent vector, and the second notation indexing a vector of α values of the length $2\times N$. In the implementation for the original parametrization coded for this thesis, (10) is implemented as

$$\begin{aligned}\boldsymbol{f}^{new} &= (K^{-1}+\mathcal{I})^{-1}(\underbrace{\mathcal{I}\boldsymbol{f}+\nabla\log p(\boldsymbol{y}|\boldsymbol{f})}_{\coloneqq b}) \\ &= ((I+\mathcal{I}K)K^{-1})^{-1}b \\ &= K\underbrace{(I+\mathcal{I}K)^{-1}b}_{\coloneqq a},\end{aligned} \tag{30}$$

where it is possible to rewrite the middle term as

$$\begin{aligned}(I+\mathcal{I}K)^{-1} &= (I+LL^TK)^{-1} \\ &= (L(L^{-1}+L^TK))^{-1} \\ &= (L^{-1}+L^TK)^{-1}L^{-1} \\ &= (L^{-1}+L^TKLL^{-1})^{-1}L^{-1} \\ &= ((I+L^TKL)L^{-1})^{-1}L^{-1} \\ &= L(\underbrace{I+L^TKL}_{CC^T})^{-1}L^{-1} \\ &= L(C^T)^{-1}C^{-1}L^{-1},\end{aligned} \tag{31}$$

where L and C are the lower triangular matrices of Cholesky decomposition for $\mathcal{I}$ and $(I+L^TKL)$ respectively.

Positive definiteness The Newton method and natural gradient adaptation requires inverting two matrices (stated above). The Cholesky decomposition stabilizes the inversion. The problem is that term $(I+L^TKL)$ is not always positive definite, and thus the Cholesky decomposition CC^T is not possible. To overcome this problem an R library called *Matrix*'s method *nearPD()* was used. This method searches for the nearest positive definite matrix in terms of Frobenius norm. The details of the algorithm can be found from Higham (2002b).

Marginal likelihood Following the approximation introduced in (26), it is possible to get approximations of the marginal likelihood. The determinant term $\det(K)\det(K^{-1} + W(\hat{\boldsymbol{f}}))$ in terms of natural gradient adaptation can be written as

$$\begin{aligned} \det(K)\det(K^{-1} + \hat{\mathcal{I}}) &= \det(I + K\hat{\mathcal{I}}) \\ &= \det(\hat{\mathcal{I}}^{-1} + K)\det(\hat{\mathcal{I}}), \end{aligned} \tag{32}$$

to avoid inverting the K matrix, which is usually computationally heavier than inverting the $\hat{\mathcal{I}}$, as it is possible to benefit from the structure of the $\hat{\mathcal{I}}$. Notice that $\hat{\mathcal{I}} := \mathcal{I}(\hat{\boldsymbol{f}})$. It can also be numerically more difficult to invert the K than it is to invert the $\mathcal{I}$. When optimizing the posterior distribution with natural gradient adaptation, the $W(\hat{\boldsymbol{f}})$ is replaced by $\hat{\mathcal{I}}$, which is positive definitive by definition.

Overshooting The Newton method and natural gradient adaptation (Hartmann & Vanhatalo, 2018) is known to overshoot in spesific situations, and thus try to diverge from the solution. For this reason the implementation has a fallback method, that monitors that the objective $\log q(\boldsymbol{y}|X, \theta)$ increases at each step. Mathematically this is implemented so, that if the $\log q(\boldsymbol{y}|X, \theta)$ did not increase, then we introduce a new a-term by averaging over the current and the previous a-terms. Notice that the a-term is defined in (30), and the point of the averaging is that if the method is trying to overshoot, then it will not go as far as it first intended.

The reparametrization of the observation model will be discussed in Section 3.2, and the motivation for it is to make the Fisher information diagonal. When the model is reparametrized, the Laplace approximation using natural gradient adaptation, and predictions can be implemented same way as in Rasmussen & Williams (2005), and thus make the algorithms as fast as they can be. However the avoidance of overshoot and the method of finding the nearest positive definite matrix are still implemented as a fallback method.

Other tricks to enhance the numerical stability Numerical stability of Gaussian processes is always a tricky part. The computation involves inversion, matrix multiplication, product of large and small numbers, initialization and so on. For these reasons I replaced the *Inf* values of the gradient and Fisher information with numerical maximum of the computer and *-Inf* values with the negative counterpart. The initialization of the hyperparameters were encircled with a try-catch statement to try different initial values until convergence of the optimization.

3 Gaussian processes in scale and shape varying Weibull models

Theoretical aspects of Gaussian processes were presented above in a very general context in terms of the observation model. When modeling with Gaussian process,

a decision about the observation model must be made. The choice can be clear by the underlying phenomena, or it can be complicated due to different aspects of the data and the nature that generated the data. Difficulties can arise from the fact that there are many options for the observation model, and the decision on which to choose is not clear. Another way the observations can cause difficulties to the choice of the observation model is when there is some additional knowledge about the phenomena, for example physical laws, that must be taken into account in the observation model.

This thesis studies the Weibull distribution's properties as an observation model, and if the reparametrization of the model can yield to more effective approximations. The Weibull observation model was chosen as it is asymmetrical and far away from a normal distribution. It has been suggested that the Weibull observation model yields to better approximations, when reparameterized to an orthogonal parameterization (Hartmann & Vanhatalo, 2018). The following sections provide the definition of the Weibull distribution and the orthogonal reparameterization.

3.1 Weibull distribution

Weibull distribution is a very flexible distribution being the generalization the exponential distribution. Its support is in the range $[0, \infty)$, and it is typically parametrized with two positive real parameters called shape and scale. The density function of a Weibull distributed random variable $Y \sim Weibull(\alpha_1, \alpha_2)$ is

$$p(y|\boldsymbol{\alpha}) = \begin{cases} \alpha_1 \alpha_2 \left(a_2 y\right)^{\alpha_1 - 1} \exp\left(-(\alpha_2 y)^{\alpha_1}\right) & \text{, if } y \geq 0 \\ 0 & \text{, otherwise.} \end{cases} \tag{33}$$

Traditional way to parametrize the Weibull distribution is with $\alpha_1 = \alpha$ and $\alpha_2 = \frac{1}{\beta}$ (Gelman et al., 2013) but it is not used in this thesis. Weibull distribution is used in many real life applications for example survival analysis and engineering, see Peltola et al. (2014) and Kotilainen et al. (2018).

Observation model's notation $p(\boldsymbol{y}|\boldsymbol{f}, \boldsymbol{X})$ include the dependence on the latent vector $\boldsymbol{f}$. What is meant by the dependence is that the Weibull distribution parameters are dependent on the $\boldsymbol{f}$. The dependence can be almost arbitrarily decided by the modeler as will be demonstrated in the reparametrization case in Chapter 3.2. To emphasize the dependence of the parameters on the latent vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, the parameters are denoted by

$$\begin{aligned} \alpha_1 &= g(\boldsymbol{f}_1, \boldsymbol{f}_2) \\ \alpha_2 &= h(\boldsymbol{f}_1, \boldsymbol{f}_2), \end{aligned} \tag{34}$$

as functions of the latent vectors. In the standard parametrization each of the parameters depend only on one latent vector by $\alpha_1 = g(\boldsymbol{f}_1)$ and $\alpha_2 = h(\boldsymbol{f}_2)$. As the differentiability of g and h is often desired, an obvious choice of function for α_1 and

α_2 so that the parameters lie on the positive real line is (Vanhatalo et al., 2013)

$$\begin{aligned} \alpha_1 &= \exp(\boldsymbol{f}_1) \\ \alpha_2 &= \exp(\boldsymbol{f}_2). \end{aligned} \tag{35}$$

The above dependence of parameters α_1 and α_2 is just one way how the observation model can depend on the latent vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, and how the posterior is adjusted by the data. This parametrization is referred as original parametrization throughout the thesis.

3.2 Reparametrization

Weibull distribution acts as an perfect example for reparametrization study. The parameters α_1 and α_2 are not naturally orthogonal, which would be beneficial for the computations in natural gradient adaptation as the Fisher information matrix would become diagonal, and the computation could be accelerated. Hypothesis of this thesis is that the reparametrization also affects the asymptotical behavior of Laplace approximation, yielding to a faster convergence towards normal distribution when the parametrization is chosen correctly. Derivation of the orthogonal parametrization follows the papers by Huzurbazar (1956), Cox & Reid (1987) and Hartmann & Vanhatalo (2018) for which the calculations are presented in detail.

Definition 3.1. Elements of a Fisher information matrix $\mathcal{I}$ of an observation model $p(\boldsymbol{y}|\boldsymbol{\alpha})$ are defined as the expectation

$$\mathcal{I}_{ij} = \mathbb{E}\left[-\frac{\partial^2}{\partial\alpha_i \partial\alpha_j} \log p(\boldsymbol{y}|\boldsymbol{\alpha})\right], \tag{36}$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots \alpha_p)$.

The definition of orthogonal parameters is given next.

Definition 3.2. Parameters α_i and α_j are said to be orthogonal if and only if

$$\mathcal{I}_{ij} = 0,$$

for $i \neq j$.

The sufficient regularity conditions are assumed in the following calculations, yielding to an orthogonal parametrization. The derivation starts by defining the new parametrization as $\boldsymbol{\eta} = (\eta_1, \dots, \eta_p) = F(\boldsymbol{\alpha})$, where $F : A \to H$ is assumed to be bijective function, and $\alpha \in A$ and $\eta \in H$. The logarithmic observation model of the new parametrization can then be rewritten as

$$\log p^*(\boldsymbol{y}|\boldsymbol{\eta}) = \log p(\boldsymbol{y}|F^{-1}(\boldsymbol{\eta})) = \log p(\boldsymbol{y}|\boldsymbol{\alpha}), \tag{37}$$

where $p^*(\boldsymbol{y}|\boldsymbol{\eta})$ emphasizes the fact that $p^*(\boldsymbol{y}|\boldsymbol{\eta}) \neq p(\boldsymbol{y}|\boldsymbol{\eta})$. From the above equation the first derivative of the log likelihood with respect to $\boldsymbol{\eta}$, also called the score function, is derived

$$\begin{aligned}\frac{\partial}{\partial \eta_i} \log p^*(\boldsymbol{y}|\boldsymbol{\eta}) &= \frac{\partial}{\partial \eta_i} \log p(\boldsymbol{y}|\alpha_1(\boldsymbol{\eta}), \ldots, \alpha_p(\boldsymbol{\eta})) \\ &= \sum_{k=1}^{p} \frac{\partial \log p(\boldsymbol{y}|\boldsymbol{\alpha})}{\partial \alpha_k} \frac{\alpha_k}{\eta_i}.\end{aligned} \tag{38}$$

From the first derivative, the second derivative is calculated by

$$\begin{aligned}\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log p^*(\boldsymbol{y}|\boldsymbol{\eta}) &= \sum_{k=1}^{p} \frac{\partial}{\partial \eta_j} \left(\underbrace{\frac{\partial \log p(\boldsymbol{y}|\alpha_1(\boldsymbol{\eta}), \ldots, \alpha_p(\boldsymbol{\eta}))}{\partial \alpha_k}}_{:= g(\alpha_1(\boldsymbol{\eta}), \ldots, \alpha_p(\boldsymbol{\eta}))} \frac{\partial \alpha_k}{\partial \eta_i} \right) \\ &= \sum_{k=1}^{p} \left(\frac{\partial g(\boldsymbol{\alpha})}{\partial \eta_j} \frac{\partial \alpha_k}{\partial \eta_i} + g(\boldsymbol{\alpha}) \frac{\partial^2 \alpha_k}{\partial \eta_i \partial \eta_j} \right) \\ &= \sum_{k=1}^{p} \left(\left(\sum_{l=1}^{p} \frac{\partial g(\boldsymbol{\alpha})}{\partial \alpha_l} \frac{\partial \alpha_l}{\partial \eta_j} \frac{\partial \alpha_k}{\partial \eta_i} \right) + g(\boldsymbol{\alpha}) \frac{\partial^2 \alpha_k}{\partial \eta_i \partial \eta_j} \right) \\ &= \sum_{k=1}^{p} \sum_{l=1}^{p} \frac{\partial^2 \log p(\boldsymbol{y}|\boldsymbol{\alpha})}{\partial \alpha_k \partial \alpha_l} \frac{\partial \alpha_l}{\partial \eta_j} \frac{\partial \alpha_k}{\partial \eta_i} + \sum_{k=1}^{p} \frac{\partial \log p(\boldsymbol{y}|\boldsymbol{\alpha})}{\partial \alpha_k} \frac{\partial^2 \alpha_k}{\partial \eta_i \partial \eta_j}.\end{aligned} \tag{39}$$

Then taking the expectation of the negative second derivative of the log likelihood yields to Fisher information for an arbitrary parametrization $\boldsymbol{\eta} = F(\boldsymbol{\alpha})$

$$\begin{aligned}\mathbb{E}\left[-\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log p^*(\boldsymbol{y}|\boldsymbol{\eta}) \right] &= \sum_{k=1}^{p} \sum_{l=1}^{p} \mathbb{E}\left[-\frac{\partial^2 \log p(\boldsymbol{y}|\boldsymbol{\alpha})}{\partial \alpha_k \partial \alpha_l} \right] \frac{\partial \alpha_l}{\partial \eta_j} \frac{\partial \alpha_k}{\partial \eta_i} \\ &\quad - \sum_{k=1}^{p} \underbrace{\mathbb{E}\left[\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{\alpha})}{\partial \alpha_k} \right]}_{=0} \frac{\partial^2 \alpha_k}{\partial \eta_i \partial \eta_j},\end{aligned} \tag{40}$$

which can be rewritten with the elements of the Fisher information parametrized by the original parametrization as

$$\mathcal{I}_{ij}(\boldsymbol{\eta}) = \sum_{k=1}^{p} \sum_{l=1}^{p} \frac{\partial \alpha_l}{\partial \eta_j} \frac{\partial \alpha_k}{\partial \eta_i} \mathcal{I}_{k,l}(\boldsymbol{\alpha}). \tag{41}$$

Matrix notation of (41) is simply

$$\mathcal{I}_{\eta}(\boldsymbol{\eta}) = J(\boldsymbol{\alpha})^T \mathcal{I}_{\alpha}(\boldsymbol{\alpha}) J(\boldsymbol{\alpha}), \tag{42}$$

where

$$J(\boldsymbol{\alpha}) = \begin{bmatrix} \frac{\partial \alpha_1}{\partial \eta_1} & \cdots & \frac{\partial \alpha_1}{\partial \eta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial \alpha_p}{\partial \eta_1} & \cdots & \frac{\partial \alpha_p}{\partial \eta_p} \end{bmatrix} \tag{43}$$

is the Jacobian of the function $\boldsymbol{\alpha}(\boldsymbol{\eta})$. Notice that $\mathbb{E}\left[\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{\alpha})}{\partial \alpha_k}\right] = 0$ because of the regularity conditions that were assumed. Majority of the named probability distributions fulfill these conditions, and the Weibull distribution is among them. Now that the analytical form for the Fisher information in arbitrary parametrization has been obtained, choices about the reparametrization to lead the way for the diagonal Fisher information matrix can be made.

In the Weibull model $p = 2$, which combined with the goal that $\mathcal{I}_{1,2}(\boldsymbol{\eta}) = \mathcal{I}_{2,1}(\boldsymbol{\eta}) = 0$ and (41) leads to the equation

$$\begin{aligned} \mathcal{I}_{1,2}(\eta_1, \eta_2) &= \frac{\partial \alpha_1}{\partial \eta_1}\frac{\partial \alpha_1}{\partial \eta_2}\mathcal{I}_{1,1}(\boldsymbol{\alpha}) + \frac{\partial \alpha_1}{\partial \eta_1}\frac{\partial \alpha_2}{\partial \eta_2}\mathcal{I}_{1,2}(\boldsymbol{\alpha}) + \frac{\partial \alpha_2}{\partial \eta_1}\frac{\partial \alpha_1}{\partial \eta_2}\mathcal{I}_{2,1}(\boldsymbol{\alpha}) + \frac{\partial \alpha_2}{\partial \eta_1}\frac{\partial \alpha_2}{\partial \eta_2}\mathcal{I}_{2,2}(\boldsymbol{\alpha}) \\ &= 0. \end{aligned} \tag{44}$$

Since (44) has more unknowns, α_1 and α_2, than equations, some decisions about the unknown variables α_1 and α_2 can be freely made. A choice that $\alpha_1 = h(\eta_1)$ and $\alpha_2 = h(\eta_1, \eta_2)$ is then made. This choice is arbitrary and could be done in other ways as well. By the above assumptions Equation (44) takes the form of

$$\begin{aligned} \mathcal{I}_{1,2}(\eta_1, \eta_2) &= \frac{\partial \alpha_1}{\partial \eta_1}\underbrace{\frac{\partial \alpha_1}{\partial \eta_2}}_{=0}\mathcal{I}_{1,1}(\boldsymbol{\alpha}) + \frac{\partial \alpha_1}{\partial \eta_1}\frac{\partial \alpha_2}{\partial \eta_2}\mathcal{I}_{1,2}(\boldsymbol{\alpha}) + \frac{\partial \alpha_2}{\partial \eta_1}\underbrace{\frac{\partial \alpha_1}{\partial \eta_2}}_{=0}\mathcal{I}_{2,1}(\boldsymbol{\alpha}) + \frac{\partial \alpha_2}{\partial \eta_1}\frac{\partial \alpha_2}{\partial \eta_2}\mathcal{I}_{2,2}(\boldsymbol{\alpha}) \\ &= \frac{\partial \alpha_1}{\partial \eta_1}\frac{\partial \alpha_2}{\partial \eta_2}\mathcal{I}_{1,2}(\boldsymbol{\alpha}) + \frac{\partial \alpha_2}{\partial \eta_1}\frac{\partial \alpha_2}{\partial \eta_2}\mathcal{I}_{2,2}(\boldsymbol{\alpha}) \\ &= 0. \end{aligned} \tag{45}$$

A further assumption that $\alpha_1(\eta_1) = \exp(\eta_1)$ is made, and the values for $\mathcal{I}_{1,2}$ and $\mathcal{I}_{2,2}$ are derived in Appendix A to get Equation (45) in to the form of

$$\begin{aligned} \exp(\eta_1)\frac{\partial \alpha_2}{\partial \eta_2}\frac{1}{\alpha_2}\underbrace{(1 + \Psi(1))}_{:=c} + \frac{\partial \alpha_2}{\partial \eta_1}\frac{\partial \alpha_2}{\partial \eta_2}\frac{\alpha_1^2}{\alpha_2^2} &= 0 \\ \iff \exp(\eta_1)\frac{1}{\alpha_2}c + \frac{\partial \alpha_2}{\partial \eta_1}\frac{\alpha_1^2}{\alpha_2^2} &= 0 \\ \iff c\partial\eta_1 \exp(-\eta_1) &= -\frac{\partial \alpha_2}{\alpha_2}, \end{aligned} \tag{46}$$

where $\Psi(\cdot)$ is the digamma function. Continuing by solving the above partial differential equation for α_2 yields to an orthogonal reparametrization

$$\begin{aligned} c\partial\eta_1 \exp(-\eta_1) &= -\frac{\partial\alpha_2}{\alpha_2} \\ \iff c\int \exp(-\eta_1)\partial\eta_1 &= -\int \frac{1}{\alpha_2}\partial\alpha_2 \\ \iff -c\exp(-\eta_1) - cz(\eta_2) &= -\log\alpha_2 - a(\eta_1) \\ \iff \alpha_2 &= \exp\left(c\exp(-\eta_1) + cz(\eta_2) - a(\eta_1)\right). \end{aligned} \tag{47}$$

It is typical to differential equations, that they have infinite amount of solutions if no initial value is provided. To make the above solution unique, a choice needs to made since all of the solutions of the above form (47) solve the partial differential equation (46). If initial values $z(\eta_2) = \eta_2$ and $a(\eta_1) = 0$ are given, then the solution becomes unique

$$\begin{aligned} \alpha_1 &= \exp(\eta_1) \\ \alpha_2 &= \exp(c\exp(-\eta_1) + c\eta_2), \end{aligned} \tag{48}$$

and the orthogonal parametrization is complete.

