Efficient and Scalable Inference for Generalized Student-t Process Models

A thesis submitted to the University of Dublin, Trinity College
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Department of Statistics, Trinity College Dublin

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

December 2020

Gernot Roetzer

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

The copyright belongs jointly to the University of Dublin and Gernot Roetzer.

Gernot Roetzer

Dated: 25.03.2019

Abstract

Gaussian Processes are a popular, nonparametric modelling framework for solving a wide range of regression problems. However, they are suffering from 2 major shortcomings. On the one hand, they require efficient, approximate inference for non-Gaussian observation likelihoods (the Generalized Gaussian Process Regression problem) and, on they other hand, their cubic run time in the number of observations is a major obstacle to large-scale inference tasks.

In recent years, the development of efficient and scalable inference methods for the Generalized Gaussian Process Regression problem has progressed steadily. However, the more robust generalization of the Gaussian Process, the Student-t Process, while suffering under the same shortcomings, has not been given the same amount of attention with respect to more general likelihoods.

In this thesis, we utilize the mathematical framework of q-algebra to extend some of the efficient and scalable methods for Generalized Gaussian Process Regression to the case of Generalized Student-t Process Regression.

We demonstrate in experiments that some of our Student-t based methods can compete with their Gaussian counterparts and that they can be be more robust to mislabelled data. However, we also see that the new methods are suffering under severe convergence problems and need considerable effort to tune them properly.

Acknowledgements

A PhD is a daunting journey with up and downs. Many people have accompanied me on this endeavour, have helped me through the downs, have celebrated the ups with me. There were too many of them, who had an impact on me, too many to name them all without forgetting anyone. Therefore, I will forego a long list of names and extend my gratitude to all of them.

However, there are 3 people I would like to thank separately. My advisor Simon Wilson, without him, I would not have reached the point where I am today. My mother, who has always supported me unconditionally. My wife, who endures me bravely and brings light to my world.

Finally, I would also like to thank the Insight Centre for Data Analytics for their financial support that allowed me to embark on this journey.

Gernot Roetzer
Trinity College Dublin
December 2020

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Learning a latent function from training data is a common task in Machine Learning and Statistics. A nonparametric Bayesian solution to this problem is Gaussian Processes (GPs) (Rasmussen & Williams (2005)). By using GPs as a prior over the space of real-valued functions, they can be used to solve a variety of different types of problems like regression, classification, or even reinforcement learning (e.g. Rasmussen & Williams (2005) and Rasmussen et al. (2003)). Due to their flexibility, ease of conducting inference, and principled way of quantifying uncertainty, they enjoy ever increasing popularity.

A natural generalization of GPs is the Student-t Process (TP) (see Fang et al. (1990) or Shah et al. (2014)). While a GP neglects the uncertainty associated with choosing a covariance function, a parametric function that often solely defines the GP, a TP takes this uncertainty into account. This allows for a more genuine representation of the underlying uncertainty. A consequence of this is that the TP has heavier-tails than the GP, which makes the TP more robust against outliers in the data (e.g. Shah et al. (2014), Solin & Särkkä (2015)).

However, these beneficial properties come at a high price, TPs are generally not as mathematically tractable as GPs. This might explain why there has been limited work on extending some of the advancements for GP models of the last decade to TP models. In particular, important developments to improve the scalability of GPs (e.g. Titsias (2009)) and to solve more general regression problems efficiently (e.g. Shang & Chan (2013)) lack their corresponding TP couterpart.

In this thesis, we aim to close the gap between GPs and TPs when it comes to efficient and scalable inference. Our hope is that the methods presented in this thesis allow to utilize the TP and benefit from its properties in a broader range of applications.

1.1 Thesis Contributions

In this thesis, the following contributions have been made:

1. In Section 4.1, we introduce t-relexation, a simple, but general framework for obtaining tractable approximations for TP models based on q-algebra.
2. In Chapter 4, we develop the t-Laplace approximation, whose defining property is that, in contrast to the standard Laplace approximation, a multivariate Student-t distribution is used as approximate posterior distribution.
3. We introduce two variants of variational Student-t approximation in Chapter 5. These methods allow for efficient inference in latent Student-t models and are a conceptual generalization of variational Gaussian approximation (Opper & Archambeau (2009)).
4. Finally, in Chapter 6, we present the first scalable Generalized Student-t Process Regression (GTPR) models based on sparse inducing point methods.

1.2 Thesis Outline

- Chapter 2 gives an overview of the important concepts used in this thesis.
- Chapter 3 motivates the development of new methods based on an application of Laplace approximation.
- Chapter 4 presents the t-relaxation and develops the t-Laplace approximation.
- Chapter 5 shows the derivation of the variational Student-t approximations.
- Chapter 6 extends the concept of sparse inducing point methods to GTPR models.

- Chapter 7 compares the performace of the developed methods to their Gaussian counterparts.
- Chapter 8 provides an outlook on future research opportunities.
- Chapter 9 summarizes the thesis and gives some concluding remarks.

Chapter 2

Background

2.1 Gaussian Processes

2.1.1 Basics

A Gaussian Process (GP) is a stochastic process that has as defining property that any of its finite samples are distributed according to a multivariate Gaussian distribution (Rasmussen & Williams (2005)). That is, any finite sample has the following density:

$$p(\boldsymbol{f}) = \frac{1}{(2\pi)^{\frac{n}{2}} \det |\boldsymbol{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{f} - \boldsymbol{\mu})^T \boldsymbol{K}^{-1}(\boldsymbol{f} - \boldsymbol{\mu})\right), \tag{2.1}$$

where $\boldsymbol{\mu}$ is the mean of the distribution, n is the number of dimensions, and $\boldsymbol{K}$ is the covariance matrix. We abbreviate this expression with either $\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \boldsymbol{K})$ for multivariate Gaussian distributions or $\mathcal{GP}(\boldsymbol{f}; \boldsymbol{\mu}, \boldsymbol{K})$ for Gaussian Processes. A GP forms a distribution over the functions $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \to \mathbb{R}$.

The GP is specified by a mean function $\boldsymbol{\mu} : \boldsymbol{\mathcal{X}} \to \mathbb{R}$ and a covariance function $\boldsymbol{K} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}} \to \mathbb{S}^+$ for some input space $\boldsymbol{\mathcal{X}}$, whereas $\mathbb{S}^+$ is the space of symmetric, positive semi-definite matrices (potentially infinite dimensional). It is a common assumption to set the mean function to zero (Rasmussen & Williams (2005)). As a result, the GP is solely defined by the covariance function.

A covariance function that is often used (Rasmussen & Williams (2005)) and that we will use for our experiments as well, is the squared exponential covariance function:

$$\boldsymbol{K}_{ij} = s^2 \exp\left(-\frac{\boldsymbol{d}_{ij}^T \boldsymbol{d}_{ij}}{2l^2}\right), \tag{2.2}$$

with

$$\boldsymbol{d}_{ij} = \boldsymbol{x}_i - \boldsymbol{x}_j, \tag{2.3}$$

whereas s^2 is the (signal) variance parameter that controls the dispersion of samples from the Gaussian Process. The lengthscale l works as smoothing parameter, the higher l the smoother the functions that are sampled from the Gaussian Process. The different $\boldsymbol{x}$ represent our inputs. That is, given a finite collection of inputs $\boldsymbol{X} = \{x_i : i = 1, \ldots, n\}$ and $\widetilde{\boldsymbol{X}} = \{\tilde{x}_j : j = 1, \ldots, m\}$, $\boldsymbol{K}(\boldsymbol{X}, \widetilde{\boldsymbol{X}})$ returns an $n \times m$ dimensional covariance matrix, this is commonly abbreviate by $\boldsymbol{K_{nm}}$ (e.g. Titsias (2009), Hensman, Matthews & Ghahramani (2015)). Consequently, $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})$ is denoted by $\boldsymbol{K_{nn}}$. Frequently, we will drop the subscripts for the symmetric matrices, when their dimensionality is clear from the context. Importantly, while the bold subscripts refer to the dimensionality of the matrix, the non-bold subscript used in 2.3 denote individual elements of the matrix.