Notice that the Gaussian process prior is assumed for the $\boldsymbol{\eta}$ in the reparametrization case.

4 Experiments

The theory of Gaussian processes and its application to the Weibull model was introduced in the previous chapter. The data used for the research and how it is acquired is presented next. After the introduction of the data the research question is studied and the methods of comparisons are presented in detail. The research question of this study is to compare original and orthogonal parametrizations of Gaussian processes with Weibull likelihood model. The comparison is done with respect to the computational times as well as the accuracy of the posterior predictive distribution of $\boldsymbol{f}_*$. The key hypothesis is that the orthogonal parametrization outperforms the original parametrization in both categories, posterior predictive accuracy as well as in computational time.

4.1 Data

The data used in this thesis is simulated one dimensional data. The aim is to study two different kind of datasets, one where the latent processes are smooth and another one where the latent processes are formed so that the response is changing rapidly in location and in the magnitude of variation. The former will be called smooth data and the latter step data from now on. The step data is short for step function

generated data, where the data is formed so that the other generating latent function follows a step function. Both of the Weibull distribution's parameters are modeled with a Gaussian process, and thus, two latent posterior distributions are learned. The data is produced by first limiting the x-axis to the range of $[0, 1]$ and then divided into 200 equally spaced locations. For each of the 200 location x_i the corresponding y_i value is generated by drawing one sample from the $Weibull(\exp(f_1(x_i)), \exp(f_2(x_i)))$ distribution, where f_1 and f_2 are from either the smooth or step function. From the 200 $\{x_i, y_i\}_{i=1}^{200}$ pairs, N randomly chosen pairs are provided for the inference of the latent vectors' posteriors with varying N. Both of the parametrizations use the same exact data.

4.1.1 Hyperparameters for data

Hyperparameters are optimized with respect to hyperposteriors that use the approximative marginal likelihood (26) as a likelihood and a prior that prefers short length scales l and large variance parameter σ^2. For a fixed variation in the shape parameter of Weibull distribution, the variation of $\boldsymbol{\eta}_2$ is much higher than the variation of $\boldsymbol{f}_2$. If the GP hyperparameters were given the same prior, the prior would restrict the adaptation of hyperparameters more in the latter case than in the former case. For this reason the hyperpriors for $\boldsymbol{\eta}_2$ and $\boldsymbol{f}_2$ were chosen so that the marginal prior for the variation in scale parameter of the Weibull distribution were the same in both parametrizations. A Gamma prior was chosen for all of the hyperparameters, using $l_1, l_2 \sim Gamma(20, 0.03)$ and $\sigma_1^2, \sigma_2^2 \sim Gamma(5, 1)$ for original parametrization, and $l_1, l_2 \sim Gamma(20, 0.03)$, $\sigma_1^2 \sim Gamma(5, 1)$ and $\sigma_2^2 \sim Gamma(5.8, 3.25)$ for the orthogonal parametrization. There are two different pairs of hyperparameters for each parametrization as there are two processes being modeled. Notice that the Gamma distribution is parametrized with shape parameter k and scale parameter ϕ, having a density function $f_{Gamma}(x) = \frac{1}{\Gamma(k)\phi^k} x^{k-1} \exp(-\frac{x}{\phi}) \mathbb{1}_{\{x>0\}}$. In order to find the MAP estimate of the hyperparameters, the hyperposterior is optimized using *optim()* method from *stats* library (R Core Team, 2018), and especially the conjugate gradient method. More on the conjugate gradient method for optimizing is given by Fletcher & Reeves (1964). Should the prior be the same for both of the parametrizations, then the basis for comparison is not fair and one parametrization would be doomed to perform badly. When there is no possible mapping between the hyperparameters to yield to the same probability distribution for the latent processes, one needs to rely on optimization. This is also what I did in the thesis, choose a hyperprior for original parametrization and then optimized the orthogonal hyperprior σ_2^2 parameters k and ϕ so that the (tested empirically) cumulative probabilities at the absolute maximum value of the respective generating process would be as close to each other as possible.

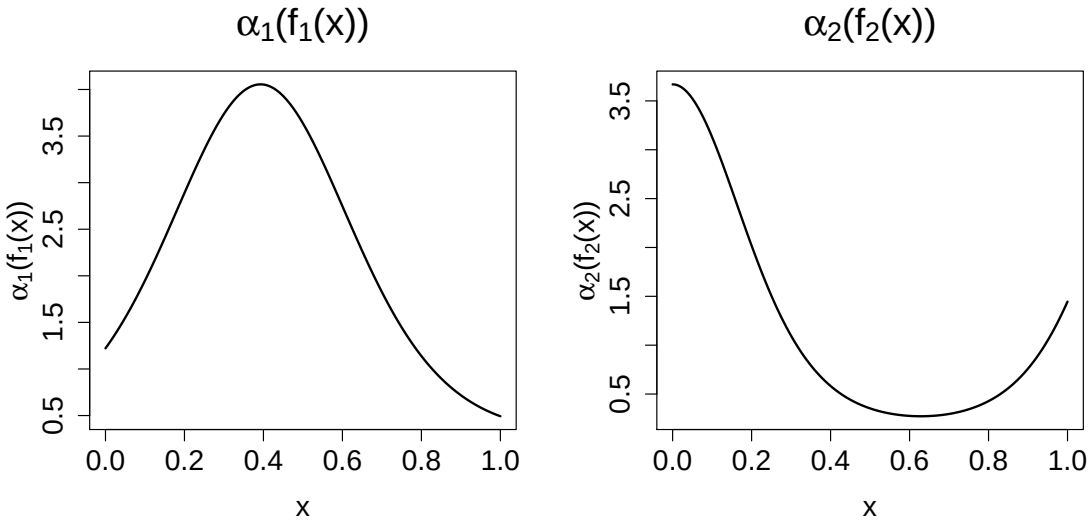


Figure 6: A figure of the generating processes for the smooth data ran through the functions $\alpha_1(x) = \alpha_2(x) = \exp(x)$. Notice that this plot is about the generative process of the training data, so it is independent from the assumptions of the parametrization and modeling in general.

4.1.2 Smooth data

The smooth data is formed through functions

$$\begin{aligned} f_1(x) &= 1.2\sin(4x) + 0.2 \qquad \text{and} \\ f_2(x) &= 1.3\cos(5x), \end{aligned} \tag{49}$$

and the plots for these ran through the exponential activation functions $\alpha_1(x) = \exp(x)$ and $\alpha_2(x) = \exp(x)$ are visible in Figure 6.

The smooth latent processes will still give rise to non smooth response variable $\boldsymbol{y}$ because of the random draws from a Weibull distribution and the large variance at the region where $x \gtrapprox 0.8$, where $\gtrapprox$ stands for "approximately greater than". An example of how a smooth data, and one dataset randomly selected for training with $N = 30$, can look like is visible in Figure 7.

4.1.3 Step data

Step data is produced by taking random draws from a Weibull distribution given two latent functions, same way as the smooth data. The only exception is that the underlying latent functions differ so that the data generated by the process has sharply alternating magnitude both in mean and variance. The latent functions are

$$\begin{aligned} f_1(x) &= 0 \qquad \text{and} \\ f_2(x) &= \begin{cases} 0.2, & \text{if } 0.4 \leq x \leq 0.6 \\ 3, & \text{otherwise.} \end{cases} \end{aligned} \tag{50}$$

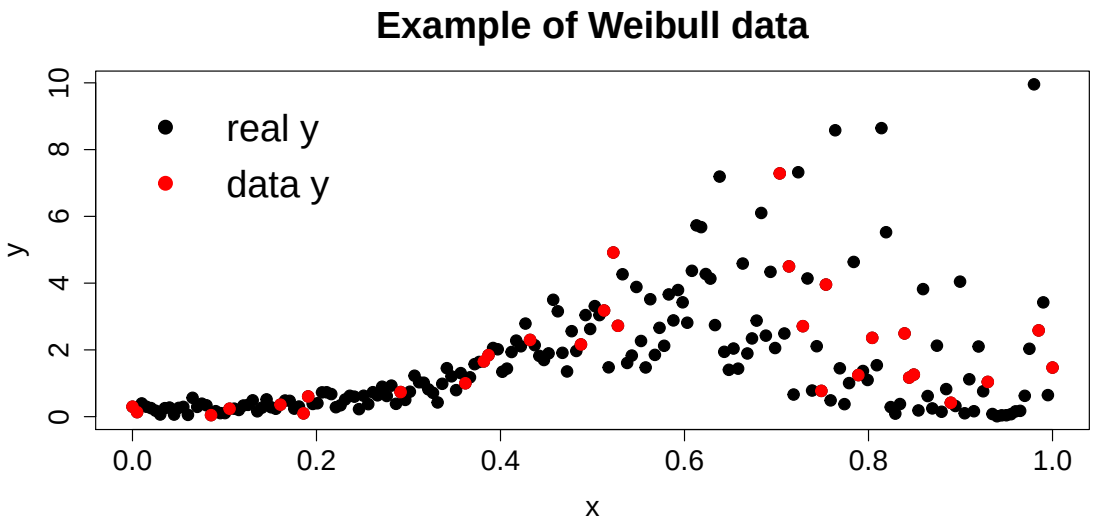


Figure 7: One dataset drawn from the Weibull distribution given the smooth latent processes. The black dots show the simulated data in each of the discretized $\boldsymbol{x}$ location and the red dots show one possibility of training data size $N = 30$.

Similarly to the smooth data, the latent functions ran through an exponential activation function are visible in Figure 8 and one dataset that is generated by them is shown in Figure 9.

4.2 Methods to validate and compare the goodness of the posterior approximations

Comparison of posterior predictive distribution of $\boldsymbol{f}_*|\boldsymbol{y}$ indicates how well does Laplace approximation interpolate the latent function in unseen regions of the feature space. The baseline distribution that Laplace approximated predictive distribution is compared to is the MCMC predictive distribution. In this thesis the comparison methods are done with respect to the posterior predictive mean and covariance, as well as with Kullback-Leibler divergence (Kullback & Leibler, 1951) between the two predictive distributions. The posterior predictive distribution $\boldsymbol{f}_*|\boldsymbol{y}$ is evaluated at $N_* = 101$ points $\boldsymbol{x}_*$ equally spread in the range of $[0, 1]$. The analysis is repeated 100 times for each N. All of the comparative measures will have two different approaches, one comparing the performance in the native space of the parametrization, that is original posterior is in $\boldsymbol{f}$-space, and orthogonal in $\boldsymbol{\eta}$-space. The second way to compare these measures is to map the posterior predictive distributions to the same space and do the comparison. In this thesis, the orthogonal parametrization's posterior predictive distribution is mapped to original parametrization's space and the comparison is done there. Details of the comparisons are presented next.

Covariance function used for the calculations of the comparisons was the squared exponential covariance function (2). Squared exponential was chosen as it is a

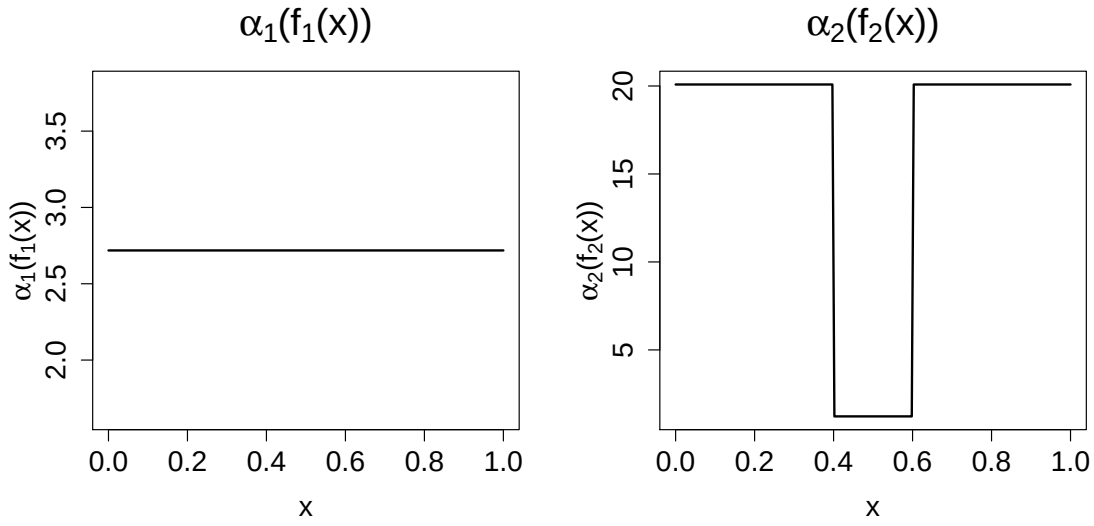


Figure 8: Step data's latent functions ran through exponential activation function, that is, $\alpha_1(x) = \alpha_2(x) = \exp(x)$. Notice that this is again emphasizing the generation of the training data.

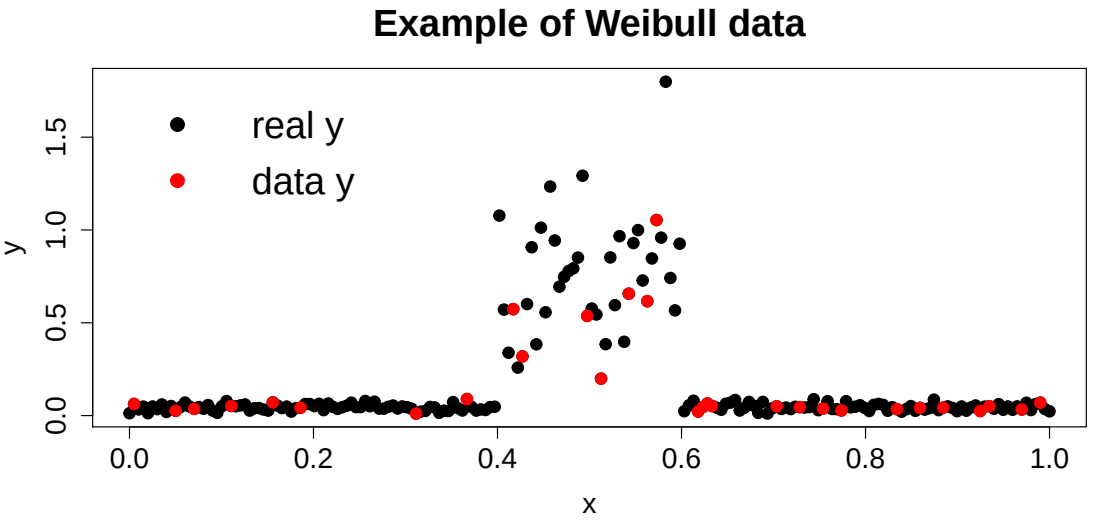


Figure 9: A dataset generated by the step data's latent process and an example how could a 30-point sample from it look like.

smooth and widely used kernel function.

4.2.1 Mean and covariance differences

Mean of the posterior predictive distribution is a N_*-dimensional vector, and given two of the possible mean vectors $\boldsymbol{\mu}_{Laplace}$ and $\boldsymbol{\mu}_{MCMC}$, the mean squared error (MSE) of them is computed. Notice that since there are two parametrizations, there will be one pair of mean vectors for each of the parametrizations. Because the posterior predictive distributions $\boldsymbol{f}$ and $\boldsymbol{\eta}$ of different parametrizations do not represent the same phenomena, two different kind of approaches are taken in order to compare them. In first approach they are mapped to the same space and compared there without any normalization. Second approach keeps the posterior predictive distributions in their native spaces, that is, original parametrization in $\boldsymbol{f}$-space and orthogonal parametrization in $\boldsymbol{\eta}$-space, and then normalizes the difference between MCMC and Laplace approximation by the magnitude of the MCMC posterior predictive samples. When the two comparisons are written in formulas, the first approach in terms of original parametrization becomes

$$MSE_{original_1}(\boldsymbol{f}^*_{MCMC}, \boldsymbol{f}^*_{Laplace}) = \left\| M(\boldsymbol{f}^*_{MCMC}) - \mathbb{E}[\boldsymbol{f}^*_{Laplace}] \right\|^2, \tag{51}$$

where $\boldsymbol{f}^*_{MCMC}$ stands for the predictive sample distribution of $\boldsymbol{f_*}$ for MCMC and similarly $\boldsymbol{f}^*_{Laplace}$ for Laplace. As described above, the original parametrization's posterior predictive distribution is not mapped to anywhere. The first approach for the orthogonal parametrization is

$$MSE_{orthogonal_1}(\boldsymbol{\eta}^*_{MCMC}, \boldsymbol{\eta}^*_{Laplace}) = \left\| M(\boldsymbol{h}(\boldsymbol{\eta}^*_{MCMC})) - M(\boldsymbol{h}(\boldsymbol{\eta}^*_{Laplace})) \right\|^2, \tag{52}$$

where $\boldsymbol{h}(\eta_1, \eta_2) = (\eta_1, c\exp(-\eta_1) + c\eta_2)$ is the inverse mapping from $\boldsymbol{\eta}$-space to $\boldsymbol{f}$-space and $c = 1 + \Psi(1)$. $M(\cdot)$ stands for sample mean. Notice that in the above notation, both of the functions $h(\cdot)$ and $M(\cdot)$ operate over matrices. Here $\boldsymbol{\eta_*}$ is a matrix of posterior predictive samples and the function h is applied for each (η_1, η_2) pair, resulting in the same size matrix as $\boldsymbol{\eta}^*_{Laplace}$ (or $\boldsymbol{\eta}^*_{MCMC}$). For the matrix $h(\boldsymbol{\eta}^*_{Laplace})$ function $M(\cdot)$ computes the column-wise mean, resulting in a vector of length N_*. Similarly $M(\cdot)$ is applied to different matrices, with or without the mapping $h(\cdot)$.

The second approach in terms of equations is

$$MSE_{original_2}(\boldsymbol{f}^*_{MCMC}, \boldsymbol{f}^*_{Laplace}) = \frac{\left\| M(\boldsymbol{f}^*_{MCMC}) - \mathbb{E}[\boldsymbol{f}^*_{Laplace}] \right\|^2}{\left\| M(\boldsymbol{f}^*_{MCMC}) \right\|^2}, \tag{53}$$

for the original parametrization and

$$MSE_{orthogonal_2}(\boldsymbol{\eta}^*_{MCMC}, \boldsymbol{\eta}^*_{Laplace}) = \frac{\left\| M(\boldsymbol{\eta}^*_{MCMC}) - \mathbb{E}[\boldsymbol{\eta}^*_{Laplace}] \right\|^2}{\left\| M(\boldsymbol{\eta}^*_{MCMC}) \right\|^2}, \tag{54}$$

for the orthogonal parametrization.

The covariance matrices are compared in same two ways, except that the (sample) mean is changed to (sample) covariance and the element-wise difference of matrices are compared with matrix version of squared euclidean distance, often stated as squared Frobenius norm. Similarly to the mean comparison, the posterior predictive samples are mapped back from the orthogonal parametrization space to the same space with the original parametrization in the first approach. The normalization is done with respect to the Frobenius norm of posterior predictive sample covariance of MCMC in the second approach.

4.2.2 Kullback-Leibler divergence from the true conditional posterior to the Gaussian approximations

Kullback-Leibler divergence (hereafter KL-divergence) is a traditional and robust way to compare two probability distributions. The scale of KL-divergence goes from zero to infinity, zero standing for a situation comparing a distribution with itself. KL-divergence from q to p is defined as

$$KL(p||q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx. \tag{55}$$

The idea is to calculate the KL-divergence between the true posterior density p and the Laplace approximated density q. However the density function p is unknown in this thesis, and an approximation of it is required. MCMC methods' ability to produce infinite amount of samples from the posterior is not enough. The approximation that was done in this thesis was another Gaussian approximation since the KL-divergence between two Gaussian distribution is analytical.

The goal of the approximation is to find a normal distribution with a density function p_N which minimizes the KL divergence between the true distribution's density p and the approximative density p_N. Notice that as the parameters of the normal distribution identify the normal distribution, it is equivalent to look for a normal distribution or its parameters μ and Σ, that fulfill the minimization goal. For this approximation a theorem is introduced.

Theorem 1. *A normal distribution $p_N(\theta) \sim N(\mu, \Sigma)$ that minimizes the Kullback-Leibler divergence for a density function $p(\theta)$ is such that*

$$\begin{aligned} \mu &= \mathbb{E}[\theta] \\ \Sigma &= Cov[\theta]. \end{aligned}$$

Proof. The definition of KL-divergence states that

$$\begin{aligned} KL(p(\theta)||p_N(\theta)) &= \int p(\theta) \log\left(\frac{p(\theta)}{p_N(\theta)}\right) d\theta \\ &= \int p(\theta) \log(p(\theta))\, d\theta - \int p(\theta) \log(p_N(\theta))\, d\theta, \end{aligned} \tag{56}$$

where only the second term depends on the approximative distribution, and it should be maximized with respect to the μ and Σ to minimize the KL-divergence, as the sign of it is negative. The $p_N(\theta)$ is normally distributed so that the second term can be written as

$$\begin{aligned}\int p(\theta)\log\left(p_N(\theta)\right)d\theta &= \int p(\theta)\left(-\frac{d}{2}\log(2\pi)-\frac{1}{2}\log\det(\Sigma)\right.\\ &\qquad\left.-\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right)d\theta\\ &= -\frac{d}{2}\int p(\theta)\log(2\pi)d\theta\\ &\quad+\underbrace{\int p(\theta)\left(-\frac{1}{2}\log\det(\Sigma)-\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right)d\theta}_{:=U},\end{aligned} \tag{57}$$

where again only U depends on the μ and Σ. After this, it is possible to maximize U with respect to μ, by setting its first partial derivative to zero, and get

$$\begin{aligned}\frac{\partial U}{\partial\mu} &= \int -\frac{1}{2}p(\theta)(-2\Sigma^{-1}(\theta-\mu))d\theta\\ &= \int p(\theta)(\Sigma^{-1}(\theta-\mu))d\theta = 0\\ \iff \int p(\theta)\Sigma^{-1}\theta d\theta &= \int p(\theta)\Sigma^{-1}d\theta\mu\\ \iff \int p(\theta)\theta d\theta &= \int p(\theta)d\theta\mu\\ \iff \mu &= \mathbb{E}\left[\theta\right].\end{aligned} \tag{58}$$

Similarly the differentiation of U with respect to Σ yields that

$$\begin{aligned}\frac{\partial U}{\partial\Sigma} &= \int p(\theta)\left(-\frac{1}{2}\Sigma^{-1}+\frac{1}{2}\Sigma^{-1}(\theta-\mu)(\theta-\mu)^T\Sigma^{-1}\right)d\theta\\ &= -\frac{1}{2}\int p(\theta)d\theta\Sigma^{-1}+\frac{1}{2}\Sigma^{-1}\int p(\theta)\left((\theta-\mu)(\theta-\mu)^T\right)d\theta\Sigma^{-1} = 0\\ \iff -\Sigma^{-1} &+ \Sigma^{-1}Cov\left[\theta\right]\Sigma^{-1} = 0\\ \iff -I &+ Cov\left[\theta\right]\Sigma^{-1} = 0\\ \iff Cov\left[\theta\right]\Sigma^{-1} &= I\\ \iff \Sigma &= Cov\left[\theta\right].\end{aligned} \tag{59}$$

Notice that in the above calculations Equations (86), (57) and (61) were used from Petersen & Pedersen (2012)

In order for these critical points to maximize the U, the second derivatives of them need to be negative definite at the critical point. First the second derivative of U

gives

$$\begin{aligned}\frac{d^2U}{d\mu^2} &= \left(\int p(\theta)\Sigma^{-1}\theta d\theta - \Sigma^{-1}\mu\right) \\ &= -\Sigma^{-1}.\end{aligned}$$

The second derivative of U with respect to Σ is

$$\begin{aligned}\frac{d^2U}{d\Sigma^2} &= \frac{1}{2}\Sigma^{-1}\Sigma^{-1} + \frac{1}{2}\frac{d^2}{d\Sigma^2}(\Sigma^{-1}C\Sigma^{-1}) \\ &\quad \frac{1}{2}\Sigma^{-1}\Sigma^{-1} + \frac{1}{2}\left(-\Sigma^{-1}\Sigma^{-1}C\Sigma^{-1} + \Sigma^{-1}C(-\Sigma^{-1}\Sigma^{-1})\right) \\ &\quad \frac{1}{2}\Sigma^{-1}\Sigma^{-1} - \Sigma^{-1}\Sigma^{-1}C\Sigma^{-1},\end{aligned}$$

where $C = \int p(\theta)\left((\theta-\mu)(\theta-\mu)^T\right)$ and the result is obtained by using symmetry of Σ and C. Now substituting the Σ with critical point $Cov(\theta) = C$ the above equation becomes

$$\begin{aligned}\frac{1}{2}C^{-1}C^{-1} - C^{-1}C^{-1}CC^{-1} &= \frac{1}{2}C^{-1}C^{-1} - C^{-1}C^{-1} \\ &= -\frac{1}{2}C^{-1}C^{-1},\end{aligned}$$

from which it is possible to conclude that both of the second derivatives are negative definite by using the symmetry, positive definiteness of C and substituting Σ with C. Notice that above calculations used Equations (37) and (59) from Petersen & Pedersen (2012). □

The point of Theorem 1 is to discover the best Gaussian approximation for the posterior distribution in terms of KL-divergence. This approximative density function is denoted as p_N. When the approximation p_N has been discovered, the Laplace approximations are compared to the distribution p_N in terms of KL-divergence. Notice that since there are two parametrizations, all of the approximations will be doubled, that is, there will be two approximations $p_{N_{original}}$ and $p_{N_{orhtogonal}}$, and the Laplace approximations for the respective parametrizations. This point requires some attention, as often in literature, a similar kind of approach is taken when comparing two approximations of the same posterior. These comparisons are possible because the parametrizations do not change. An example of the approach taken in literature is when there are two approximations $q_{Laplace}$ and q_{VB}, and the desire is to compare these to the true distribution p. In this thesis the parametrization of posterior p changes, and therefore it is not possible to directly compare the Laplace approximations $q_{Laplace_{original}}$ and $q_{Laplace_{orthogonal}}$.

With the help of Theorem 1 and the knowledge that KL-divergence between two normal distributions is analytical (see Appendix D), it is easy to approximate the KL-divergence between the Laplace approximation and the Gaussian approximation

of MCMC sampled posterior. Notice that μ and Σ are approximated by the sample mean and covariance from the MCMC samples respectively.

Notice that also in KL-divergence comparison two approaches are taken. They follow the similar tracks as was described in Section 4.2.1. First approach is to map the posterior predictive samples of each method, that is Laplace and MCMC, to the same space and compute the approximative KL-divergence from those samples. Again, the orthogonal posterior predictive samples are mapped from $\boldsymbol{\eta}$-space to $\boldsymbol{f}$-space before the comparison is done. Another approach is not to map the posterior predictive samples to the same space and just compare their divergences calculated in different spaces. No normalization is done in the KL-divergence comparison.

4.2.3 Mapping reparametrized posterior back to the original space

The above comparisons uses mapping $\boldsymbol{\eta}$ to $\boldsymbol{f}$-space. To be more precise the original parametrization produces posterior distributions for $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, while orthogonal parametrization produces posterior distributions for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. If they would not be mapped to the same space, the results would only be directional and not accurate as the magnitude of variation in $\boldsymbol{\eta}_2$ is much higher than what it is with $\boldsymbol{f}_2$. Motivation for this is that closeness of Laplace approximation is computed to the MCMC posterior statistics in $\boldsymbol{f}$-space for original parametrization and in $\boldsymbol{\eta}$-space for orthogonal parametrization, the closenesses of Laplace approximations would not be comparable between the different parametrizations. In order to make the parametrizations comparable, posterior predictive distribution for latent variables $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are mapped back to the same space with the original parametrized process $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$. The mapping from $\boldsymbol{\eta}$ back to the $\boldsymbol{f}$ can be recovered from Equation (48) to give

$$\begin{aligned} \boldsymbol{f}_1 &= \boldsymbol{\eta}_1, \quad \text{and} \\ \boldsymbol{f}_2 &= c\exp(-\boldsymbol{\eta}_1) + c\boldsymbol{\eta}_2, \end{aligned} \tag{60}$$

where $c = 1 + \Psi(1)$, where $\Psi(\cdot)$ is again the digamma function. Notice that this mapping was earlier denoted as $h(\cdot)$ for notational purposes and to emphasize the difference between two parametrizations. The above mapping was done for orthogonal Laplace and MCMC posterior predictive sample distributions before doing the comparison in the first approach. For the Laplace case the sampling was done from the posterior predictive distribution given in Section 2.3.1 and then each of the samples were transformed back to the $\boldsymbol{f}$-space using the above transformation. Same procedure was done for the MCMC posterior predictive distribution with the exception that the samples are coming from Stan, instead of analytical Gaussian distribution.

4.2.4 Computing times

The orthogonal parametrization with natural gradient adaptation affects the algorithm for finding the MAP estimate of the posterior distribution required for Laplace

approximation introduced in Chapter 2.2.1. For this reason the computational times for finding the MAP estimate were recorded and the difference will be emphasized in the next chapter. Notice that the optimization of the hyperparameters requires the location of the MAP, and therefore the advantage given by the orthogonalization will multiply in each iteration of conjugate gradient method used for the optimization.

4.2.5 Convergence of the MCMC chains

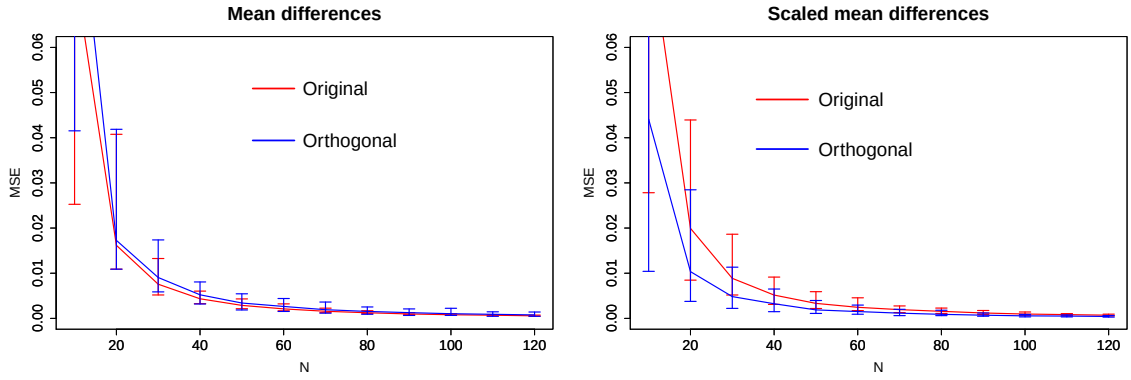
The convergence of the MCMC chains was monitored by the $\hat{R}$-values. Each posterior having any of the dimensions' $\hat{R}$-value above 1.05 led to the disqualification of the corresponding MCMC samples and posterior approximation. Most of the posterior sampling repetitions for sample sizes had no problems in converging and had zero discarded samplings. Maximum number of discarded posterior samplings were 4 for one sample size, where then there were 96 comparisons done instead of 100. Orthogonal parametrization seemed to have problems with converging more often than the original one.

5 Results

The following experiments study the question whether or not the specific orthogonal parametrization increases the rate that the posterior distribution approaches the limiting normal distribution or is it better to stay in the original parametrization. The results include comparative plots for the measures introduced in the previous section as well as some example plots of the posterior predictive distribution's characteristics for illustrative purposes. For each comparative measure, two approaches are taken. Label (a) is used when the posterior predictive distribution of orthogonal parametrization is mapped to the $\boldsymbol{f}$-space before doing the comparison. Label (b) is used, when there is no mapping of orthogonal parametrization, and the difference is normalized with a statistic of MCMC posterior predictive distribution. Notice that the solid line in the upcoming performance comparing plots denotes the median of the corresponding measure over $\sim$100 posterior predictive distribution comparisons on different datasets. Notice that some of the comparisons were dropped because of convergence issues as explained above. The confidence interval around the line denotes the region between 2.5% and 97.5% quantiles, unless otherwise stated. In the plots for posterior predictive distributions the solid line denotes mean value for Laplace approximated distribution and the median value for MCMC predictive distribution. Notice that Laplace approximation's mean equals to the median since the predictive distribution is normally distributed.

5.1 Smooth data

The mean squared error of the posterior predictive distribution means for concatenated latent processes $\boldsymbol{f} = [\boldsymbol{f}_1, \boldsymbol{f}_2]$ in Figure 10(a) shows only marginally better

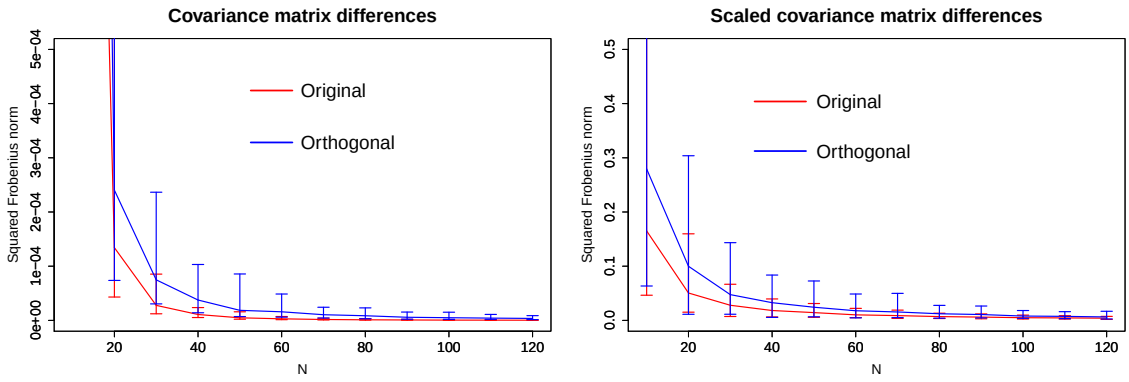


(a) (b)

Figure 10: (a) Mean squared errors of the smooth data for latent processes posterior predictive means. The different colors correspond to different parametrization's differences. Notice that the orthogonal parametrization is transformed back to $\boldsymbol{f}$-space in order for the means to be comparable without the need for scaling. Notice also that there are two processes and their MSE is compared by concatenating the two processes $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ into a long vector $\boldsymbol{f}$. (b) Scaled mean squared errors of the predictive mean of Laplace approximation and MCMC predictive sample mean. Each MSE is normalized by the squared L2-norm of the respective parametrization's MCMC predictive sample mean vector.

performance for the original parametrization than for the orthogonal one. When N is sufficiently modest, then median of the mean squared errors, denoted as line, is visually smaller when the parameters are original. When N grows, the differences become impossible to detect. Therefore, both of the parametrizations' differences seem to converge to a same small value. Orthogonal parametrization has wider confidence interval at every sample size, whereas original parametrization's confidence interval becomes so narrow that it is impossible to detect it from the plot when the sample size is large. Figure 10(b), is otherwise similar to 10(a) except the comparison is done in the parametrizations' own spaces. The MSE of mean vectors is normalized with the squared L2-norm of the MCMC mean vector. Interestingly in this comparison the orthogonal parametrization performs better in terms of rate of convergence and stability. The major difference is at low sample sizes in favor of the orthogonal parametrization. However both of the parametrizations' MSE of means still converge to a same region around zero. Notice that the formulas for the comparisons were presented in Equations (51)-(54).