2.1.2 Classical Regression

The popularity of GPs stems, partly, from their simplicity in the classical regression case, i.e. real-valued observation with additive, Gaussian noise. That is, for the following model:

$$\boldsymbol{f} \sim \mathcal{GP}(\boldsymbol{0}, \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})) \tag{2.4}$$

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{f}, \sigma^2 \boldsymbol{I}), \tag{2.5}$$

we can obtain the prediction mean and prediction covariance for function values $\boldsymbol{f}_*$ at some new inputs x_* in closed form by:

$$\mathbb{E}[\boldsymbol{f}_* | \boldsymbol{X}, \boldsymbol{y}, x_*] = \boldsymbol{K_{n_* n}} (\sigma^2 \boldsymbol{I} + \boldsymbol{K_{nn}})^{-1} \boldsymbol{y} \tag{2.6}$$

$$\mathbb{V}[\boldsymbol{f}_* | \boldsymbol{X}, \boldsymbol{y}, x_*] = \boldsymbol{K_{n_* n_*}} - \boldsymbol{K_{n_* n}} \boldsymbol{K_{nn}^{-1}} \boldsymbol{K_{n n_*}}. \tag{2.7}$$

This is possible, because the observations and the function values are jointly multivariate Gaussian. As a result, by computing the expectation and variance of the conditional distribution, we have specified all the uncertainty associated with $\boldsymbol{f}_*$.

2.1.3 Model Selection

The fact that a GP can be fully specified by its covariance function reduces the problem of model selection to finding appropriate hyperparameters for the covariance function, e.g. finding an appropriate signal variance and lengthscale for the squared exponential kernel, and for (potentially) the observation model. This is usually done by optimizing the (log) marginal likelihood with respect to the hyperparameters. Where the log marginal likelihood is given by:

$$\log p(\boldsymbol{y}|\boldsymbol{X}) = \log \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f}. \tag{2.8}$$

For the classical regression case, the marginal likelihood is available in closed form as a multivariate Gaussian distribution (Rasmussen & Williams (2005)):

$$\log p(\boldsymbol{y}|\boldsymbol{X}) \sim \mathcal{MVN}(\boldsymbol{0}, \sigma^2\boldsymbol{I} + \boldsymbol{K_{nn}}). \tag{2.9}$$

Taking the logarithm gives:

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{X}) = & -\frac{1}{2}\boldsymbol{y}^T(\sigma^2\boldsymbol{I} + \boldsymbol{K_{nn}})^{-1}\boldsymbol{y} - \frac{1}{2}\log\det\left|\sigma^2\boldsymbol{I} + \boldsymbol{K_{nn}}\right| \\ & - \frac{n}{2}\log 2\pi,\end{aligned} \tag{2.10}$$

which needs to be optimized with respect to θ, the hyperparameters of the covariance function (e.g. lengthscale l and signal noise s^2 in 2.2), and the noise variance σ^2. In the literature, this approach is referred to as type-2 maximum likelihood estimation (Rasmussen & Williams (2005))[1].

2.1.4 Generalized Regression

Generalized Gaussian Process Regression (GGPR) generalizes the ordinary Gaussian Process Regression (GPR) by taking into account observation models that are

[1] In contrast to the type-1 maximum likelihood estimation, we marginalize over the other latent variables, i.e. $\boldsymbol{f}$, instead of using the full, unmarginalized likelihood for optimization.

not Gaussian (Shang & Chan (2013)). That is, a GGPR model has the following form:

$$\boldsymbol{f} \sim \mathcal{GP}(\mathbf{0}, \boldsymbol{K}) \tag{2.11}$$

$$\boldsymbol{y}_i \sim g(\boldsymbol{f}_i), \tag{2.12}$$

where $g(\cdot)$ is the observation model that is used to model, e.g., binary or count data. It is common to assume that the observations are independent given the unobserved function values (Shang & Chan (2013)).

For a concrete example, we can look at the GP binary classification problem, where the observation vector is a vector of $\{0, 1\}^n$and the observation model is a Bernoulli distribution in combination with a sigmoid function (see below) in order to map the random function $\boldsymbol{f}$ onto the range $(0, 1)$:

$$\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})) \tag{2.13}$$

$$\boldsymbol{y}_i \sim \mathrm{Ber}(\mathrm{sigm}(\boldsymbol{f}_i)), \tag{2.14}$$

with

$$\mathrm{sigm}(x) = \frac{1}{1 + \exp^{-x}}. \tag{2.15}$$

The main problem of GGPR models is that the conjugacy property of the Gaussian observation model is lost. That is, there are no closed form solutions for the conditional predictive distribution and the log marginal likelihood anymore. In the last two decades the literature on approximation methods for particular cases of GGPR models has grown steadily, e.g.:

- Early approaches in the geostatistics community used MCMC, see Diggle et al. (1998) for more details.

- An overview of approximation methods for the GP binary classification problem can be found in Nickisch & Rasmussen (2008).

- Lloyd et al. (2015) provides a variational algorithm for GP modulated Poisson Processes.

- Student-t distributions are used as observation model in Jylänki et al. (2011) and Vanhatalo et al. (2009).

- Khan et al. (2012) and Shang & Chan (2013) deal with efficient inference for general observation models.

The advantage of developing an algorithm for the abstract class of GGPR models, instead of a concrete example, is that the algorithm is applicable to many different models. In Section 2.3, we will present some methods that allow for efficient inference in the non-conjugate, generalized regression case in more detail. Nevertheless, this generality comes at the expense that a tailor-made approach is likely to be faster and more accurate.

However, the big bottleneck, in both generalized and ordinary regression, that prevents the application of GPs for large dataset is that they require the inversion of an $n \times n$ covariance matrix, $\boldsymbol{K_{nn}}$, during model selection/learning and prediction (Rasmussen & Williams (2005)). That is, the complexity of computing the posterior or marginal likelihood is cubic in the number of observations. While this is still a manageable computing load for a small number of observations (e.g. a thousand observations), it gets quickly prohibitive with increasing number of observations (e.g. ten or even a hundred thousand observations). In Section 2.3.4, we will give an overview over the most prominent methods to overcome this bottleneck, the sparse inducing point methods.

2.2 The Student-t Process

2.2.1 Basics

In contrast to the GP, the Student-t Process (TP) is a stochastic process that has as defining property that any finite collection of samples from the process is distributed according to a multivariate Student-t distribution. That is, any finite sample has the following density (Kotz & Nadarajah (2004)):

$$p(\boldsymbol{f}) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{\frac{n}{2}} \det|\boldsymbol{K}|^{\frac{1}{2}}} \left(1 + \frac{1}{\nu}(\boldsymbol{f}-\boldsymbol{\mu})^T \boldsymbol{K}^{-1}(\boldsymbol{f}-\boldsymbol{\mu})\right)^{-\frac{\nu+n}{2}}, \quad (2.16)$$

whereas ν is the degrees of freedom, $\boldsymbol{\mu}$ is the mean of the distribution, n is the number of dimensions, and $\boldsymbol{K}$ is a scaling or dispersion matrix. We abbreviate this expression with either $\mathcal{MVT}(\boldsymbol{f}; \nu, \boldsymbol{\mu}, \boldsymbol{K})$ for multivariate Student-t distributions or $\mathcal{TP}(\boldsymbol{f}; \nu, \boldsymbol{\mu}, \boldsymbol{K})$ for Student-t Processes.

The TP is obtained by placing an Inverse Wishart Process prior on the covariance function to model the uncertainty that is associated with choosing a particular kernel function (Shah et al. (2014)).

Compared to the GP, the TP has two interesting properties. On the one hand, due to the Student-t nature of any finite sample of the TP, the TP puts more probability mass into the tails of its finite sample distributions. Therefore, the mean function needs to shift less to explain outliers in the observed function values. From this perspective, the TP is more robust against outliers than the GP (Shah et al. (2014)).

On the other hand, the conditional predictive distribution of a multivariate Student-t distribution is multivariate Student-t and the covariances of the conditional distribution depends directly on the variables conditioned on, that is, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly multivariate Student-t (Shah et al. (2014) or Ding (2016)), then

$$\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{MVT}(\nu + n_x, \boldsymbol{\mu}_y + \boldsymbol{K}_{yx}\boldsymbol{K}_{xx}(\boldsymbol{x} - \boldsymbol{\mu}_x), \beta(\boldsymbol{K}_{yy} - \boldsymbol{K}_{yx}\boldsymbol{K}_{xx}^{-1}\boldsymbol{K}_{xy})) \tag{2.17}$$

$$\beta = \frac{\nu + (\boldsymbol{x} - \boldsymbol{\mu}_x)^T \boldsymbol{K}_{xx}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x)}{\nu + n_x}, \tag{2.18}$$

where the subscript indicates from which random variable the magnitude has to be taken. An important feature of the TP is that, in contrast to the GP, the conditional predictive covariance depends on the observed data $\boldsymbol{x}$. Shah et al. (2014) connect this feature to improved predictive covariances that they observe in their comparative experiments between GP and TP.