In Figure 11(a), the squared Frobenius norm of the elementwise difference between Laplace approximated posterior predictive sample covariance and sample covariance of MCMC drawn posterior predictive samples is being compared in both parametrizations. In the figure the decreasing trend is present, meaning that as more and more data is observed, then the Laplace approximation's covariance approaches the MCMC sample covariance in both parametrizations. However it is evident that the original parametrization performs better than orthogonal at least

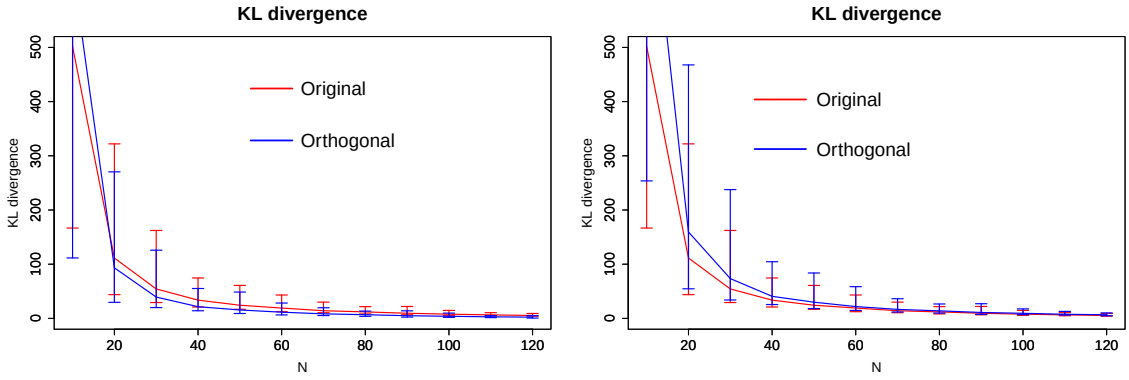


(a) (b)

Figure 11: (a) Squared Frobenius norm of the difference between Laplace approximation's posterior predictive covariance and posterior predictive sample covariance of MCMC samples. Orthogonal samples are first projected back to the $\boldsymbol{f}$-space before the comparison. Different colors denote different parametrizations. Notice that when $N = 10$ the value of the measure is beyond the margins of this plot, however the behavior is similar as in the plot otherwise, that is, orthogonal parametrization having larger median value of the Frobenius norm of the differences and wider confidence interval. (b) Squared Frobenius norm of the difference between Laplace approximated posterior predictive covariance and MCMC posterior predictive sample covariance normalized by the squared Frobenius norm of the MCMC posterior predictive sample covariance. Notice that in this comparison the parametrizations are compared in their native parameter space and no mapping is done.

when the number of observations is modest. Both parametrization's median of covariance differences converge close to the same number, orthogonal parametrization having slightly wider confidence interval. Notice that the comparison is not visible when $N = 10$ as the change in scale is so big. However the orthogonal parametrization performs worse also when $N = 10$. Figure 11(b) is the squared Frobenius norm of the difference between the posterior predictive sample covariance between Laplace approximation and MCMC normalized by the squared Frobenius norm of the MCMC posterior predictive sample covariance. The plot is very similar to the plot in 11(a) with the exception that the differences are actually visible for $N = 10$ as well. Notice that the change in y-scale is huge. See also Equations (51)-(54) and the discussion after them for mathematical intuition for the comparison.

The final measure of predictive accuracy is the KL-divergence, presented in Figure 12. In Figure 12(a) the KL-divergence shows that onwards from when $N = 20$ the orthogonal parametrization actually performs better. The orthogonal parametrization has narrower confidence intervals and lower median divergence than the original parametrization. When $N = 10$ the median of the orthogonal parametrization is higher and the confidence interval is wider and the results for that are not visible in the plot. Figure 12(b) shows contradictory behavior from 12(a) as the orthogonal parametrization is performing worse at all samples sizes. Notice that the parameters



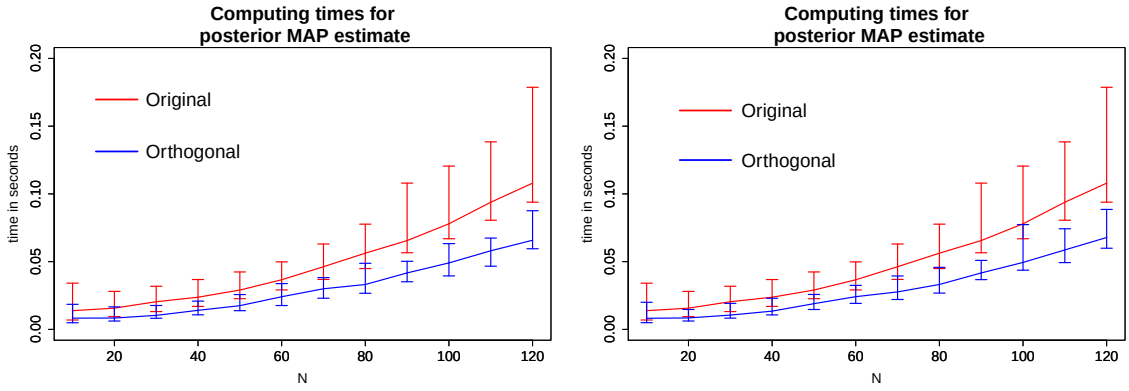
(a) (b)

Figure 12: (a) Approximative KL-divergence between the Laplace approximation and MCMC sampling with the smooth data. The approximative KL-divergence was presented in Chapter 4, and the approximation is done for both parametrizations, denoted with different colors. Before the approximation is done the orthogonal parametrization's posterior predictive distributions are projected from $\boldsymbol{\eta}$-space to the original parametrization's $\boldsymbol{f}$-space. (b) Approximative KL-divergence between the Laplace approximation and the Laplace approximated and MCMC posterior predictive distributions. Notice that in this comparison both of the parametrizations are kept in their native spaces.

are kept in their native space. The performance of the orthogonal parametrization is also less stable as it has wider confidence intervals.

The computing times of the MAP estimate using the natural gradient adaptation, introduced in Section 2.2.1, are plotted in Figure 13. As expected, when the Fisher information is diagonal, the computational complexity decreases, and thus the orthogonal parametrization is faster than the original parametrization in wall clock time. All of the comparisons were computed with a single Intel i7-4770 processor, and the computation times were recorded from the iterations of the actual computations used for the predictive accuracy comparisons. Notice that the computing times were saved for both of the approaches mentioned above, the one where the posterior predictive distribution samples are mapped back to the same space (labelled as (a)) and the one where they are not (labelled as (b)). The mapping does not affect the computing times of the posterior MAP as the procedure is done after finding the MAP. For completeness and robustness, the times for both of the approaches are plotted.

To give the reader more insight about what is actually being compared above in the (a) plots, an example plots of posterior predictive distribution for latent processes for the smooth data are provided in Figure 14. Notice that in the figure the two processes $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are in separate plots, separated to different columns by the process and different rows by the parametrization. Each plot have three lines to denote the generating function of the response variable from equations in (49), posterior



(a) (b)

Figure 13: Wall clock times of finding the maximum of posterior distribution using natural gradient adaptation with the smooth data. Different colors denote the different parametrizations. Notice that (a) and (b) show the same computing times of the same data and the only difference is the stochastic nature of computing. Both of the plots are included for coherency to other result plots and to provide more confidence on the result.

predictive mean for Laplace and posterior predictive median for MCMC. The dashed lines correspond to the 95% confidence interval. These predictive distributions are then computed and compared for each varying N and for each repetition of N. Notice also that the whole posterior predictive covariance matrix is not presented in the aforementioned figure which rather shows a marginal variance of each predictive location. An example of what is being compared in the covariance comparison, the full covariance matrices are visualized in Figure 15. The Frobenius norm is calculated from the elementwise difference of the covariance matrices. As each row of the figure represents one parametrization, the comparison takes place row-wise. Notice that all of the example visualizations discussed in this paragraph are taken from a single run with sample size $N = 30$ and their purpose is to clarify the meaning of the comparative measures. The difference in these plots is really marginal and if looked accurately enough one can spot some differences in the latent posterior predictive distributions depending on the parametrization.

In the (b) plots the comparison is done in the native spaces of the parametrizations. That is, MCMC statistics in $\boldsymbol{f}$-space are compared to Laplace statistics in $\boldsymbol{f}$-space for original parametrization and same thing is done for orthogonal MCMC and Laplace posterior predictive distributions that lie in the $\boldsymbol{\eta}$-space. Comparing these disrtibutions in their native spaces in a same plot would not be informative, which is why there is no corresponding plots for the (b) comparisons.

Finally an example of how Gaussian process predictions on the response variable can look like is visible in Figure 16(a) for original parametrization and Figure 16(b) for orthogonal parametrization, when the data is generated by the smooth processes. Both parametrizations seem to fit the data well and almost equally. MAP prediction

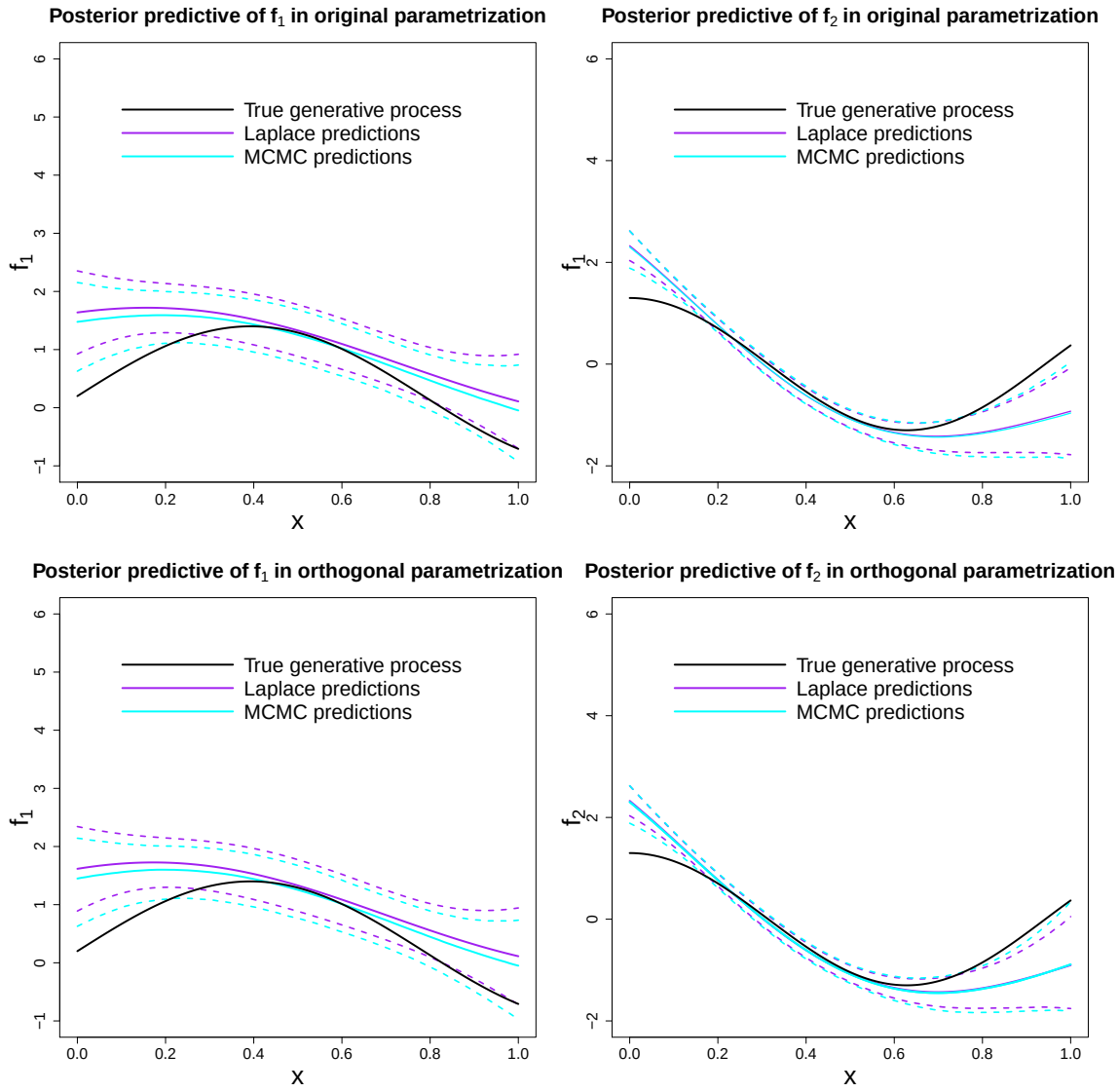


Figure 14: A plot of posterior predictive distributions for smooth data. First row denoting the original parametrization and second row denoting the orthogonal parametrization. The columns stand for different processes. The dashed lines stand for 2.5% and 97.5% percent quantiles for respective posterior distributions. The process that generated the data is plotted in black. Notice that these distributions are only compared in the (a) plots of Results section. In the (b) plots the $\boldsymbol{\eta}_2$ differs largely from the $\boldsymbol{f}_2$ and a plot showing both of them is not informative.

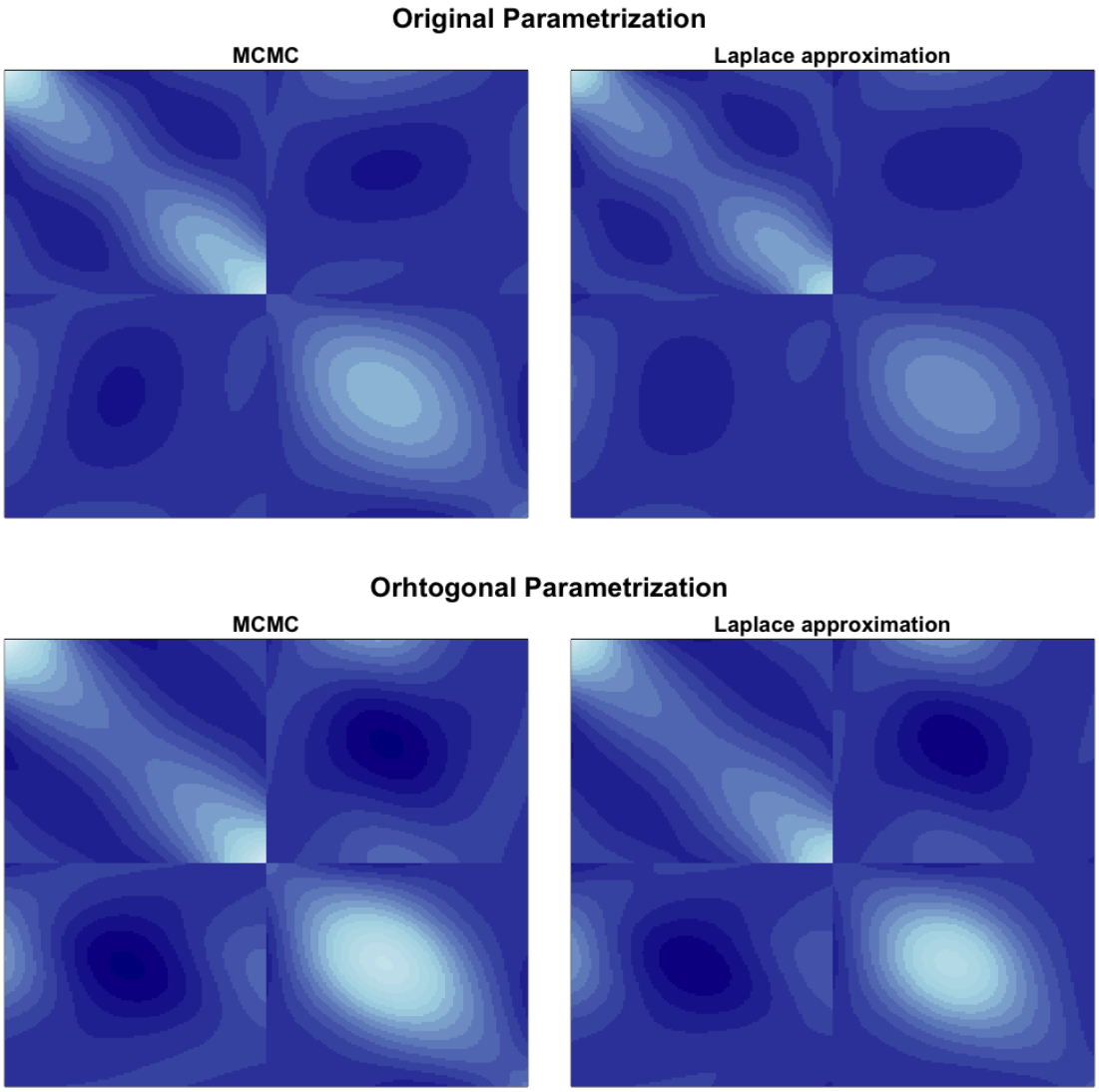


Figure 15: A plot of posterior predictive covariance matrices for smooth data for latent processes $\boldsymbol{f}$. Notice that the covariance matrix is block diagonal, where first diagonal block determine the amount of dependence within the first process. Similarly second element in the diagonal of the block diagonal matrix stands for the covariance matrix for posterior of the second process. The thing to notice is that they also share information between the two processes. Even that they are independent a priori, they become dependent in posterior distribution. The covariance comparison in Figure 11 is formed by comparing these covariance matrices row-wise for each N and for each iteration with N. The color scale goes from blue to white in ascending order. Again the (a) plots in Results section are the plots where these kind of matrices are compared. The (b) plots both of the parametrizations stay in their native spaces and a comparison between them is not meaningful.

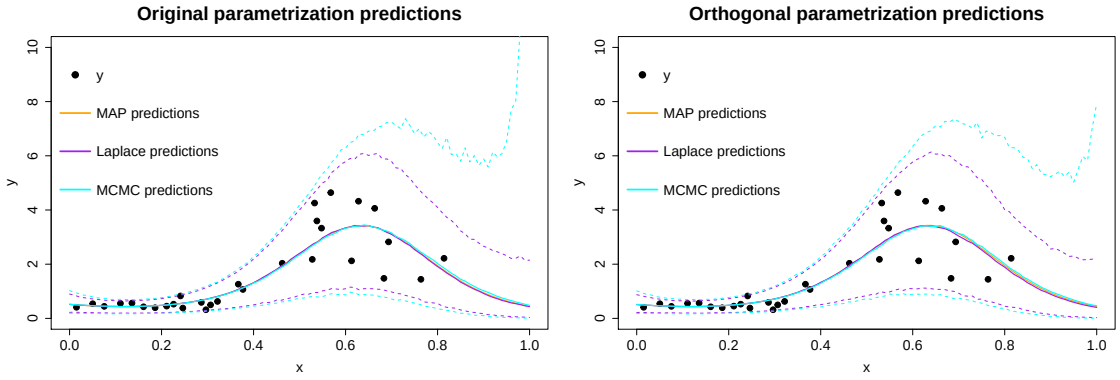


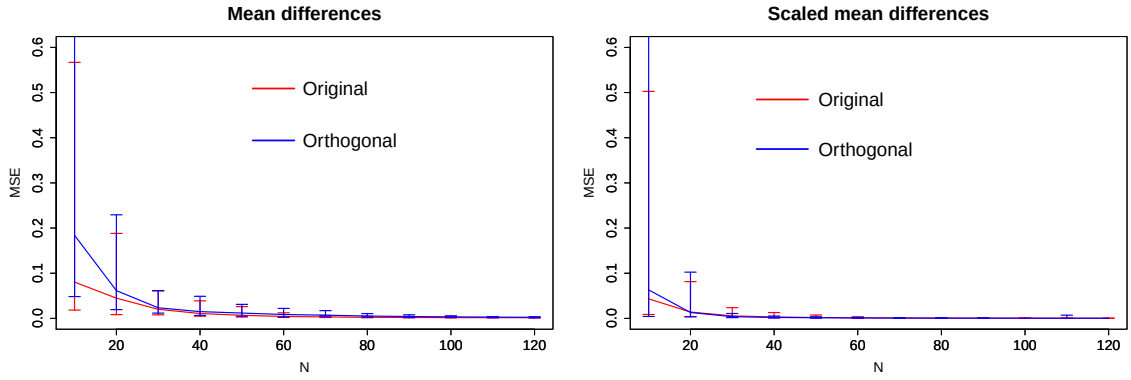
Figure 16: (a) Posterior predictive distribution for the response variable y in original parametrization. An example with $N = 30$ shows the median as a solid line and the 95% quantiles of the distribution. MAP predictions are plotted to emphasize the fact that many different kind of predictions are plausible, and in this case it agrees with the median of the MCMC and the mean of the Laplace approximation. (b) Posterior predictive distribution for the response variable in Orthogonal parametrization.

stands for a point estimate made through the expectation $E[\boldsymbol{y}_*|\hat{\boldsymbol{f}}]$ and is presented here as a alternative method to produce points predictions. Laplace predictions and MCMC predictions are produced as described in Section 4.2. The two dashed lines represent the 95% quantile of posterior predictive distribution $p(y_*|\boldsymbol{X}, \boldsymbol{y}, x_*)$ for each x_*. Notice that these plots are not compared in the thesis and they are presented for the readers interested in the applications of Gaussian processes.

5.2 Step data

The same comparisons and examples that were given for the smooth data are also given for the step data. The first comparison is again the mean squared errors of the posterior predictive means. Figure 17 shows the comparison between the two parametrizations. In Figure 17(a) the means are compared in the $\boldsymbol{f}$-space and it appears that the original parametrization achieves lower MSE measured by the median of the errors across the 100 repetitions. The confidence interval of the original parametrization is also narrower at all sample sizes N. Both of the parametrizations converge to the same region. However in Figure 17(b) the means are compared in their native spaces, and the MSE is normalized with the squared L2-norm of the MCMC posterior predictive mean. The plot shows that the orthogonal performs slightly better from $N = 30$ onwards at least in terms of narrower confidence interval. However the performances are practically equivalent.

The squared Frobenius norms of covariance matrices' elementwise difference are visible in Figure 18. In Figure 18(a) the orthogonal parametrization achieves lower differences between the MCMC and Laplace approximated covariance than orthogonal parametrization in terms of median differences apart from when $N = 10$. Orthogonal parametrization confidence intervals also seem to shrink faster than the confidence



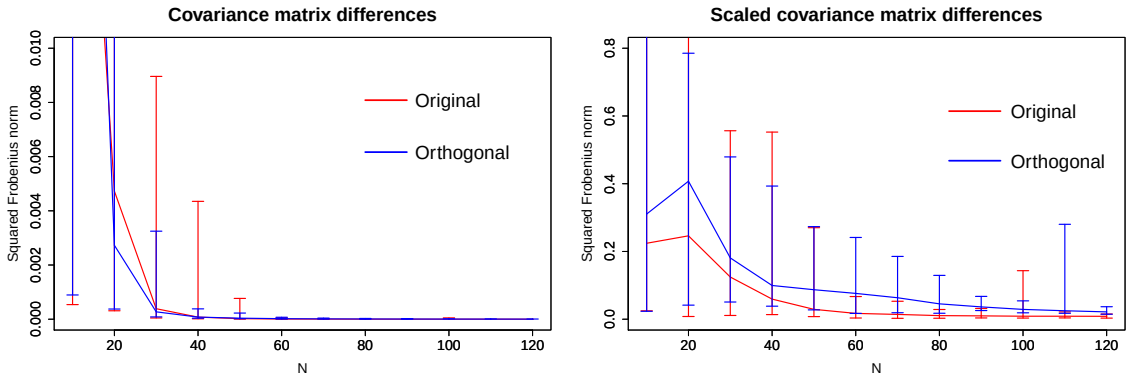
(a) (b)

Figure 17: (a) Mean squared errors of the posterior predictive means of the step data for latent processes. The different colors correspond to different parametrization's differences. Notice that the orthogonal parameters are transformed to $\boldsymbol{f}$-space before comparison is done. (b) Same statistic as in the (a) with the exception that orthogonal posterior predictive distribution is not mapped to $\boldsymbol{f}$-space, and instead is compared in $\boldsymbol{\eta}$-space. The difference is then normalized with squared L2-norm of the MCMC mean.

intervals for original parametrization. As the N grows, both of the parametrizations seem to converge close to each other. In Figure 18(b) orthogonal parametrization performs worse. The confidence interval is wider at most sample sizes and the median is above the original parametrization's counterpart.

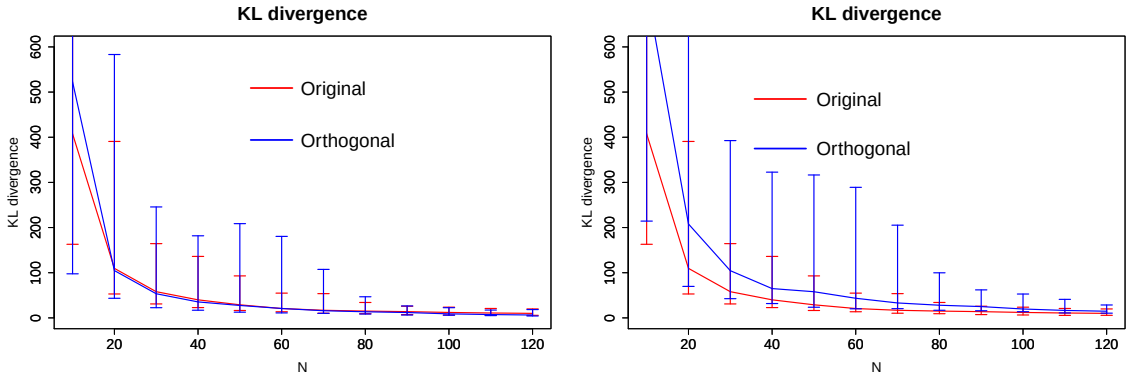
The posterior predictive distributions' KL-divergence comparison is presented in Figure 19. In Figure 19(a) the behavior is slightly different from the smooth data as neither parametrization seems to outperform the other significantly in terms of median of the divergences. However the confidence intervals of divergences are wider in the orthogonal parametrization until when $N = 90$ after which the performance is almost identical. If the plot is examined very precisely, then the orthogonal median is lower after $N = 10$ and the quantiles are also lower with large sample size. However these remarks are only marginal and probably reveal more about the nature of randomness than the posterior predictive accuracy. In Figure 19(b) the KL-divergence is computed in the native space of the parametrization and no mapping is done for the orthogonal posterior predictive distribution. The KL-divergence here is much worse in the orthogonal parametrization. The median is higher and the confidence intervals are much wider.

Finally, the computing times of the posterior MAP estimate in the aforementioned comparisons are presented in Figure 20. The figure supports the theoretical result that the orthogonal parametrization should perform faster due to its diagonal Fisher information matrix in natural gradient adaptation. However the difference is not as clear as it was with the smooth data, and the confidence intervals cross each other with the majority of N. Median values still favor orthogonal parametriza-



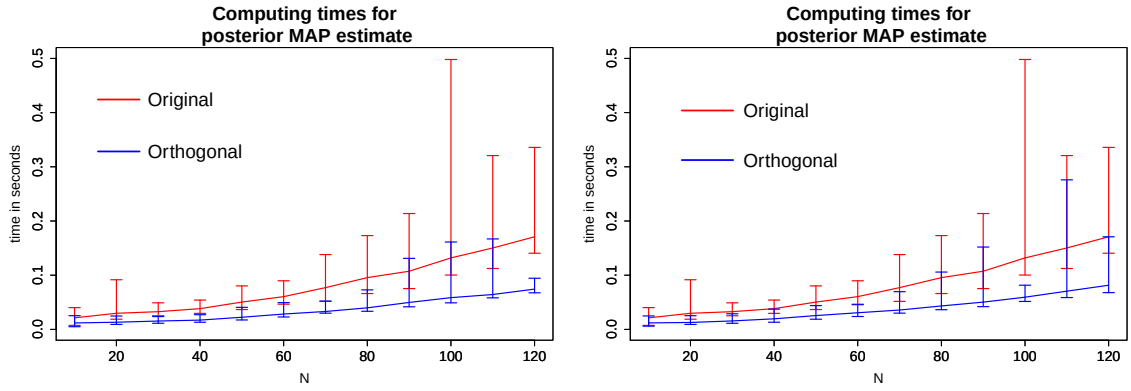
(a) (b)

Figure 18: (a) Squared Frobenius norm of the difference between Laplace approximation's posterior predictive covariance and posterior predictive sample covariance of MCMC samples. The orthogonal parametrization's posterior predictive distribution is transformed from $\boldsymbol{\eta}$ to $\boldsymbol{f}$-space before the comparison is done. (b) The squared Frobenius norm of the difference between MCMC and Laplace approximated posterior predictive covariances. The value is then normalized with the squared Frobenius norm of the MCMC posterior predictive covariance.



(a) (b)

Figure 19: (a) Approximative KL-divergence between the Laplace approximation and MCMC posterior predictive distributions with the step data. The approximative KL-divergence was presented in Chapter 4, and the approximation is done for both parametrizations, denoted with different colors. The orthogonal posterior is mapped back to the original parametrization's space before the comparison. (b) KL-divergence between the Laplace approximated and MCMC approximated posterior predictive distributions. Notice that the orthogonal parametrization is not mapped to $\boldsymbol{f}$-space and instead the comparison is done in $\boldsymbol{\eta}$-space.



(a) (b)

Figure 20: Wall clock times of finding the maximum of posterior distribution using natural gradient adaptation with the step data. Different colors denote the different parametrizations. Again (a) and (b) just show different passes over the same data set, where in (a) the posterior predictive is mapped to the same space with original parameters. However the procedure does not affect the computing times of posterior MAP and hence these plots are showing the trend of computing times over two computations of the same thing.

tion as being faster than the original parametrization. Notice that there are two saved computing times, one for each comparison method. Actually only the orthogonal parametrization was computed again, so the computing times for original parametrization should be the same.

For completeness, the example plots of posterior predictive distribution of latent processes are presented in Figure 21 for the predictive distribution quantiles, and in Figure 22 for the full covariance matrix. Notice that in these plots the $\boldsymbol{\eta}$-space is transformed to $\boldsymbol{f}$-space for the orthogonal parametrization.

Finally the posterior predictive distributions for the response variable are visible in Figure 23(a) for original parametrization and Figure 23(b) for orthogonal parametrization. Like before, the solid lines in Laplace and MCMC denote for the median values and the dashed lines denote the central 95% confidence interval of the posterior predictive distribution.

6 Discussion

The hypothesis of this thesis was that diagonalizing the Fisher information would not only increase the speed of computation, but also the rate of convergence towards a normal distribution. Hence, the Laplace approximation would achieve more accurate results in orthogonal case where the posterior would be a product of two almost normally distributed distributions. Because Laplace approximation is a normal approximation, and thus the predictive distributions for latent process and for

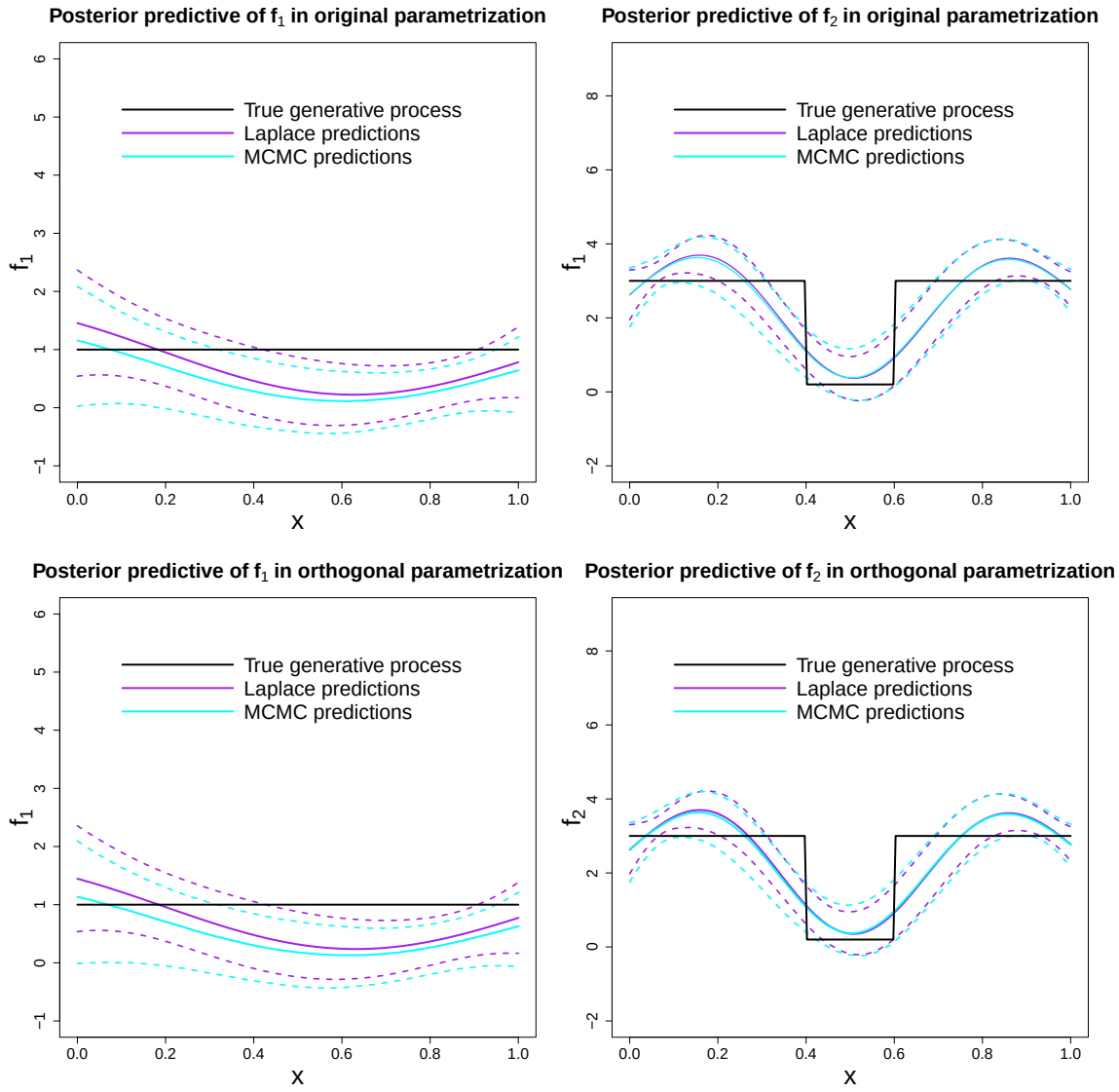


Figure 21: A plot of posterior predictive distributions for the step data. The first process is flat and the posterior predictive distribution is not fitting very well as it would require very large length scale to locally approximate the straight line. For the second process the length scale should be very small for posterior predictive distribution to be able to follow sharp fluctuations. However the hyperprior for the length scale is discouraging both of these values for length scale and thus the posterior predictive is trying to balance with the fit.