2.2.2 Classical and Generalized Regression

While the GP allows for a closed form solution of the classical regression problem (for conditional distribution and marginal likelihood), the TP does not share this property for its adjusted version, i.e. with Student-t errors. The problem is that Student-t distributions are not closed under addition (Kotz & Nadarajah (2004)).

Zhang & Yeung (2010) and Shah et al. (2014) overcome this problem by incorporating the additive noise term directly into the kernel function. Predictions are conducted by using the conditional distribution presented in the previous section.

The method has the shortcoming that the error term and the contribution of the kernel function are not independent, due to the degrees of freedom parameter ν (Shah et al. (2014)).

For the task of deriving methods for Generalized Student-t Process Regression (GTPR) of the form,

$$\boldsymbol{f} \sim \mathcal{TP}(\nu, \mathbf{0}, \boldsymbol{K}) \tag{2.19}$$

$$\boldsymbol{y}_i \sim g(\boldsymbol{f}_i), \tag{2.20}$$

the lack of conjugacy in the classical regression model is not that critical. It is just another potential test case for the methods.

2.2.3 Robustness

The term robustness is associated with many different definitions. E.g. Huber (2011) defines robustness as insensitivity to small deviations from the assumptions. For this thesis, the focus lies on one specific form of deviations: Outliers.

The usage of Student-t distributions as a means of mitigating the impact of outliers has a long history in the GP community. Neal introduced the Student-t distribution as an observation model for the GP regression problem (Neal (1997)). This approach was relying on MCMC. To improve on the high computational cost of MCMC, approximate methods like factorized variational inference (Tipping & Lawrence (2005)), expectation propagation (Jylänki et al. (2011)), or Laplace approximation (Vanhatalo et al. (2009), Mair & Brefeld (2018)) were introduced. Recent GP-based models that utilize a GP prior with a Student-t observation model are, e.g., Mattos et al. (2017) and Ranjan et al. (2016).

Generally, the application of Student-t distributions for data sets with outliers is often (e.g. Jylänki et al. (2011), Vanhatalo et al. (2009)) justified with the theoretical work of O'Hagan (1979). O'Hagan established that the Student-t distribution is outlier-prone, that is, an inference method applied onto a Student-t sample can ignore extreme outliers. In contrast, O'Hagan (1979) demonstrated that the Gaussian is outlier-resistant, which means that an inference method applied onto a Gaussian sample never rejects outliers.

Nevertheless, while the Student-t likelihood has been used extensively with GP priors in the ordinary regression case, robust methods for the generalized regression case are less prominent. For the classification case, robust methods were introduced by Kim & Ghahramani (2008) for the binary problem and by Hernández-Lobato et al. (2011) for the multi-class problem. Both methods rely on additional latent variables that represent the labelling error to obtain robustness. Another approach by L. Wauthier & Jordan (2010) for the regression and classification case used copulas to obtain a heavy-tailed process from the GP.

However, there has not been much effort in the literature in utilizing TPs for robust generalized regression. A notable exception to this is the work of Futami et al. (2017), where they developed an expectation propagation algorithm for the TP classification problem. They showed that the Student-t approach is more robust to outliers than its Gaussian counterpart in this case.

2.2.4 Student-t Process Based Models

The applications of TPs in the literature are dominated by 2 fields, multi-task learning and Bayesian optimization.

Multi-task learning (MTL) assumes that different tasks share some common structure and modelling them jointly improves the performance on the individual tasks (Zhang & Yang (2017)). That is, in MTL we have multiple sets[2] of observations and inputs and the TP approach to it is:

$$\boldsymbol{f}^j \sim \mathcal{T}(\nu, \mathbf{0}, \boldsymbol{K}(\boldsymbol{X}^j, \boldsymbol{X}^j)), \tag{2.21}$$

$$\boldsymbol{y}^j \sim g(\boldsymbol{f}^j) \tag{2.22}$$

where j indicates the set of observations and inputs and g represents the observation likelihood. In Yu et al. (2007), g was chosen to be a Gaussian distribution, i.e.:

$$\boldsymbol{y}^j \sim \mathcal{N}(\boldsymbol{f}^j, \sigma), \tag{2.23}$$

[2] There can be an overlap between the different sets, e.g. a concrete input can be in multiple input sets

while Zhang & Yeung (2010) used a Student-t observation likelihood implicitly by incorporating the Student-t errors into the TP, i.e.:

$$\boldsymbol{y}^j \sim \mathcal{T}(\nu, \boldsymbol{0}, \boldsymbol{K}(\boldsymbol{X}^j, \boldsymbol{X}^j) + \sigma \boldsymbol{I}), \tag{2.24}$$

where $\boldsymbol{I}$ refers to the identity matrix. In contrast to Yu et al. (2007), this approach has the advantage that it does not require approximate inference. Shah et al. (2014) use a similar approach as Zhang & Yeung (2010) , with the difference that there is only one set of inputs and observations, i.e. an ordinary regression problem. However, Shah et al. (2014) note that this approach leads to errors that are not independent.

The second field where TPs had a larger impact is Bayesian optimization (BO). BO is a derivative-free optimization method for computationally expensive black-box functions (Shahriari et al. (2015), Snoek et al. (2012)). The general idea is to fit a surrogate model to previous evaluations of the black-box function and use this model to choose promising next inputs for the function based on some evaluation method called acquisition function. This procedure of fitting and finding new inputs is repeated until convergence or some budget of, e.g., function evaluations, is exhausted. Originally, GPs were used as surrogate model Močkus (1975), whereas Shah et al. (2014) and Shah et al. (2013) showed that TPs can be a viable alternative. Their work has been extended to multi-objective optimization (van der Herten et al. (2016)), inventory control and optimization (Xie & Chen (2017)), and aerospace optimization (Tracey & Wolpert (2018)).

Besides of multi-task learning and Bayesian optimization, TPs have been used for, e.g. financial time series models (Ruxanda et al. (2019)), stochastic block models (Xu et al. (2011)), Bayesian quadrature (Prüher et al. (2017)), functional ANOVA (Zhang, Chen, Wang & Wu (2018)), and anomaly detection in data streams (Xu et al. (2017)). Moreover, Shah et al. (2014) found that their ordinary TP regression outperforms a GP with Student-t likelihood for a specific dataset with change-points. However, to our best knowledge, this superiority of the TP under change-points has not been further explored in the literature.

Finally, examples of TP regression models beyond Gaussian or Student-t observation likelihoods are rare. As mentioned in the previous section, Futami et al.

(2017) worked on efficient inference for TP based binary classification. Additionally, Archambeau & Bach (2011) arrived at a TP binary classification model via their work on multiple kernel learning.

2.3 Approximate Inference

2.3.1 Laplace Approximation

The Laplace approximation is a deterministic inference method that fits a Gaussian distribution to the posterior. This is done by using a second order Taylor expansion centered at the posterior mode to approximate the posterior and the log marginal likelihood (Rasmussen & Williams (2005)).

The Laplace approximation requires 2 steps:

1. Find the posterior mode (MAP) of the logarithm of the unnormalized posterior distribution $p(\theta|y)$[3]. That is, from Bayes' theorem we obtain:

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} && (2.25) \\ &\propto p(y|\theta)p(\theta). && (2.26) \end{aligned}$$

   Taking the $\log$ of this expression gives:

$$\log p(\theta|y) \propto \underbrace{\log p(y|\theta) + \log p(\theta)}_{:=\varphi(\theta)}. \tag{2.27}$$

   We denote the arguments of the maximum of this expression with respect to θ as $\hat{\theta}$

2. Substitute into the marginal loglikelihood integral and use a truncated Taylor expansion at the MAP to simplify the integral.

[3] In this section, θ refers to some arbitrary values of interest. That is, θ represents more general parameters and is not limited to the kernel parameters.

$$
\begin{aligned}
p(y) &= \int p(y|\theta)p(\theta)d\theta && (2.28)\\
&= \int \exp\left\{\log\phi(\theta)\right\}d\theta && (2.29)\\
&= \int \exp\left\{\log\phi(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2\frac{\partial^2}{\partial\theta^2}\phi(\hat{\theta})\right\}d\theta, && (2.30)\\
&= \exp\left\{\log\phi(\hat{\theta})\right\}\int \exp\left\{\frac{1}{2}(\theta-\hat{\theta})^2\frac{\partial^2}{\partial\theta^2}\phi(\hat{\theta})\right\}d\theta, && (2.31)
\end{aligned}
$$

where $\frac{\partial^2}{\partial\theta^2}$ is the second derivative with respect to θ. This last expression is a Gaussian integral and can be solved readily (see Abramowitz & Stegun (1964)).

In the case of a GGPR model, this approximation of the log marginal likelihood can be used to optimize the kernel parameters. However, this optimization alters the GP prior, which might cause the approximation to deteriorate. Consequently, we have to repeat steps 1. and 2. until convergence of the kernel parameters is reached.

Our motivation for considering the Laplace approximation in a TP setting stems from three points. Firstly, Shang & Chan (2013) show that the Laplace approximation performs as well as more sophisticated variational approximation methods for GGPR models. Additionally, in contrast to the second family of algorithms, which we will cover in the next section, Laplace approximation based solutions for the GGPR and GTPR problem do not require numerical integration methods like Monte Carlo or numerical quadrature Burden & Faires (1997). Thirdly, there have been many extensions and improvements to the classical algorithm, e.g. Cseke & Heskes (2011), Fog (2008), Rue et al. (2009), over the years. A successful adaption of the Laplace approximation would allow for future research in these directions as well.

From an historical point of view, the method was first introduced in a statistical inference setting by Tierney & Kadane (1986). However, its origin as an approach to approximate integrals dates back to Laplace himself (Laplace (1986)). In the context of GPs, Williams & Barber (1998) were the first to use the Laplace approximation for the binary and multi-classification problem.

2.3.2 Variational Inference In General

Our treatment of variational inference based methods is split into three parts. In this section, we cover the general concept of variational inference. Thereafter, we will show how these general concepts apply to the variational Gaussian approximation (Barber & Bishop (1998)) and its generalization, the variational sparse inducing point methods (Titsias (2009), Hensman, Matthews & Ghahramani (2015)).

The basic idea of variational inference is to find a tractable distribution (variational distribution) $q_\phi(\theta)$ that minimizes the discrepancy between the variational distribution and an intractable target distribution (Jordan et al. (1999), Wainwright & Jordan (2008)). That is, the inference problem is turned into an optimization problem. There are 2 important components for understanding variational inference, the discrepancy measure and the variational distribution.

1. **Discrepancy measure**: The discrepancy measure is used to quantify the difference between the variational distribution and the one we are interested in. In this sense, the discrepancy measure plays the role of the objective function for the optimization problem. Commonly, the reverse KL divergence is used to measure the discrepancy between the 2 distributions, the reverse KL divergence is given by:

$$KL(q_\phi(\theta) \parallel p(\theta)) = \int q_\phi(\theta) \log \frac{q_\phi(\theta)}{p(\theta|y)} d\theta, \tag{2.32}$$

   where the posterior distribution $p(\theta|y)$ is the intractable target distribution of interest.

   The reverse KL divergence is minimized when the variational distribution equals the distribution we want to approximate.

   Although the reverse KL divergence is arguably the most used divergence measure for variational inference, other measures, e.g. alpha divergence (Hernandez-Lobato et al. (2016)), Stein discrepancy (Liu & Wang (2016), Ranganath et al. (2016)), and, chi divergence (Dieng et al. (2017)), can be used as objective function for the optimization as well.

2. **Variational distribution**: When it comes to choosing a tractable family of distributions, a trade-off between flexibility/expressiveness of the family and ease of optimization with respect to the discrepancy measure needs to be balanced. While there might be a family of distributions with individual distributions (represented by the variational parameters ϕ) that are a good approximation to a problem at hand, there might not be an efficient optimization scheme to find these members and vice versa.

   For example, mean-field variational inference (Opper & Saad (2001)) makes the assumption that the joint variational distribution over the individual parameters factors. This has the advantage that for certain problems we can find the optimal family of distributions for the individual parameters. Additionally, for these cases, mean-field variational inference results in an iterative update scheme that provably minimizes the KL divergence between target and variational distribution (Bishop (2006)).

   In the mean-field case, the trade-off is in favor of an efficient optimization scheme with a convergence guarantee. The factorization assumption can lead to poor approximations for the target distribution, especially if the random variables are not independent. Because the mean-field approximation assumes independence of the variables, it cannot approximate any dependencies of the target distribution.

   In the next section, we will introduce variational Gaussian approximation, which does not impose a factorization assumption on the joint variational distribution, but requires numerical integration.

There is an alternative view on variational inference, one that does not directly rely on the KL divergence, but on a lower bound on the log marginal likelihood of the data, which needs to be maximized. As this approach to variational inference will be directly used for deriving the variational methods in this thesis, we give a short overview based on Zhang, Butepage, Kjellstrom & Mandt (2018).

Starting with the log marginal likelihood, we can introduce the variational distribution, as follows[4]

[4] The dependence of the variational distribution on the variational parameters ϕ has been suppressed.

$$\log p(y) = \log \int p(y|\theta)p(\theta)d\theta \tag{2.33}$$
$$= \log \int \frac{q(\theta)p(y|\theta)p(\theta)}{q(\theta)}d\theta, \tag{2.34}$$

whereas $q(\theta)$ is the variational approximation to the intractable posterior $p(\theta|y)$. Using Jensen's inequality (Gradshteyn & Ryzhik (2014)), we can lower bound the log marginal likelihood:

$$\log p(y) \geq \int q(\theta) \log \frac{p(y|\theta)p(\theta)}{q(\theta)}d\theta \tag{2.35}$$
$$= \int q(\theta) \log p(y|\theta)d\theta + \underbrace{\int q(\theta) \log \frac{p(\theta)}{q(\theta)}d\theta}_{-KL(q(\theta) \parallel p(\theta))}. \tag{2.36}$$

As the log marginal likelihood of the data is also known as the evidence, this lower bound is commonly denoted as the Evidence Lower Bound (ELBO). Interestingly, while the first integral favors a variational distribution that fits the observed data well, the second integral, the negative, reverse KL divergence, penalizes variational distributions that deviate from the prior distribution.

It can be shown (e.g. Zhang, Butepage, Kjellstrom & Mandt (2018)) that maximizing the ELBO minimizes the KL between the variational distribution and the intractable posterior:

$$\log p(y) - ELBO = \log p(y) - \int q(\theta) \log p(y|\theta)d\theta - \int q(\theta) \log \frac{p(\theta)}{q(\theta)}d\theta \tag{2.37}$$
$$= \int q(\theta) \log p(y)d\theta - \int q(\theta) \log \frac{p(y|\theta)p(\theta)}{q(\theta)}d\theta \tag{2.38}$$
$$= \int q(\theta) \log \frac{p(y)q(\theta)}{p(y|\theta)p(\theta)}d\theta \tag{2.39}$$
$$= \int q(\theta) \log \frac{q(\theta)}{p(\theta|y)}d\theta \tag{2.40}$$
$$= KL(q(\theta) \parallel p(\theta|y)), \tag{2.41}$$

whereas the following identities have been used:

$$\log p(y) = \int q(\theta) \log p(y) d\theta, \tag{2.42}$$

this is justified by $\log p(y)$ being constant with respect to θ, and

$$\frac{p(y)}{p(y|\theta)p(\theta)} = \frac{p(y)}{p(y,\theta)} = \frac{1}{p(\theta|y)}, \tag{2.43}$$

which is based on the reciprocal of Bayes' Theorem.

Equation 2.41 shows that the difference between the log marginal likelihood and the ELBO is the KL divergence between the variational distribution and the posterior. Consequently, minimizing this difference by maximizing the ELBO reduces the KL divergence between the variational distribution and the distribution of interest.

A more detailed treatment of general variational inference, which covers its different facets, is provided by Wainwright & Jordan (2008). More information on recent advances in variational inference can be found in Zhang, Butepage, Kjellstrom & Mandt (2018).

2.3.3 Variational Gaussian Approximation

The variational Gaussian approximation is a variational method that minimizes the reverse KL divergence between the target distribution and a multivariate Gaussian distribution (e.g. Barber & Bishop (1998), Nickisch & Rasmussen (2008), and Opper & Archambeau (2009)). It is especially suited for latent Gaussian models.

The reasons for focusing on the variational Gaussian approximation as the foundation for variational methods for GTPR models are manifold. On the one hand, the variational Gaussian approximation has been extensively used and reinvented for different instances of GGPR models (e.g. Barber & Bishop (1998), Nickisch & Rasmussen (2008), and Opper & Archambeau (2009)). It provides a dense covariance matrix for the Gaussian approximation of the posterior, while many other variational approximation methods make an independence assumption and work with diagonal covariance matrices (see Wainwright & Jordan (2008)). Due to a reparameterization of parameters (Opper & Archambeau (2009)), the variational Gaussian approximation requires the optimization of only $2n$ variational parameters, which is identical to methods that utilize an indepence assumption for the variational distribution. Moreover, the method has recently been adapted for variational sparse GP models (Sheth et al. (2015)), an indicator that it could also power scalable GTPR models.