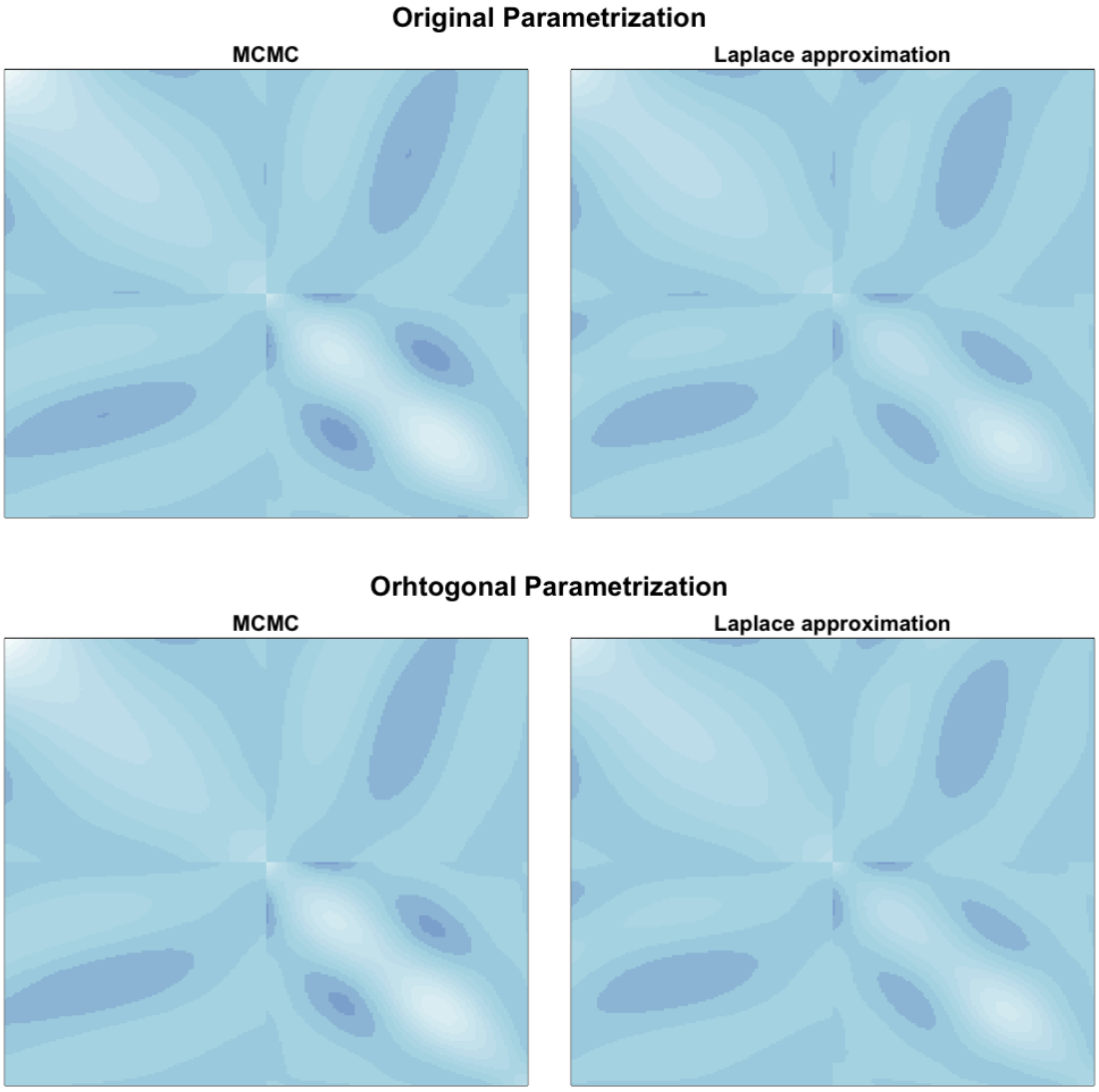
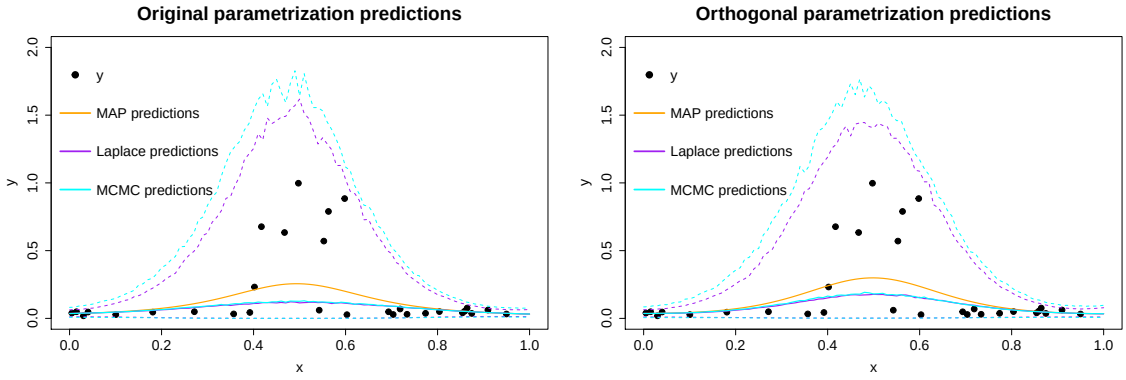


Figure 22: An example plot of posterior predictive covariance matrices for the step data when $N = 30$. The block diagonal structure comes from the fact that the processes $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ are treated as concatenated vector $\boldsymbol{f} = [\boldsymbol{f}_1, \boldsymbol{f}_2]$. The color scale goes from dark blue to white.



(a) (b)

Figure 23: (a) Original parametrization's posterior predictive distributions for the step data. Notice how the variance gets higher in the middle, but in a continuous nature, whereas the applications would wish for more accurate step in the middle to avoid misinterpretations of the situation. (b) The same predictions as made in (a) but in orthogonal parametrization.

the response variable are also normal, then having an observation model as close to normal distribution as possible would seem like an attractive approach for accurate predictions.

The question of the effect of reparametrization was studied in this thesis in empirical way, focusing only on two different kind of specific data generating processes. The smooth data having a smooth generative processes, and the step data having a constant function and a step function as a generative processes were used for the generation of the data. For both of these generative processes, a large amount of different datasets was generated by drawing samples from a Weibull distribution. The Weibull distribution was also chosen to be the observation model for the Gaussian process as well. The accuracy of the Laplace approximation was examined in a number of ways. The key concept behind the comparisons was to compare how close the Laplace approximation is to the true posterior distribution. Since the posterior distribution is not formed through conjugate prior and likelihood, it does not match any familiar distribution with a closed form probability density function. The true posterior distribution was therefore approximated by the most accurate way known, Markov chain Monte Carlo approximation. Then the comparison measures were the mean, covariance and the KL-divergence between Laplace approximated posterior predictive distribution and MCMC approximated posterior predictive distribution. All of the comparison measures were compared in two different ways, one where the comparison of the posterior predictive distributions is done in the same parameter space, having orthogonal posterior predictive samples mapped to original parameter space before the comparison. Another way was to compare the posterior predictive distributions of different parametrization in their native parameter spaces and normalize the comparison to be comparative. A smaller difference between the MCMC and the Laplace approximation would then denote a better approximation by the

Laplace method since it would be closer to the approximation of the truth, that is, the MCMC approximation.

In the example cases that were examined in this thesis, the orthogonal parametrization's effect on the accuracy is debatable. Different datasets and different comparison methods produced different kind of behavior. For example with the smooth data, the performance of the orthogonal parametrization was worse measured in mean and covariance difference, but better measured in KL-divergence when the comparison was done in the same parameter space. When the comparison was done in the native parameter spaces, the mean was better in orthogonal parametrization, but the covariance and KL-divergence was worse. When the data was generated by the process having the step function and the comparison was done in the same parameter space, the orthogonal mean was worse, orthogonal covariance was better and KL-divergence showed no significant difference. When the comparison was done in the native parameter spaces, the mean showed no significant difference but orthogonal covariance and KL-divergence were worse. One thing is sure, the accuracy of the Laplace approximation was not significantly reduced by the orthogonalization in terms of KL-divergence when the comparison was done in the same space, so in the context of predictive accuracy, the two parametrizations seemed to perform at least equivalently. When the parametrizations were compared in their native spaces, the orthogonal parametrization performed worse in the case of step data.

The computing times on the other hand were decreased significantly by the orthogonalization. This result was of course very anticipated, as the natural gradient method requires an inversion of a matrix and if the matrix is diagonal the inversion process is $O(N)$ instead of $O(N^3)$. Looking at the results for the computing times, the improvement does not seem to be as radical as going from $O(N^3)$ to $O(N)$. The reason for this is that the bottleneck of the computation is not only the inversion of the matrix, but also the matrix multiplications involved in the algorithm. Therefore, the improvement one gets from the orthogonal parametrization limits to the improvement by constant factors of the time complexity analysis. In practice this improvement might be significant, but in theory the improvement is considered to be marginal.

The overall conclusion of this theis is that one would get a faster algorithm by orthogonalizing the observation model's parameters, with reduction of predictive performance only on hard datasets. In fact it is possible that the accuracy can get better at least with modest sample sizes. However it appears that the orthogonalization could lead to instability of predictions at least with non-smooth generating processes. One should also note that the above conclusions are only made from empirical study and they should not be generalized without further examination in more general context.

The hypothesis of this thesis relied on a proposition that the orthogonalization of the parameters would improve the "normality" of the posterior and therefore the accuracy of the Laplace approximation (Hartmann & Vanhatalo, 2018). The proposition can be somewhat justified through the asymptotic behavior of the maximum

likelihood estimator. In my opinion there is a more important factor to the discussion which is rather optimization based: with a lot of freedom when making the assumptions when deriving the orthogonal parameters, one should study the possibility of making these choices so that they would be optimal in the sense of the normality of the posterior distribution. The choices that are being targeted here are for example that $\alpha_1(\eta_1) = \exp(\eta_1)$ and the initial value for the differential equation. If these choices have an effect on the normality of the posterior, they should indeed be chosen to increase the normality.

Overall, comparing the parametrization to each other seems like a hard task. The Bayesian hierarchical framework does not ease the problem, as for example the choice of hyperparameter prior seemed to have a huge effect on the accuracies. Certainly the same hyperprior is not sufficient as different parametrization's latent processes require different properties in order to fit to the data. For example in this thesis the orthogonal parametrization required much larger variance parameter as the data favored large fluctuation in the y-axis.

The effect of the hyperparameters and parametrization is an interesting area of research that could play a role in increasing accuracy of the approximations and decreasing the numerical issues and computation times. This area of research should not be overlooked, but seen as a possibility for easy improvements without the need for new data or other costly procedures.

The comparison measures of this work can be criticized. The differences between means and covariances might not reveal the accuracy of the posterior predictive distribution especially in high dimensions where the distributions can vary even if the mean and covariances match. For this reason the KL-divergence should be the measure to focus on, but since the KL-divergence treated in this thesis is approximative, one could argue about its accuracy. How well does the normal distribution parametrized with sample mean and covariance of the MCMC chains present the actual posterior? Both of these are valid points and they should indeed be studied in more detail and in a context not as empirical as in this thesis. To address this, other diagnostics were looked into in local experiments. The diagnostics were the non-normality measures from Kass & Slate (1994). However, these diagnostics were dropped as they were computationally so heavy that the computations would have taken way too long. The computing of the results for this thesis took two days for the smooth data and three days for the step data, which were not optimal for many iterations that obtaining the results required.

7 Conclusions

The aim of this study was to answer the question if reparametrization affects the performance of Laplace approximation in the Weibull observation model. The hypothesis was tested on two different generative processes by comparing the closeness of Laplace approximated posterior predictive distribution and the true posterior distribution. The true posterior distribution also required an approximation itself,

which was done by Markov chain Monte Carlo sampling. The conclusion of this thesis is that the reparametrization of the observation model decreases the computing time required for the Laplace approximation. The predictive preformance of the orthogonal parametrization showed mixed performance, having better accuracy in some measures and worse in others. Therefore, it is not possible to conclude that the orthogonalization would improve the Laplace approximation's accuracy in all cases.

Appendixes

A Fisher information for Weibull distribution

Consider the density function of a random variable X that follows a Weibull distribution with parameters α_1 and α_2

$$f_X(x) = \begin{cases} \alpha_1\alpha_2 \left(a_2 x\right)^{\alpha_1 - 1} \exp\left(-(\alpha_2 x)^{\alpha_1}\right) & \text{, if } x \geq 0 \\ 0 & \text{, otherwise.} \end{cases}$$

Then the Fisher information $\mathcal{I}$ with elements

$$\mathcal{I}_{i,j} = \mathbb{E}\left[\left(\frac{\partial}{\partial\alpha_i}\log(p(y|\boldsymbol{\alpha})\right)\left(\frac{\partial}{\partial\alpha_j}\log(p(y|\boldsymbol{\alpha})\right)\right] = \mathbb{E}\left[-\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}\log(p(y|\boldsymbol{\alpha})\right], \tag{61}$$

can be denoted by matrix notation with

$$\mathcal{I} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}. \tag{62}$$

In order to achieve the analytical solutions of the Fisher information matrix's terms some intermediate results are needed. First, the moment generating function of the logarithm of Weibull distribution is required, which is

$$\mathbb{E}\left[X^k\right] = \mathbb{E}\left[\exp\left(k\log X\right)\right] = \frac{\Gamma\left(\frac{k}{\alpha_1}+1\right)}{\alpha_2^k}. \tag{63}$$

Then, two more detailed results are required. First one of these is that

$$\begin{aligned}
\mathbb{E}\left[\log(X)X^k\right] &= \int_0^\infty \log(x)x^k\alpha_1\alpha_2(\alpha_2 x)^{\alpha_1-1}\exp(-(\alpha_2 x)^{\alpha_1})dx \\
&= \int_0^\infty \log\left(\frac{\mu^{\frac{1}{\alpha_1}}}{\alpha_2}\right)\frac{\mu^{\frac{k}{\alpha_1}}}{\alpha_2^k}\alpha_1\alpha_2\left(\mu^{\frac{1}{\alpha_1}}\right)^{\alpha_1-1}\exp(-\mu)\frac{\mu^{\frac{1}{\alpha_1}-1}}{\alpha_1\alpha_2}d\mu \\
&= \int_0^\infty \log\left(\frac{\mu^{\frac{1}{\alpha_1}}}{\alpha_2}\right)\frac{\mu^{\frac{k}{\alpha_1}}}{\alpha_2^k}\exp(-\mu)d\mu \\
&= \frac{1}{\alpha_1\alpha_2^k}\int_0^\infty \mu^{\frac{k}{\alpha_1}}\exp(-\mu)\log\left(\frac{\mu}{\alpha_2^{\alpha_1}}\right)d\mu \\
&= \frac{1}{\alpha_1\alpha_2^k}\int_0^\infty \mu^{\frac{k+\alpha_1}{\alpha_1}-1}\exp(-\mu)\left(\log\mu - \log(\alpha_2^{\alpha_1})\right)d\mu \\
&= \frac{1}{\alpha_1\alpha_2^k}\left(\Gamma^{(1)}\left(\frac{k+\alpha_1}{\alpha_1}\right) - \log\left(\alpha_2^{\alpha_1}\right)\Gamma^{(0)}\left(\frac{k+\alpha_1}{\alpha_1}\right)\right),
\end{aligned} \tag{64}$$

where on the second equality we used the change of variables with

$$\begin{aligned}
\mu &= (\alpha_2 x)^{\alpha_1} \iff x = \frac{\mu^{\frac{1}{\alpha_1}}}{\alpha_2} := \phi(\mu) \\
\implies \phi'(\mu) &= \frac{\mu^{\frac{1}{\alpha_1}-1}}{\alpha_1\alpha_2}
\end{aligned}$$

and on the last equality the property of derivatives of the gamma function

$$\frac{\partial^n}{\partial x^n}\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}(\log t)^n dt$$

were used.

The second result is similar to the first result, but the logarithm in the expectation is squared, yielding to the result

$$\begin{aligned}
\mathbb{E}\left[\log^2(X)X^k\right] &= \int_0^\infty \log^2(x)x^k\alpha_1\alpha_2(\alpha_2 x)^{\alpha_1-1}\exp(-(\alpha_2 x)^{\alpha_1})dx \\
&= \int_0^\infty \log^2\left(\frac{\mu^{\frac{1}{\alpha_1}}}{\alpha_2}\right)\frac{\mu^{\frac{k}{\alpha_1}}}{\alpha_2^k}\alpha_1\alpha_2\left(\mu^{\frac{1}{\alpha_1}}\right)^{\alpha_1-1}\exp(-\mu)\frac{\mu^{\frac{1}{\alpha_1}-1}}{\alpha_1\alpha_2}d\mu \\
&= \frac{1}{\alpha_1^2\alpha_2^k}\int_0^\infty \mu^{\frac{k}{\alpha_1}}\exp(-\mu)\log^2\left(\frac{\mu}{\alpha_2^{\alpha_1}}\right)d\mu \\
&= \frac{1}{\alpha_1^2\alpha_2^k}\int_0^\infty \mu^{\frac{k+\alpha_1}{\alpha_1}-1}\exp(-\mu)\left(\log^2\mu - 2\log(\alpha_2^{\alpha_1})\log(\mu)\right. \\
&\qquad\qquad \left. + \log^2(\alpha_2^{\alpha_1})\right)d\mu \\
&= \frac{1}{\alpha_1^2\alpha_2^k}\left(\Gamma^{(2)}\left(\frac{k+\alpha_1}{\alpha_1}\right) - 2\log(\alpha_2^{\alpha_1})\Gamma^{(1)}\left(\frac{k+\alpha_1}{\alpha_1}\right)\right. \\
&\quad \left. + \log^2\left(\alpha_2^{\alpha_1}\right)\Gamma^{(0)}\left(\frac{k+\alpha_1}{\alpha_1}\right)\right),
\end{aligned} \tag{65}$$

which is achieved with the same treatment as the first one.

With these results, the elements of the Fisher information matrix for Weibull distribution are straightforward to calculate.