Due to the importance of the method for this thesis, we briefly demonstrate the derivation of the basic bound for the variational Gaussian approximation.

Starting with the evidence lower bound, which we have derived in the previous section:

$$\log p(\boldsymbol{y}) \geq \int q(\boldsymbol{\theta}) \log p(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} - KL(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})), \tag{2.44}$$

where $q(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{V})$ is the variational approximation to the intractable posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. The variables are in bold to emphasize the multivariate setting. Importantly, both integrals can be efficiently solved under two conditions. Firstly, in case of a latent Gaussian model, that is $p(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{a}, \boldsymbol{K})$, the closed form solution for the KL divergence between two multivariate Gaussian distributions can be utilized (Kullback (1997)):

$$\begin{aligned} KL(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})) =& 0.5 \Bigg(\log \frac{\det |\boldsymbol{K}|}{\det |\boldsymbol{V}|} + \\ & Tr\left\{\boldsymbol{V}\boldsymbol{K}^{-1}\right\} + (\boldsymbol{a}-\boldsymbol{m})^T \boldsymbol{K}^{-1} (\boldsymbol{a}-\boldsymbol{m}) - n \Bigg), \end{aligned} \tag{2.45}$$

where n refers to the number of dimensions of the multivariate Gaussian distribution. Secondly, the first integral does not usually have a closed-form solution. However, under the assumption that the model likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ factorizes and each observation depends on only one component of $\boldsymbol{\theta}$, i.e. $p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_i p(\boldsymbol{y}_i|\boldsymbol{\theta}_i)$, we can express the n-dimensional integral as n 1-dimensional integrals:

$$\int q(\boldsymbol{\theta}) \log p(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} = \int q(\boldsymbol{\theta}) \log \prod_i p(\boldsymbol{y}_i|\boldsymbol{\theta}_i) d\boldsymbol{\theta} \tag{2.46}$$

$$= \sum_i \int q(\boldsymbol{\theta}) \log p(\boldsymbol{y}_i|\boldsymbol{\theta}_i) d\boldsymbol{\theta}. \tag{2.47}$$

For each of the individual integrals, we can marginalize over the components of $\boldsymbol{\theta}$ that $\boldsymbol{y}_i$ is not depending on. Additionally, it is common (e.g. Opper & Archambeau (2009), Challis & Barber (2013)) to use the relationship between the univariate, standard Gaussian distribution and any other univariate Gaussian distribution to simplify the expression even further:

$$\int q(\boldsymbol{\theta}) \log p(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} = \sum_i \int q(\boldsymbol{\theta}_i) \log p(\boldsymbol{y}_i|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \tag{2.48}$$

$$= \sum_i \int \mathcal{N}(\boldsymbol{\theta}_i; 0, 1) \log p(\boldsymbol{y}_i|\boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}}\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \tag{2.49}$$

Putting it all together, we obtain for the ELBO:

$$\log p(\boldsymbol{y}) \geq \sum_i \int \mathcal{N}(\boldsymbol{\theta}_i; 0, 1) \log p(\boldsymbol{y}_i|\boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}}\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i - 0.5 \Bigg(\log \frac{\det |\boldsymbol{K}|}{\det |\boldsymbol{V}|} + Tr\left\{\boldsymbol{V}\boldsymbol{K}^{-1}\right\} + (\boldsymbol{a} - \boldsymbol{m})^T \boldsymbol{K}^{-1} (\boldsymbol{a} - \boldsymbol{m}) - n \Bigg). \tag{2.50}$$

This bound can be efficiently computed and optimized via numerical quadrature and gradient-based optimization methods.

2.3.4 Variational Sparse Inducing Point Methods

As mentioned in the section about the GP, the primary bottleneck of GP models is their cubic run time in terms of number of observations.

One way to overcome this problem was the development of sparse approximation methods, also known as pseudo-inputs or inducing point methods (Quiñonero-Candela & Rasmussen (2005)). The basic idea of these methods is to find m inducing points, where $m << n$, at which the GP is evaluated. This augmentation of the ordinary GP model allows for a low-rank approximation of the covariance matrix of the Gaussian posterior distribution (Quiñonero-Candela & Rasmussen (2005), Bauer et al. (2016)). This usually leads to a reduction of the complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(nm^2)$, where m is the number of inducing points. What distinguishes the different sparse methods is how they select the inducing points.

Early methods focused on selecting a subset from the observed inputs values as inducing points (e.g. Williams & Seeger (2001), Herbrich et al. (2003)). For example, in practice, $\boldsymbol{K_{nn}}$ is not directly inverted, but $\sigma\boldsymbol{I} + \boldsymbol{K_{nn}}$, where σ is some small jitter to make the inversion more numerically stable. This can be used to form an approximation that can be efficiently inverted:

$$(\sigma\boldsymbol{I} + \boldsymbol{K_{nn}})^{-1} \approx (\sigma\boldsymbol{I} + \boldsymbol{K_{nm}}\boldsymbol{K_{mm}^{-1}}\boldsymbol{K_{mn}})^{-1}, \tag{2.51}$$

the approximate form can be inverted in $\mathcal{O}(nm^2)$ via Sherman–Morrison–Woodbury matrix identiy (e.g. Golub & Van Loan (2012)), that is:

$$(\sigma\boldsymbol{I} + \boldsymbol{K_{nm}}\boldsymbol{K_{mm}^{-1}}\boldsymbol{K_{mn}})^{-1} = \sigma^{-1}I - \sigma^{-2}\boldsymbol{K_{nm}}(\boldsymbol{K_{mm}} + \sigma^{-1}\boldsymbol{K_{mn}}\boldsymbol{K_{nm}})^{-1}\boldsymbol{K_{mn}}, \tag{2.52}$$

Williams & Seeger (2001) used random sampling to choose the observations for $\boldsymbol{K_{mm}}$, while later approaches of that era utilized forms of greedy search (e.g. Herbrich et al. (2003)).

In contrast to the early approaches, Snelson & Ghahramani (2005) generalized the idea of sparse approximation by relaxing the requirement of chosing the inducing points among the observed input values. They developed a principled way to learn the inducing points. Later on this method was renamed ”Fully Independent Training Conditional” (FITC) (Quiñonero-Candela & Rasmussen (2005)) and was arguably the most prominent sparse inducing point method of the last decade (Bauer et al. (2016)). However, there is evidence that the FITC method tends to overfit to the data (e.g. Naish-Guzman & Holden (2008), Matthews (2017)).

The next breakthrough in the field of sparse inducing point methods was Titsias' variational sparse inducing point method. This approach embeds the idea of inducing points in a variational framework, that is, the inducing points are chosen to minimize a KL divergence. The general idea of the variational sparse inducing point method has been used in the development of efficient algorithms for different problems, such as GP latent variable models (Titsias & Lawrence (2010)), deep GP models (Damianou & Lawrence (2012)), or streaming approximations (Bui et al.

(2017)). Additionally, the variational sparse inducing point method forms the basis of the methods that have inspired our development of the sparse inducing point methods for GTPR models, the scalable variational GP classification approach from Hensman, Matthews & Ghahramani (2015) and the sparse variational inference approach for GGPR models from Sheth et al. (2015).

Originally, our approach was based on the work of Hensman, Matthews & Ghahramani (2015), which focuses only on the binary classification problem. In this sense, our sparse variational inducing point method generalizes their approach to TPs and beyond classification problems. However, the latter has already been achieved by Sheth et al. (2015). As this approach is essential for this thesis, we briefly cover the derivation of the variational bound based on Hensman, Matthews & Ghahramani (2015).

Starting with the logarithm of the conditional likelihood of the observations with respect to the function values $\boldsymbol{u}$ at the inducing points and using Jensen's inequality (suppressing the dependence on the data and inducing points):

$$\log p(\boldsymbol{y}|\boldsymbol{u}) = \log \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})d\boldsymbol{f} \tag{2.53}$$

$$\geq \int p(\boldsymbol{f}|\boldsymbol{u}) \log p(\boldsymbol{y}|\boldsymbol{f})d\boldsymbol{f} \tag{2.54}$$

$$= \mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})} \left[\log p(\boldsymbol{y}|\boldsymbol{f})\right]. \tag{2.55}$$

In the next step, an intractable variational bound for the log marginal likelihood is formed:

$$\log p(\boldsymbol{y}) = \log \int p(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u} \tag{2.56}$$

$$= \log \int q(\boldsymbol{u})\frac{p(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{u})}{q(\boldsymbol{u})}d\boldsymbol{u} \tag{2.57}$$

$$\geq \int q(\boldsymbol{u}) \log \frac{p(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{u})}{q(\boldsymbol{u})}d\boldsymbol{u} \tag{2.58}$$

$$= \int q(\boldsymbol{u}) \log p(\boldsymbol{y}|\boldsymbol{u})d\boldsymbol{u} + \int q(\boldsymbol{u}) \log \frac{p(\boldsymbol{u})}{q(\boldsymbol{u})}d\boldsymbol{u} \tag{2.59}$$

$$= \mathbb{E}_{q(\boldsymbol{u})} \left[\log p(\boldsymbol{y}|\boldsymbol{u})\right] - KL(q(\boldsymbol{u}) \parallel p(\boldsymbol{u})). \tag{2.60}$$

The expectation in the above equation is intractable. However, we can use (2.55) to lower bound it and obtain a variational lower bound for the log marginal likelihood.

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_{q(\boldsymbol{u})}\left[\mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right]\right] - KL(q(\boldsymbol{u}) \parallel p(\boldsymbol{u})) \tag{2.61}$$

$$= \mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right] - KL(q(\boldsymbol{u}) \parallel p(\boldsymbol{u})), \tag{2.62}$$

where $q(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}$. Assuming that $p(\boldsymbol{f}|\boldsymbol{u})$ and the variational distribution $q(\boldsymbol{u})$ are Gaussian, $q(\boldsymbol{f})$ is Gaussian as well. Similarly to the variational Gaussian approximation in the previous section, this bound can be written as:

$$\log p(\boldsymbol{y}) \geq \sum_i \int \mathcal{N}(\boldsymbol{f}_i; 0, 1) \log p(\boldsymbol{y}_i | \boldsymbol{b}_i + \sqrt{\boldsymbol{B}_{ii}}\boldsymbol{f}_i) d\boldsymbol{f}_i - 0.5 \Bigg(\log \frac{\det|\boldsymbol{K_{mm}}|}{\det|\boldsymbol{B}|} + Tr\left\{\boldsymbol{B}\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\right\} + (\boldsymbol{a}-\boldsymbol{b})^T \boldsymbol{K}_{\boldsymbol{mm}}^{-1}(\boldsymbol{a}-\boldsymbol{b}) - m \Bigg), \tag{2.63}$$

where,

$$\boldsymbol{b} = \boldsymbol{K_{nm}}\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\boldsymbol{m} \tag{2.64}$$

$$\boldsymbol{B} = \boldsymbol{K_{nn}} + \boldsymbol{K_{nm}}\boldsymbol{K}_{\boldsymbol{mm}}^{-1}(\boldsymbol{V} - \boldsymbol{K_{mm}})\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\boldsymbol{K_{nm}}. \tag{2.65}$$

By amending this method to work with TP models, we hope to be able to use the TP in more versatile settings with larger datasets as well.

2.3.5 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) allows the generation of samples from arbitrary posterior distributions (Gelman et al. (2013)). This is in contrast to the methods presented so far, which can only provide approximations to the posterior distribution [5]. MCMC is used in this work to provide the gold standard to assess the quality of the posterior approximations provided by approximation methods.

The basic idea is to generate a Markov chain that has the target posterior distribution as its stationary distribution. E.g. in the Metropolis algorithm (Metropolis

[5] MCMC is also an approximation, but one that is getting arbitrarily good with more iterations

et al. (1953)), an MCMC variant, this is done by sampling from a proposal distribution and then adjust the sample to fit the target posterior distribution. In particular, a sample, θ_P, is generated from the proposal distribution $p(\theta_P|\theta_{t-1})$[6], where θ_{t-1} is the previous state of the Markov chain. Afterwards, the new sample is either rejected or accepted, based on the ratio of the posterior densities [7]. That is,

$$a = \frac{p(\theta_P|y)}{p(\theta_{t-1}|y)}, \tag{2.66}$$

and

$$\theta_t = \begin{cases} \theta_P & \text{if } a > 1 \\ \theta_P & \text{with probability } a\text{, if } a \leq 1 \\ \theta_{t-1} & \text{otherwise.} \end{cases} \tag{2.67}$$

This procedure guarantees that the resulting Markov chain has the posterior distribution as its stationary distribution. That is, we can start the algorithm at arbitrary parameter values and eventually, the method will converge to the posterior distribution [8].

The main disadvantage of the Metropolis algorithm is its random walk behaviour Gelman et al. (2013), which makes it hard to efficiently traverse high-dimensional and/or highly correlated parameter spaces. Non-conjugate GGPR problems require the exploration of a large parameter space, as inference needs to be done over each function value and these values can also be highly correlated. Consequently, MCMC methods are difficult to use for these kind of problems, which led to the development of approximate inference schemes in the first place. However, there are MCMC methods that can manage high-dimensional, correlated parameter spaces more efficiently. One example of these methods is Hamiltonian Monte Carlo (HMC) (Duane et al. (1987)). HMC uses a variable-augmentation scheme and

[6] For the Metropolis algorithm, the proposal distribution needs to be symmetric, i.e. $p(\theta_P|\theta_{t-1}) = p(\theta_{t-1}|\theta_P)$. An extension for non-symmetric proposal distributions is given by the Metropolis-Hastings algorithm (Hastings (1970))

[7] Only the unnormalized posterior densities via Bayes Theorem are necessary as the normalizing constants are cancelling each other out

[8] Discarding warm up, that is, an initial phase where the chain has not yet converged to its stationary distribution (see Gelman et al. (2013) for details)

deterministic proposals based on Hamiltonian dynamics to circumvent the random walk behaviour and to allow for efficient inference.

The first one to use HMC in a non-conjugate GGPR setting was Radford Neal Neal (1997). In Radfords work, HMC was used to conduct inference over the hyperparameters of the kernel function, while the latent function values, $\boldsymbol{f}$, were updated via Gibbs sampling (see Geman & Geman (1984) for details about Gibbs sampling). This was done, because Gibbs sampling requires the conditional distributions of the hyperparameters given all the other parameters and latent function values, which is not readily available in the GP case.

Part of the reason for the rise in popularity of HMC is the development of software packages in the last 10 years, which make it straightforward to use HMC for different problems. For this work, we will use the STAN implementation of HMC Carpenter et al. (2016).

2.3.6 Expectation Propagation

Another popular approximation method is expectation propagation (EP) Minka (2001). Similarly to the variational methods presented in section 2.3.2, EP minimizes a divergence measure (usually KL) between an intractable target posterior distribution and a tractable approximation. However, while the variational methods optimize $KL(q(\theta) \parallel p(\theta))$, EP minimizes $KL(p(\theta) \parallel q(\theta))$. As the KL divergence is not symmetric, these two expressions differ.

EP is based on an online posterior approximation method called assumed density filtering (ADF) (see Maybeck (1982)). ADF is online in the sense that it updates its posterior approximation with each new observation. To be more precise, ADF starts with the unnormalized posterior based in Bayes' Theorem:

$$p(\theta|y) \propto p(\theta) \prod_{i}^{n} p(y_i|\theta). \tag{2.68}$$

For each data point, y_i, the posterior approximation, $q_{i-1}(\theta)$, is updated to form the new approximation $q_i(\theta)$. This is done by solving:

$$q_i(\theta) = \underset{q(\theta)}{\operatorname{argmin}}\, KL(q(\theta) \parallel q_{i-1}(\theta)p(y_i|\theta)). \tag{2.69}$$

For members of the exponential family, the minimization of this KL divergence is equal to moment matching. That is, we need to set the moments of the approximating distribution to the moments of the true posterior.

The main drawback of ADF is that the approximation depends on the order of the observations, that is, different permutations of the observations, in general, give different approximations for the posterior distribution after n steps. EP overcomes this shortcoming by repeatedly iterating and updating the so called site approximations g_i until convergence Minka (2001). Where the site approximations are the approximations for the individual observation likelihood $p(y_i|\theta)$, which form together the overall approximation of the posterior, $q(\theta)$. That is, EP assumes that:

$$q(\theta) \propto p(\theta) \prod_{i}^{n} g_i(\theta). \tag{2.70}$$

In the EP algorithm, the site approximations are updated by first removing the current site approximation from the overall approximation, that is:

$$q_{-i}(\theta) \propto \frac{q(\theta)}{g_i(\theta)}, \tag{2.71}$$

where the $-i$ indicates that the ith site approximation has been removed. The distribution $q_{-i}(\theta)$ is known as the cavity distribution. Afterwards, the overall approximation can be updated by minimizing $KL(q_{-i}(\theta)p(y_i|\theta) \parallel q(\theta))$. Similarly to the ADF case, this minimization simplifies to moment matching for the exponential family. Finally, the update of the site approximation is completed by setting

$$g_i(\theta) \propto \frac{q(\theta)}{q_{-i}(\theta)}. \tag{2.72}$$

EP was generalized to work with Student-t distributions as site approximations by Futami et al. (2017).