Starting the calculations with a_{11}, the log-density is differentiated twice and the first derivative is given by

$$\frac{\partial}{\partial\alpha_1}\log f_X(x) = \frac{1}{\alpha_1} + \log\alpha_2 + \log x - \log(\alpha_2 x)(\alpha_2 x)^{\alpha_1}. \tag{66}$$

Differentiating the first gradient again will lead the second derivative to take the form of

$$\begin{aligned}
\frac{\partial^2}{\partial\alpha_1^2}\log f_X(x) &= -\frac{1}{\alpha_1^2} - \left(\log\alpha_2 + \log x\right)^2(\alpha_2 x)^{\alpha_1} \\
&= -\frac{1}{\alpha_1^2} - \left(\log^2\alpha_2 + 2\log\alpha_2\log x + \log^2 x\right)\alpha_2^{\alpha_1}x^{\alpha_1} \\
&= -\frac{1}{\alpha_1^2} - \alpha_2^{\alpha_1}\left(\log^2(\alpha_2)x^{\alpha_1} + 2\log\alpha_2\log(x)x^{\alpha_1} + \log^2(x)x^{\alpha_1}\right).
\end{aligned} \tag{67}$$

Now taking the expectation of the negative second derivative will lead to the final form of the first top left element of the Fisher information matrix

$$
\begin{aligned}
a_{11} &= \mathbb{E}\left[-\frac{\partial^2}{\partial\alpha_1^2}\log f_X(x)\right] \\
&= \frac{1}{\alpha_1^2} + \alpha_2^{\alpha_1}\left(\log^2(\alpha_2)\mathbb{E}\left[x^{\alpha_1}\right] + 2\log\alpha_2\mathbb{E}\left[\log(x)x^{\alpha_1}\right] + \mathbb{E}\left[\log^2(x)x^{\alpha_1}\right]\right) \\
&= \frac{1}{\alpha_1^2} + \alpha_2^{\alpha_1}\left(\log^2(\alpha_2)\frac{\Gamma(2)}{\alpha_2^{\alpha_1}} + 2\log\alpha_2\frac{1}{\alpha_1\alpha_2^{\alpha_1}}\left(\Gamma'(2) - \log(\alpha_2^{\alpha_1})\Gamma(2)\right)\right. \\
&\qquad\qquad \left. + \frac{1}{\alpha_1^2\alpha_2^{\alpha_1}}\left(\Gamma^{(2)}(2) - 2\log(\alpha_2^{\alpha_1})\Gamma^{(1)}(2) + \log^2(\alpha_2^{\alpha_1})\Gamma(2)\right)\right) \\
&= \frac{1}{\alpha_1^2} + \alpha_2^{\alpha_1}\left(\frac{\log^2\alpha_2}{\alpha_2^{\alpha_1}} + \frac{2\log(\alpha_2)\Gamma^{(1)}(2)}{\alpha_1\alpha_2^{\alpha_1}} - \frac{2\alpha_1\log^2(\alpha_2)}{\alpha_1\alpha_2^{\alpha_1}}\right. \\
&\qquad\qquad \left. + \frac{\Gamma^{(2)}(2)}{\alpha_1^2\alpha_2^{\alpha_1}} - \frac{2\log(\alpha_2^{\alpha_1})\Gamma^{(1)}(2)}{\alpha_1^2\alpha_2^{\alpha_1}} + \frac{\log^2(\alpha_2^{\alpha_1})}{\alpha_1^2\alpha_2^{\alpha_1}}\right) \\
&= \frac{1}{\alpha_1^2} + \log^2\alpha_2 + \frac{2\log(\alpha_2)\Gamma^{(1)}(2)}{\alpha_1} - 2\log^2\alpha_2 + \frac{\Gamma^{(2)}(2)}{\alpha_1^2} - \frac{2\log(\alpha_2)\Gamma^{(1)}(2)}{\alpha_1} + \log^2\alpha_2 \\
&= \frac{1}{\alpha_1^2}\left(1 + \Gamma^{(2)}(2)\right) \\
&= \frac{1}{\alpha_1^2}\left(1 + \Psi^{(1)}(2) + \Psi^2(2)\right) \\
&= \frac{1}{\alpha_1^2}\left(1 + \Psi^{(1)}(1) - 1 + \Psi^2(2)\right) \\
&= \frac{1}{\alpha_1^2}\left(\Psi^{(1)}(1) + \Psi^2(2)\right).
\end{aligned}
\tag{68}
$$

Notice that on the third equality we used the results (63), (64) and (65) defined above. The 4th equality follows from the fact that $\Gamma(1) = \Gamma(2) = 1$. The 7th equality arises from the relationship between Gamma function and Digamma function $\Psi(x)$

$$
\begin{aligned}
\Psi(x) &= \frac{\partial}{\partial x}\log\Gamma(x) = \frac{\Gamma^{(1)}(x)}{\Gamma(x)} \\
\implies \Psi^{(1)}(x) &= \frac{\Gamma^{(2)}(x)}{\Gamma(x)} - \Psi^2(x) \\
\iff \Gamma^{(2)}(x) &= \Gamma(x)\left(\Psi^{(1)}(x) + \Psi^2(x)\right),
\end{aligned}
$$

and the 8th equality follows from the property of the trigamma function $\Psi^{(1)}(z+1) = \Psi^{(1)}(z) - \frac{1}{z^2}$.

The second term $a_{12} = a_{21}$ follows from the similar straightforward, but laborious calculations. The first derivative is already derived in (66) and taking the derivative

with respect to α_2 yields to the following equation

$$\begin{aligned}
\frac{\partial^2}{\partial\alpha_1\partial\alpha_2}\log f_X(x) &= \frac{1}{\alpha_2} - x^{\alpha_1}\left(\frac{1}{\alpha_2 x}x\alpha_2^{\alpha_1} + \log(\alpha_2 x)\alpha_1\alpha_2^{\alpha_1-1}\right) \\
&= \frac{1}{\alpha_2} - x^{\alpha_1}\left(\alpha_2^{\alpha_1-1} + (\log(\alpha_2)+\log(x))\alpha_1\alpha_2^{\alpha_1-1}\right) \\
&= \frac{1}{\alpha_2} - x^{\alpha_1}\left(\alpha_2^{\alpha_1-1} + \log(\alpha_2)\alpha_1\alpha_2^{\alpha_1-1} + \log(x)\alpha_1\alpha_2^{\alpha_1-1}\right) \\
&= \frac{1}{\alpha_2} - \alpha_2^{\alpha_1-1}x^{\alpha_1} - \alpha_1\alpha_2^{\alpha_1-1}\log(\alpha_2)x^{\alpha_1} - \alpha_1\alpha_2^{\alpha_1-1}\log(x)x^{\alpha_1} \\
&= \frac{1}{\alpha_2} + \left(-\alpha_2^{\alpha_1-1} - \alpha_1\alpha_2^{\alpha_1-1}\log(\alpha_2)\right)x^{\alpha_1} - \alpha_1\alpha_2^{\alpha_1-1}\log(x)x^{\alpha_1}.
\end{aligned} \tag{69}$$

Taking the expectation of the above partial derivative multiplied by -1 will lead to the solution of non-diagonal elements of the Fisher information matrix

$$\begin{aligned}
a_{12} &= \mathbb{E}\left[-\frac{\partial^2}{\partial\alpha_1\partial\alpha_2}\log f_X(x)\right] \\
&= -\frac{1}{\alpha_2} - \left(-\alpha_2^{\alpha_1-1} - \alpha_1\alpha_2^{\alpha_1-1}\log(\alpha_2)\right)\mathbb{E}\left[x^{\alpha_1}\right] + \alpha_1\alpha_2^{\alpha_1-1}\mathbb{E}\left[\log(x)x^{\alpha_1}\right] \\
&= -\frac{1}{\alpha_2} - \left(-\alpha_2^{\alpha_1-1} - \alpha_1\alpha_2^{\alpha_1-1}\log(\alpha_2)\right)\frac{\Gamma(2)}{\alpha_2^{\alpha_1}} \\
&\quad + \alpha_1\alpha_2^{\alpha_1-1}\frac{1}{\alpha_1\alpha_2^{\alpha_1}}\left(\Gamma^{(1)}(2) - \log(\alpha_2^{\alpha_1})\Gamma(2)\right) \\
&= -\frac{1}{\alpha_2} + \frac{1}{\alpha_2} + \frac{\alpha_1\log(\alpha_2)}{\alpha_2} + \frac{1}{\alpha_2}\left(\Gamma^{(1)}(2) - \log(\alpha_2^{\alpha_1})\right) \\
&= \frac{\alpha_1\log(\alpha_2)}{\alpha_2} + \frac{\Gamma^{(1)}(2)}{\alpha_2} - \frac{\alpha_1\log(\alpha_2)}{\alpha_2} \\
&= \frac{\Gamma^{(1)}(2)}{\alpha_2} \\
&= \frac{\Psi(2)}{\alpha_2} \\
&= \frac{\Psi(1)+1}{\alpha_2},
\end{aligned} \tag{70}$$

which again uses the properties of digamma function and the results (63) and (64) presented at the beginning of this appendix.

The third and last element of the Fisher information matrix follows a similar procedure as the ones before it, first taking a derivative of the log density function with

respect to α_2

$$\begin{aligned}\frac{\partial}{\partial\alpha_2}\log f_X(x) &= \frac{1}{\alpha_2} + \frac{\alpha_1 - 1}{\alpha_2} - \alpha_1 \left(\alpha_2 x\right)^{\alpha_1 - 1} x \\ &= \frac{1}{\alpha_2} + \frac{\alpha_1 - 1}{\alpha_2} - \alpha_1 x^{\alpha_1} \alpha_2^{\alpha_1 - 1},\end{aligned} \tag{71}$$

and then the second derivative

$$\begin{aligned}\frac{\partial^2}{\partial\alpha_2^2}\log f_X(x) &= -\frac{1}{\alpha_2^2} - \frac{\alpha_1 - 1}{\alpha_2^2} - \alpha_1 x^{\alpha_1} \left(\alpha_1 - 1\right) \alpha_2^{\alpha_1 - 2} \\ &= -\frac{\alpha_1}{\alpha_2^2} - \alpha_2^{\alpha_1 - 2}\alpha_1^2 x^{\alpha_1} + \alpha_1 \alpha_2^{\alpha_1 - 2} x^{\alpha_1}.\end{aligned} \tag{72}$$

After this taking the negative expectation of the above equation will lead to the final form of bottom right element of the Fisher information matrix

$$\begin{aligned}a_{22} &= \mathbb{E}\left[-\frac{\partial^2}{\partial\alpha_2^2}\log f_X(x)\right] \\ &= \frac{\alpha_1}{\alpha_2^2} + \alpha_2^{\alpha_1 - 2}\alpha_1^2 \frac{1}{\alpha_2^{\alpha_1}} - \alpha_2^{\alpha_1 - 2}\alpha_1 \frac{1}{\alpha_2^{\alpha_1}} \\ &= \frac{\alpha_1^2}{\alpha_2^2}.\end{aligned}$$

These results were presented in Gupta & Kundu (2006) and they are revised with calculations in this thesis. With the justification of the above calculations, the analytical forms for elements of the Fisher information matrix (62) for Weibull distribution are

$$a_{11} = \frac{1}{\alpha_1^2}\left(\Psi^{(1)}(1) + \Psi^2(2)\right) \tag{73}$$

$$a_{12} = \frac{\Psi(1) + 1}{\alpha_2} \tag{74}$$

$$a_{22} = \frac{\alpha_1^2}{\alpha_2^2}. \tag{75}$$

B Predictive distribution given $\boldsymbol{f}$

Consider that a set of training inputs $\boldsymbol{f}$ at locations $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ have been obtained and the future interest is in the values of $\boldsymbol{f}_*$ at test locations $\boldsymbol{X}_* \in \mathbb{R}^{N_* \times p}$. The probabilistic predictions' target is to solve the posterior distribution of the test values $p(\boldsymbol{f}_* | \boldsymbol{X}_*, \boldsymbol{f})$. To achieve this, the focus is to first take a look of the joint distribution as it is just constant away from the posterior, when $\boldsymbol{f}_*$ is thought as the argument of

the joint density function. Prior belief is, that the distribution is normal regardless of the split between training and test points, the joint distribution is then

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{\boldsymbol{f}} \\ \mu_{\boldsymbol{f}_*} \end{bmatrix}, \underbrace{\begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) & K(\boldsymbol{X}, \boldsymbol{X}_*) \\ K(\boldsymbol{X}_*, \boldsymbol{X}) & K(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix}}_{:=\Sigma} \right). \tag{76}$$

Before jumping in to the result of the posterior predictive $p(\boldsymbol{f}_*|\boldsymbol{X}_*, \boldsymbol{f})$, some results are presented.

Lemma 2. *With all vectors u and v and for symmetric matrix Z, that is $Z = Z^T$, we have the following result*

$$\begin{aligned} u^T Z u - 2u^T Z v + v^T Z v &= (u-v)^T Z (u-v) \\ &= (v-u)^T Z (v-u). \end{aligned}$$

Proof.

$$\begin{aligned} u^T Z u - 2u^T Z v + v^T Z v &= u^T Z u - u^T Z v - u^T Z v + v^T Z v \\ &= u^T Z (u-v) - (u^T - v^T) Z v \\ &= u^T Z (u-v) - (u-v)^T Z v \\ &= u^T Z (u-v) - v^T Z (u-v) \\ &= (u^T - v^T) Z (u-v) \\ &= (u-v)^T Z (u-v), \end{aligned} \tag{77}$$

and the last equality of the lemma follows if the middle term in the beginning is rearranged as $-2u^T Z v = -2v^T Z u$, which is fine, because both are the same real number, and then doing the same steps of the proof with this term. □

Lemma 3. *For any block matrix $Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}$ the determinant has the following property*

$$\det(Z) = \det(Z_{11}) \det(Z_{22} - Z_{12}^T Z_{11}^{-1} Z_{12}).$$

Proof. Block matrix can be written as

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} = \begin{bmatrix} Z_{11} & 0 \\ Z_{12}^T & I \end{bmatrix} \begin{bmatrix} I & Z_{11}^{-1} Z_{12} \\ 0 & Z_{22} - Z_{12}^T Z_{11}^{-1} Z_{12} \end{bmatrix}. \tag{78}$$

Then, with the properties of determinant it is possible to write that the determinant of Z is

$$\begin{aligned} \det\left(\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \right) &= \det\left(\begin{bmatrix} Z_{11} & 0 \\ Z_{12}^T & I \end{bmatrix} \begin{bmatrix} I & Z_{11}^{-1} Z_{12} \\ 0 & Z_{22} - Z_{12}^T Z_{11}^{-1} Z_{12} \end{bmatrix} \right) \\ &= \det\left(\begin{bmatrix} Z_{11} & 0 \\ Z_{12}^T & I \end{bmatrix} \right) \det\left(\begin{bmatrix} I & Z_{11}^{-1} Z_{12} \\ 0 & Z_{22} - Z_{12}^T Z_{11}^{-1} Z_{12} \end{bmatrix} \right) \\ &= \det(Z_{11}) \det(Z_{22} - Z_{12}^T Z_{11}^{-1} Z_{12}). \end{aligned} \tag{79}$$

□

The following calculations will take up a lot of space and for that reason abbreviations for $\Sigma^{-1} = \begin{bmatrix} \tilde{A} & \tilde{C} \\ \tilde{C}^T & \tilde{B} \end{bmatrix}$ and $\Sigma = \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}$ are made. Deriving the posterior predictive distribution starts by rewriting the joint density function of $\boldsymbol{f}$ and $\boldsymbol{f}_*$

$$\begin{aligned} & p_{\boldsymbol{f},\boldsymbol{f}_*}(\boldsymbol{f},\boldsymbol{f}_*) \\ & = (2\pi)^{-n/2}\det(\Sigma)^{-1/2}\exp\left(-\frac{1}{2}\left(\begin{bmatrix}\boldsymbol{f}\\ \boldsymbol{f}_*\end{bmatrix}-\begin{bmatrix}\mu_{\boldsymbol{f}}\\ \mu_{\boldsymbol{f}_*}\end{bmatrix}\right)^T\begin{bmatrix}\tilde{A} & \tilde{C}\\ \tilde{C}^T & \tilde{B}\end{bmatrix}\left(\begin{bmatrix}\boldsymbol{f}\\ \boldsymbol{f}_*\end{bmatrix}-\begin{bmatrix}\mu_{\boldsymbol{f}}\\ \mu_{\boldsymbol{f}_*}\end{bmatrix}\right)\right) \\ & = (2\pi)^{-n/2}\det(\Sigma)^{-1/2}\exp\left(-\frac{1}{2}\left[(\boldsymbol{f}-\mu_{\boldsymbol{f}})^T \quad (\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*})^T\right]\begin{bmatrix}\tilde{A} & \tilde{C}\\ \tilde{C}^T & \tilde{B}\end{bmatrix}\begin{bmatrix}\boldsymbol{f}-\mu_{\boldsymbol{f}}\\ \boldsymbol{f}_*-\mu_{\boldsymbol{f}_*}\end{bmatrix}\right) \\ & := (2\pi)^{-n/2}\det(\Sigma)^{-1/2}\exp\left(-\frac{1}{2}Q(\boldsymbol{f},\boldsymbol{f}_*)\right). \end{aligned} \tag{80}$$

The term $Q(\boldsymbol{f},\boldsymbol{f}_*)$ can be presented in the form of

$$\begin{aligned} Q(\boldsymbol{f},\boldsymbol{f}_*) &= (\boldsymbol{f}-\mu_{\boldsymbol{f}})^T\tilde{A}(\boldsymbol{f}-\mu_{\boldsymbol{f}}) + (\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*})^T\tilde{C}^T(\boldsymbol{f}-\mu_{\boldsymbol{f}}) \\ &\quad + (\boldsymbol{f}-\mu_{\boldsymbol{f}})^T\tilde{C}(\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*}) + (\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*})^T\tilde{B}(\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*}) \\ &= (\boldsymbol{f}-\mu_{\boldsymbol{f}})^T\tilde{A}(\boldsymbol{f}-\mu_{\boldsymbol{f}}) + 2(\boldsymbol{f}-\mu_{\boldsymbol{f}})^T\tilde{C}(\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*}) + (\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*})^T\tilde{B}(\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*}), \end{aligned} \tag{81}$$

where last equality follows from the fact that

$$\begin{aligned} (\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*})^T\tilde{C}^T(\boldsymbol{f}-\mu_{\boldsymbol{f}}) &= \left(\tilde{C}(\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*})\right)^T(\boldsymbol{f}-\mu_{\boldsymbol{f}}) \\ &= \left(\underbrace{(\boldsymbol{f}-\mu_{\boldsymbol{f}})^T\tilde{C}(\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*})}_{\in\mathbb{R}}\right)^T \\ &= (\boldsymbol{f}-\mu_{\boldsymbol{f}})^T\tilde{C}(\boldsymbol{f}_*-\mu_{\boldsymbol{f}_*}). \end{aligned} \tag{82}$$

From inverse of partitioned matrices results (Gentle, 2007) we know that

$$\begin{aligned} \tilde{A} &= A^{-1} + A^{-1}CMC^TA^{-1} \\ \tilde{C} &= -A^TCM \\ \tilde{C}^T &= -MC^TA^{-1} \\ \tilde{B} &= M = (B - C^TA^{-1}C)^{-1}, \end{aligned}$$

which yields to $Q(\boldsymbol{f}, \boldsymbol{f}_*)$ to take the form of

$$\begin{aligned}
Q(\boldsymbol{f}, \boldsymbol{f}_*) &= (\boldsymbol{f} - \mu_{\boldsymbol{f}})^T \left[A^{-1} + A^{-1}CMC^TA^{-1}\right](\boldsymbol{f} - \mu_{\boldsymbol{f}}) \\
&\quad - 2(\boldsymbol{f} - \mu_{\boldsymbol{f}})^T A^{-1}CM(\boldsymbol{f}_* - \mu_{\boldsymbol{f}_*}) + (\boldsymbol{f}_* - \mu_{\boldsymbol{f}_*})^T M(\boldsymbol{f}_* - \mu_{\boldsymbol{f}_*}) \\
&= (\boldsymbol{f} - \mu_{\boldsymbol{f}})^T A^{-1}(\boldsymbol{f} - \mu_{\boldsymbol{f}})^T + \underbrace{(\boldsymbol{f} - \mu_{\boldsymbol{f}})^T A^{-1}C}_{u^T} M \underbrace{C^TA^{-1}(\boldsymbol{f} - \mu_{\boldsymbol{f}})}_{u} \\
&\quad -2(\boldsymbol{f} - \mu_{\boldsymbol{f}})^T A^{-1}CM \underbrace{(\boldsymbol{f}_* - \mu_{\boldsymbol{f}_*})}_{v} + (\boldsymbol{f}_* - \mu_{\boldsymbol{f}_*})^T M(\boldsymbol{f}_* - \mu_{\boldsymbol{f}_*}) \\
&= (\boldsymbol{f} - \mu_{\boldsymbol{f}})^T A^{-1}(\boldsymbol{f} - \mu_{\boldsymbol{f}})^T \\
&\quad + \left[(\boldsymbol{f}_* - \mu_{\boldsymbol{f}_*}) - C^TA^{-1}(\boldsymbol{f} - \mu_{\boldsymbol{f}})\right]^T M \left[(\boldsymbol{f}_* - \mu_{\boldsymbol{f}_*}) - C^TA^{-1}(\boldsymbol{f} - \mu_{\boldsymbol{f}})\right] \\
&= \underbrace{(\boldsymbol{f} - \mu_{\boldsymbol{f}})^T A^{-1}(\boldsymbol{f} - \mu_{\boldsymbol{f}})^T}_{:= Q_1(\boldsymbol{f})} \\
&\quad + \underbrace{\left[\boldsymbol{f}_* - (\mu_{\boldsymbol{f}_*} + C^TA^{-1}(\boldsymbol{f} - \mu_{\boldsymbol{f}}))\right]^T M \left[\boldsymbol{f}_* - (\mu_{\boldsymbol{f}_*} + C^TA^{-1}(\boldsymbol{f} - \mu_{\boldsymbol{f}}))\right]}_{:= Q_2(\boldsymbol{f}, \boldsymbol{f}_*)} \\
&= Q_1(\boldsymbol{f}) + Q_2(\boldsymbol{f}, \boldsymbol{f}_*).
\end{aligned} \tag{83}$$

Notice that on the third equality we used Lemma 2, and to emphasize it, the underbraces are used to state the correspondence in notation for u and v in the lemma. Finally, the joint probability density function can be written as

$$\begin{aligned}
&p_{\boldsymbol{f}, \boldsymbol{f}_*}(\boldsymbol{f}, \boldsymbol{f}_*) \\
&\quad = (2\pi)^{-n/2} \det(\Sigma)^{-1/2} \exp\left(Q_1(\boldsymbol{f}) + Q_2(\boldsymbol{f}, \boldsymbol{f}_*)\right) \\
&\quad = (2\pi)^{-n_{\boldsymbol{f}}/2}(2\pi)^{-n_{\boldsymbol{f}_*}/2} \det(A)^{-1/2} \det(M^{-1})^{-1/2} \exp(Q_1(\boldsymbol{f})) \exp(Q_2(\boldsymbol{f}, \boldsymbol{f}_*)) \\
&\quad = (2\pi)^{-n_{\boldsymbol{f}}/2} \det(A)^{-1/2} \exp(Q_1(\boldsymbol{f}))(2\pi)^{-n_{\boldsymbol{f}_*}/2} \det(M^{-1})^{-1/2} \exp(Q_2(\boldsymbol{f}, \boldsymbol{f}_*)) \\
&\quad = \mathcal{N}(\boldsymbol{f} | \mu_{\boldsymbol{f}}, A)\mathcal{N}(\boldsymbol{f}_* | b, M^{-1}),
\end{aligned} \tag{84}$$

where Lemma 3 was used on the second equality, and the mean of the second term is defined as $b \coloneqq \mu_{\boldsymbol{f}_*} + C^TA^{-1}(\boldsymbol{f} - \mu_{\boldsymbol{f}})$. Now it is possible to derive the posterior from this by first noticing that the marginal of $\boldsymbol{f}$ is

$$\begin{aligned}
p_{\boldsymbol{f}}(\boldsymbol{f}) &= \int p_{\boldsymbol{f}, \boldsymbol{f}_*}(\boldsymbol{f}, \boldsymbol{f}_*) d\boldsymbol{f}_* \\
&= \int \mathcal{N}(\boldsymbol{f} | \mu_{\boldsymbol{f}}, A)\mathcal{N}(\boldsymbol{f}_* | b, M^{-1}) d\boldsymbol{f}_* \\
&= \mathcal{N}(\boldsymbol{f} | \mu_{\boldsymbol{f}}, A).
\end{aligned} \tag{85}$$

With the above results, the posterior predictive distribution for latent process can

be written as

$$\begin{aligned} p_{\boldsymbol{f}_*|\boldsymbol{f}}(\boldsymbol{f}_*|\boldsymbol{f}) &= \frac{p_{\boldsymbol{f}_*,\boldsymbol{f}}(\boldsymbol{f}_*,\boldsymbol{f})}{p_{\boldsymbol{f}}(\boldsymbol{f})} \\ &= \frac{\mathcal{N}(\boldsymbol{f}|\mu_{\boldsymbol{f}},A)\mathcal{N}(\boldsymbol{f}_*|b,M^{-1})}{\mathcal{N}(\boldsymbol{f}|\mu_{\boldsymbol{f}},A)} \\ &= \mathcal{N}(\boldsymbol{f}_*|b,M^{-1}). \end{aligned} \tag{86}$$

Rewriting b and M^{-1} in the terms from covariance matrix from (76), and setting the prior means to $\mu_{\boldsymbol{f}} = 0$ and $\mu_{\boldsymbol{f}_*} = 0$, the analytical solution of posterior predictive distribution for latent process becomes

$$\begin{aligned} \boldsymbol{f}_* \ |\boldsymbol{X}_*,\boldsymbol{X},\boldsymbol{f} \sim \mathcal{N}\big(&K(\boldsymbol{X}_*,\boldsymbol{X})K(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{f}, \\ &K(\boldsymbol{X}_*,\boldsymbol{X}_*) - K(\boldsymbol{X}_*,\boldsymbol{X})K(\boldsymbol{X},\boldsymbol{X})^{-1}K(\boldsymbol{X},\boldsymbol{X}_*)\big). \end{aligned} \tag{87}$$

C Predictive statistics given $\boldsymbol{y}$

Posterior predictive mean for latent variable given the observations $\boldsymbol{y}$ is achieved by using Fubini's theorem and the analytical solution of posterior predictive of latent process given the latent process (87)

$$\begin{aligned} \mathbb{E}\left[f_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{x}_*\right] &= \int f_* p(f_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{x}_*)df_* \\ &= \int f_* \int p(f_*,\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{x}_*)d\boldsymbol{f}df_* \\ &= \int\int f_* p(f_*|\boldsymbol{X},\boldsymbol{f},\boldsymbol{x}_*)p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})df_*d\boldsymbol{f} \\ &= \int \left(\int f_* p(f_*|\boldsymbol{X},\boldsymbol{f},\boldsymbol{x}_*)df_*\right) p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})d\boldsymbol{f} \\ &= \int \mathbb{E}\left[f_*|\boldsymbol{f},\boldsymbol{X},\boldsymbol{x}_*\right] p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})d\boldsymbol{f} \\ &= \int k_*^T K^{-1}\boldsymbol{f}\, p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})d\boldsymbol{f} \\ &= k_*^T K^{-1}\int \boldsymbol{f}\, p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y})d\boldsymbol{f} \\ &= k_*^T K^{-1}\mathbb{E}\left[\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}\right]. \end{aligned} \tag{88}$$

The posterior predictive variance for latent is achieved through the law of total variance

$$\begin{aligned} Var\left[f_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{x}_*\right] &= \mathbb{E}_{\boldsymbol{f}}\left[Var\left[f_*|\boldsymbol{X},\boldsymbol{x}_*,\boldsymbol{f}\right]\right] + Var_{\boldsymbol{f}}\left[\mathbb{E}\left[f_*|\boldsymbol{X},\boldsymbol{x}_*,\boldsymbol{f}\right]\right] \\ &= \mathbb{E}_{\boldsymbol{f}}\left[k_{**} - k_*^T K^{-1}k_*\right] + Var_{\boldsymbol{f}}\left[k_*^T K^{-1}\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}\right] \\ &= k_{**} - k_*^T K^{-1}k_* + k_*^T K^{-1}Var_{\boldsymbol{f}}\left[\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}\right]K^{-1}k_* \\ &= k_{**} - k_*^T\left(K^{-1} - K^{-1}Var_{\boldsymbol{f}}\left[\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}\right]K^{-1}\right)k_*. \end{aligned} \tag{89}$$

D Kullback-Leibler divergence between two multivariate Gaussian distributions

Theorem 4. *Given two Gaussian distributions $p(x) \sim N(\mu_1, \Sigma_1)$ and $q(x) \sim N(\mu_2, \Sigma_2)$, the KL-divergence is analytical, given by*

$$KL(p||q) = \frac{1}{2}\left(\log\frac{\det\Sigma_2}{\det\Sigma_1} - d + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right), \quad (90)$$

where d is the dimension of the Gaussian distributions.

Proof. The proof follows from matrix algebra by the following steps

$$\begin{aligned}
KL(p||q) &= \int p(x)\log\frac{p(x)}{q(x)}dx \\
&= \int p(x)\log\left(\frac{\det(\Sigma_2)^{\frac{1}{2}}}{\det(\Sigma_1)^{\frac{1}{2}}}\cdot\frac{\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1))}{\exp(-\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2))}\right)dx \\
&= \frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \frac{1}{2}\int(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)p(x)dx \\
&+ \frac{1}{2}\mathbb{E}\left[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] \\
&= \frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \frac{1}{2}\mathbb{E}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right] \\
&+ \frac{1}{2}\mathbb{E}\left[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] \\
&= \frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \frac{1}{2}\mathbb{E}\left[\mathrm{Tr}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right]\right] \\
&+ \frac{1}{2}\mathbb{E}\left[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] \\
&= \frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \frac{1}{2}\mathbb{E}\left[\mathrm{Tr}\left[(x-\mu_1)(x-\mu_1)^T\Sigma_1^{-1}\right]\right] \\
&+ \frac{1}{2}\mathbb{E}\left[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] \\
&= \frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \frac{1}{2}\mathrm{Tr}[\mathbb{E}[\overbrace{(x-\mu_1)}^{\sim N(0,\Sigma_1)}(x-\mu_1)^T]\Sigma_1^{-1}] \\
&+ \frac{1}{2}\mathbb{E}\left[(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right] \\
&= \frac{1}{2}\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - \frac{1}{2}\mathrm{Tr}\left[\Sigma_1\Sigma_1^{-1}\right] \\
&+ \frac{1}{2}\left((\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2) + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)\right) \\
&= \frac{1}{2}\left(\log\frac{\det(\Sigma_2)}{\det(\Sigma_1)} - d + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right),
\end{aligned} \quad (91)$$

where on the 7th equality the equations (377) and (380) from Petersen & Pedersen (2012) were used. Notice also that the trace is linear operation and hence it is possible to change the order of expectation and trace. □

References

Amari, S., & Douglas, S. C. (1998, May). Why natural gradient? In *Proceedings of the 1998 ieee international conference on acoustics, speech and signal processing, icassp '98 (cat. no.98ch36181)* (Vol. 2, p. 1213-1216 vol.2). doi: 10.1109/ICASSP.1998.675489

Amari, S.-I. (1998, February). Natural gradient works efficiently in learning. *Neural Comput.*, *10*(2), 251–276. doi: 10.1162/089976698300017746

Billingsley, P. (1995). *Probability and measure*. Wiley.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, *76*(1), 1–32. doi: 10.18637/jss.v076.i01

Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, *49*(1), 1–39.

de Moivre, A. (1738). *The doctrine of chances*. Woodfall.

Durrett, R. (2011). *Probability: Theory and examples.*

Fletcher, R., & Reeves, C. M. (1964). Function minimization by conjugate gradients. *The Computer Journal*, *7*(2), 149-154. doi: 10.1093/comjnl/7.2.149

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis, third edition*. Taylor & Francis.

Gelman, A., & Rubin, D. B. (1992, 11). Inference from iterative simulation using multiple sequences. *Statist. Sci.*, *7*(4), 457–472. doi: 10.1214/ss/1177011136

Gentle, J. E. (2007). *Matrix algebra: Theory, computations, and applications in statistics* (1st ed.). Springer Publishing Company, Incorporated.

Geyer, C. J. (1992, 11). Practical markov chain monte carlo. *Statist. Sci.*, *7*(4), 473–483. doi: 10.1214/ss/1177011137

Gupta, R. D., & Kundu, D. (2006). On the comparison of fisher information of the weibull and ge distributions. *Journal of Statistical Planning and Inference*, *136*(9), 3130 - 3144. doi: https://doi.org/10.1016/j.jspi.2004.11.013

Hartmann, M., & Vanhatalo, J. (2018, Oct 17). Laplace approximation and natural gradient for gaussian process regression with heteroscedastic student-t model. *Statistics and Computing*. doi: 10.1007/s11222-018-9836-0

Higham, N. J. (2002a). *Accuracy and stability of numerical algorithms* (2nd ed.). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Higham, N. J. (2002b, July). Computing the nearest correlation matrix — problem from finance. *IMA Journal of Numerical Analysis*, *22*(3), 329-343. doi: 10.1093/imanum/22.3.329

Huzurbazar, V. S. (1956). Sufficient statistics and orthogonal parameters. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, *17*(3), 217–220.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in r*. Springer Publishing Company, Incorporated.

Jaynes, E., Jaynes, E., Bretthorst, G., & Press, C. U. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Kass, R. E., & Slate, E. H. (1994, 06). Some diagnostics of maximum likelihood and posterior nonnormality. *Ann. Statist.*, *22*(2), 668–695. doi: 10.1214/aos/1176325490

Kotilainen, M., Vanhatalo, J., Suominen, M., & Kujala, P. (2018). Predicting local ice loads on ship bow as a function of ice and operational conditions in the southern sea. *Ship Technology Research*, *65*(2), 87-101. doi: 10.1080/09377255.2018.1454390

Kullback, S., & Leibler, R. A. (1951, 03). On information and sufficiency. *Ann. Math. Statist.*, *22*(1), 79–86. doi: 10.1214/aoms/1177729694

MacKay, D. J. (1998, Oct 01). Choice of basis for laplace approximation. *Machine Learning*, *33*(1), 77–86. doi: 10.1023/A:1007558615313

O'Hagan, T. (2004). Dicing with the unknown. *Significance*, *1*(3), 132-133. doi: 10.1111/j.1740-9713.2004.00050.x

Peltola, T., Havulinna, A., Salomaa, V., & Vehtari, A. (2014, 01). Hierarchical bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. *CEUR Workshop Proceedings*, *1218*, 79-88.

Petersen, K. B., & Pedersen, M. S. (2012, nov). *The matrix cookbook*. Technical University of Denmark. (Version 20121115)

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes 3rd edition: The art of scientific computing* (3rd ed.). New York, NY, USA: Cambridge University Press.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning (adaptive computation and machine learning)*. The MIT Press.

Robert, C., & Casella, G. (2005). *Monte carlo statistical methods (springer texts in statistics)*. Berlin, Heidelberg: Springer-Verlag.

Robert, C., & Casella, G. (2011, 02). A short history of markov chain monte carlo: Subjective recollections from incomplete data. *Statist. Sci.*, *26*(1), 102–115. doi: 10.1214/10-STS351

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319-392. doi: 10.1111/j.1467-9868.2008.00700.x

Stirzaker, D. (2003). *Elementary probability* (2nd ed.). Cambridge University Press. doi: 10.1017/CBO9780511755309

Tenenbaum, M., & Pollard, H. (2012). *Ordinary differential equations*. Dover Publications, Incorporated.

Vanhatalo, J., Jylänki, P., & Vehtari, A. (2009). Gaussian process regression with student-t likelihood. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 1910–1918). Curran Associates, Inc.

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., & Vehtari, A. (2013). Gpstuff: Bayesian modeling with gaussian processes. *Journal of Machine Learning Research*, *14*(Apr), 1175–1179.