2.4 Mathematical Tool Set

2.4.1 t-Exponential and t-Logarithm Function

The t-exponential and the t-logarithm are generalizations of the standard exponential and logarithm function. They play their most significant role in non-extensive statistical mechanics, where the t-logarithm is used in the Tsallis entropy and the t-exponential gives rise to the t-exponential family of distributions, a generalization of the exponential family of distributions (Tsallis (1988)). More recently, the two functions have been used to derive a more robust form of logistic regression (Ding & Vishwanathan (2010)) as well as variational inference methods for the t-exponential family of distributions (Ding et al. (2011) and Futami et al. (2017)).

They are defined as follows:

Definition 2.4.1. t-exponential:

$$\exp_t(x) = [1 + (1-t)x]_{+}^{\frac{1}{1-t}} \ , \tag{2.73}$$

for $t > 0$ and $[x]_+ = \max(0, x)$.

Moreover, it can be shown that:

$$\lim_{t \to 1} \exp_t(x) = \exp(x). \tag{2.74}$$

Definition 2.4.2. t-logarithm:

$$\log_t(x) = \frac{x^{1-t} - 1}{1-t}, \tag{2.75}$$

for $t > 0$ and

$$\lim_{t \to 1} \log_t(x) = \log(x). \tag{2.76}$$

The disadvantage of the t-exponential and t-logarithm compared to their standard counterparts is that they do not possess the property of transforming between products and additions, that is:

$$\log_t(xy) \neq \log_t(x) + \log_t(y), \tag{2.77}$$

and

$$\exp_t(x) \exp_t(y) \neq \exp_t(x + y). \tag{2.78}$$

However, this shortcoming led to the development of alternative algebras. Specifically relevant for this work is the q-algebra of Borges (2004), which will be briefly covered in the next section.

2.4.2 q-Algebra

The q-algebra is a deformed algebra that was introduced by Borges (2004) and is based on the t-exponential and t-logartihm. Borges' algebra contains different standard operators, e.g. q-addition and q-subtraction, and extensions of calculus, which have special properties with respect to the t-exponential and t-logarithm function. However, for the derivations in this thesis, only the q-product is relevant[9].

Definition 2.4.3. t-product:

$$x \otimes_t y = \left[x^{1-t} + y^{1-t} - 1\right]^{\frac{1}{1-t}}, \tag{2.79}$$

for $t > 0$ and

$$\lim_{t \to 1} x \otimes_t y = x * y. \tag{2.80}$$

For the t-product, we have that:

$$\log_t(x \otimes_t y) = \log_t(x) + \log_t(y), \tag{2.81}$$

[9] To have a uniform notation, we refer to the q-product as t-product.

Chapter 3

Ordinary Laplace Approximation

In this chapter, we will take a look at the application of the Laplace approximation method presented in the previous chapter on the problem of GTPR. This demonstration gives the motivation for developing new methods based on q-algebra. The structure of the derivations in this chapter are based on Rasmussen & Williams (2005).

3.1 Derivations

Laplace approximation uses a second order Taylor expansion to approximate the posterior and the log marginal likelihood (Rasmussen & Williams (2005)). The expansion is centered at the maximum of the posterior with respect to the function values, that is, by Bayes' Theorem:

$$p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})}{p(\boldsymbol{y})} \tag{3.1}$$

$$\propto p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X}). \tag{3.2}$$

Taking the $\log$ of this expression gives:

$$\log p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}) \propto \underbrace{\log p(\boldsymbol{y}|\boldsymbol{f}) + \log p(\boldsymbol{f}|\boldsymbol{X})}_{:=\varphi_t(\boldsymbol{f})}, \tag{3.3}$$

where

$$\boldsymbol{f} \sim \mathcal{TP}(\boldsymbol{f}; \nu, \mathbf{0}, \boldsymbol{K_{nn}}), \tag{3.4}$$

and $p(\boldsymbol{y}|\boldsymbol{f})$ can be any distribution whose second derivative of its log-density exists and is non-zero (Laplace (1986)).

In order to obtain the Hessian matrix for the Taylor expansion, the unnormalized posterior is differentiated twice with respect to $\boldsymbol{f}$, :

$$\varphi(\boldsymbol{f}) = \log p(\boldsymbol{y}|\boldsymbol{f}) + \log p(\boldsymbol{f}|\boldsymbol{X}) \tag{3.5}$$

$$\frac{\partial}{\partial \boldsymbol{f}}\varphi(\boldsymbol{f}) = \frac{\partial}{\partial \boldsymbol{f}} \log p(\boldsymbol{y}|\boldsymbol{f}) - \frac{\nu + n}{\nu + \boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f}} \boldsymbol{K}^{-1} \boldsymbol{f} \tag{3.6}$$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{f} \partial \boldsymbol{f}^T}\varphi(\boldsymbol{f}) = & \underbrace{\frac{\partial}{\partial \boldsymbol{f} \partial \boldsymbol{f}^T} \log p(\boldsymbol{y}|\boldsymbol{f})}_{:=-\boldsymbol{W}} \\ & + \underbrace{\frac{2(\nu + n)}{(\nu + \boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f})^2} \left[\boldsymbol{K}^{-1} \boldsymbol{f} \boldsymbol{f}^T \boldsymbol{K}^{-1} - \frac{\nu + \boldsymbol{f}^T \boldsymbol{K}^{-1} \boldsymbol{f}}{2} \boldsymbol{K}^{-1} \right]}_{:=-\boldsymbol{D}} \end{aligned} \tag{3.7}$$

$$= -(\boldsymbol{W} + \boldsymbol{D}). \tag{3.8}$$

3.2 Marginal Likelihood and Approximate Posterior

Taking the results from the previous section, the log marginal likelihood can be expressed as:

$$\log p(\boldsymbol{y}|\boldsymbol{X}) = \log \int p(\boldsymbol{y}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{X}) d\boldsymbol{f} \tag{3.9}$$

$$= \log \int \exp(\varphi(\boldsymbol{f})) d\boldsymbol{f} \tag{3.10}$$

$$\begin{aligned} \approx & \, \varphi(\hat{\boldsymbol{f}}) \\ & + \log \int \exp\left(-\frac{1}{2} (\boldsymbol{f} - \hat{\boldsymbol{f}})^T (\boldsymbol{W} + \boldsymbol{D}) (\boldsymbol{f} - \hat{\boldsymbol{f}}) \right) d\boldsymbol{f}, \end{aligned} \tag{3.11}$$

by second order Taylor expansion. Solving this Gaussian integral gives for the log marginal likelihood:

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{X}) \approx\ & \log p(\boldsymbol{y}|\hat{\boldsymbol{f}}) - \frac{1}{2}\log\det|\boldsymbol{K_{nn}}| \\ & - \frac{\nu+n}{2}\log\left(1+\frac{1}{\nu}\hat{\boldsymbol{f}}^T\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\hat{\boldsymbol{f}}\right) \\ & - \frac{1}{2}\log\det|\boldsymbol{D}+\boldsymbol{W}| - \frac{n}{2}\log\left(\frac{\nu}{2}\right) \\ & + \log\Gamma\left(\frac{\nu+n}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right).\end{aligned} \tag{3.12}$$

Based on the maximum of the posterior, the approximate posterior distribution can be specified as follows:

$$q(\boldsymbol{f}) = \mathcal{N}(\hat{\boldsymbol{f}}, (\boldsymbol{W}+\boldsymbol{D})^{-1}). \tag{3.13}$$

3.3 Shortcomings

In contrast to the Laplace approximation for GGPR (Shang & Chan (2013)), there are several drawbacks. Firstly, the equations of the approximation are more complex. Secondly, while the expressions for the GGPR can be made numerically stable (Rasmussen & Williams (2005)), the GTPR approximation does not seem to have a straighforward way to obtain this property. To be precise, the numerical stability refers to situations where we have to invert/factorize matrices that are ill-conditioned, that is, there is a substantial difference between the greatest and smallest eigenvalue (Burden & Faires (1997)). In particular, the kernel matrix $\boldsymbol{K}$ can have arbitrarily small eigenvalues (Rasmussen & Williams (2005)). In the GP case, some of the expressions involving inversion/factorization of $\boldsymbol{K}$ can be transformed into inversions/factorizations of matrices with better numerical properties (i.e. lower condition number). These transformations are not applicable for the TP case. Finally, as in the GGPR case, the posterior is approximated by a Gaussian, this is disappointing, as it feels that a Student-t distribution is more appropriate to approximate a TP-based model. In the next chapter, t-relaxation is used to generalize the Laplace approximation to tackle all of these problems.

Chapter 4

t-Laplace Approximation

In this chapter, the first method based on q-algebra for GTPR models is developed. We will first introduce the concept of t-relaxation and then we will use the relaxation to obtain the t-Laplace approximation, a method that is built on q-algebra.

4.1 t-Relaxation

The t-relaxation is our main tool to make derivations of algorithms for GTPR models. It is used to to simplify problems and obtain tractable results. The t-relaxation is defined as follows:

Definition 4.1.1. t-Relaxation A reformulation of a problem or parts of a problem in such a way that the original problem is recovered by taking the limit as t approaches 1.

The t-relaxation is important for this work, because it allows to simplify expressions involving the logarithm and the (multivariate) Student-t distribution. In the Gaussian case, the logarithm transforms Gaussian distributions into expressions that can be easily utilized for further derivations. E.g. within variational inference, we need to deal with expectation of the form of:

$$\begin{aligned}\mathbb{E}_q[\log p(x)] &= \int q(x)\left[-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2}\frac{x^2}{\sigma^2}\right]dx &(4.1)\\ &= -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2}\frac{\mathbb{E}_q[x^2]}{\sigma^2}, &(4.2)\end{aligned}$$

with,

$$p(x) \sim \mathcal{N}(0, \sigma^2). \tag{4.3}$$

If $q(x)$ is also a Gaussian distribution, then the expectation in 4.2 can readily be solved via the definition of the variance:

$$\mathbb{V}_q[x] = \mathbb{E}_q[x^2] - \mathbb{E}_q[x]^2.$$

In contrast, if $p(x)$ was Student-t distributed, we would have to solve an expectation of the form:

$$\mathbb{E}_q\left[\log\left(1 + \frac{1}{\nu}\frac{x^2}{\sigma^2}\right)\right].$$

Which does not have a closed-form solution for the case of $q(x)$ being Student-t distributed. However, if we rewrite the Student-t distribution (Ding & Vishwanathan (2010)), as:

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi\sigma^2}}\left(1 + \frac{x^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} = \left(\psi + \psi\frac{x^2}{\nu\sigma^2}\right)^{\frac{1}{1-t}}, \tag{4.4}$$

where

$$p(x) \sim \mathcal{T}(\nu, 0, \sigma^2) \tag{4.5}$$

$$\psi = \left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi\sigma^2}}\left(1 + \frac{x^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}\right)^{1-t}, \tag{4.6}$$

and t was chosen, such that:

$$\frac{1}{1-t} = -\frac{\nu+1}{2}. \tag{4.7}$$

It is straightforward to see, that the t-logarithm can simplify this expression to:

$$\log_t p(x) = \frac{\psi}{1-t} + \frac{\psi}{1-t}\frac{x^2}{\nu\sigma^2}. \tag{4.8}$$

Consequently, the t-logarithm allows to reduce expectations of the form $\mathbb{E}_q[\log_t p(x)]$, with $p(x)$ being Student-t[1], to the problem of computing $\mathbb{E}_q[x^2]$.

[1] Assuming that t is chosen appropriately

4.1.1 t-Divergence

An example for t-relaxation is the t-divergence developed by Ding & Vishwanathan (2010). While in their work, t-divergence is justified via Bregman divergence, it is also possible to derive the divergence as a t-relaxation of the Kullback-Leibler divergence:

$$KL(q(x) \parallel p(x)) = \int q(x) \log \frac{q(x)}{p(x)} \tag{4.9}$$

$$= \int q(x)(\log q(x) - \log p(x)) \tag{4.10}$$

$$\approx_t \int q(x)^t(\log_t q(x) - \log_t p(x)), \tag{4.11}$$

divide both sides by $\int q(w)^t dw$[2]

$$\frac{1}{\int q(w)^t dw} KL(q(x) \parallel p(x)) \approx_t \int \frac{q(x)^t}{\int q(w)^t dw}(\log_t q(x) - \log_t p(x)) \tag{4.12}$$

$$= \int \widetilde{q}(x)(\log_t q(x) - \log_t p(x)) \tag{4.13}$$

$$= D_t(q(x) \parallel p(x)). \tag{4.14}$$

where $\widetilde{q}(x) = \frac{q(x)^t}{\int q(w)^t dw}$ [3].

Taking the limit ($t \to 1$) on both sides and using the Dominated Convergence Theorem and the Monotone Convergence Theorem (Schilling (2017)) to move the limits under the integral sign, the Kullback-Leibler divergence is recovered.

4.1.2 t-Expectation Propagation

Also the t-expectation propagation (t-EP) algorithm developed by Futami et al. (2017) can be understood as an example of t-relaxation for expectation propagation. This can be seen by contrasting important parts of EP, as introduced in Section 2.3.6, with their t-EP counterparts.

[2] This assumes that $\int q(x)^t dx$ is finite for any finite t greater than 0. In case of the Student-t distribution this condition holds (Ding & Vishwanathan (2010)).

[3] This is a special distribution called escort distribution, more information regarding this distribution can be found in Section 5.2

First and foremost, the site approximations are Student-t distributions for the t-EP approach, while for the standard EP method, they are Gaussian distributions. The Student-t distribution is the t-relaxation of the Gaussian distribution.

Secondly, for the t-EP, the approximation of the posterior is not a product of the different site approximations, but a t-product.

Thirdly, while the cavity distribution is computed via a normal division (see equation 2.71) for the EP case, the t-EP uses the q-algebra equivalent of the division, the t-division Borges (2004). In a similar manner as the t-product, the t-division operator converges to the normal division operator as t approaches 1.

Finally, while EP minimizes $KL(p(\theta) \parallel q(\theta))$, t-EP minimizes $D_t(q(x) \parallel p(x))$, the t-divergence. As we have seen in the previous section, the t-divergence converges to the KL-divergence as t goes to 1.

4.2 Derivations

Compared to the classic Laplace approximation, the defining difference is that the resulting posterior approximation is a multivariate Student-t and not a Gaussian distribution.

The first step to obtain the t-Laplace approximation is to apply the t-relaxation on the unnormalized posterior, i.e. it is assumed that the posterior factorizes as follows:

$$p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}) \propto_t p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X}), \tag{4.15}$$

where $\propto_t$ is defined as first taking a proportional operation, i.e. left hand side is proportional to right hand side, and then t-relaxing the right hand side. Taking the $\log_t$ of this expression gives:

$$\log_t p(\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{X}) \propto_t \underbrace{\log_t p(\boldsymbol{y}|\boldsymbol{f}) + \log_t p(\boldsymbol{f}|\boldsymbol{X})}_{:=\varphi_t(\boldsymbol{f})}. \tag{4.16}$$

As in the case of the ordinary Laplace approximation, the gradient and Hessian of the (t-relaxed) unnormalized posterior is required:

$$\varphi_t(\boldsymbol{f}) = \log_t p(\boldsymbol{y}|\boldsymbol{f}) + \frac{\Psi_p}{1-t} + \frac{\Psi_p}{1-t}\boldsymbol{f}^T(\nu\boldsymbol{K_{nn}})^{-1}\boldsymbol{f} - \frac{1}{1-t} \tag{4.17}$$

$$\frac{\partial\varphi_t(\boldsymbol{f})}{\partial\boldsymbol{f}} = \frac{\partial\log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial\boldsymbol{f}} + \frac{2\Psi_p}{1-t}(\nu\boldsymbol{K_{nn}})^{-1}\boldsymbol{f} \tag{4.18}$$

$$\frac{\partial\varphi_t(\boldsymbol{f})}{\partial\boldsymbol{f}\partial\boldsymbol{f}^T} = \frac{\partial\log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial\boldsymbol{f}\partial\boldsymbol{f}^T} + \frac{2\Psi_p}{1-t}(\nu\boldsymbol{K_{nn}})^{-1} \tag{4.19}$$

$$= \frac{\partial\log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial\boldsymbol{f}\partial\boldsymbol{f}^T} - \frac{2\Psi_p}{t-1}(\nu\boldsymbol{K_{nn}})^{-1} \tag{4.20}$$

$$= -\boldsymbol{W} - \boldsymbol{D}^{-1}, \tag{4.21}$$

where,

$$\boldsymbol{W} = -\frac{\partial\log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial\boldsymbol{f}\partial\boldsymbol{f}^T} \tag{4.22}$$

$$\boldsymbol{D} = \frac{t-1}{2\Psi_p}(\nu\boldsymbol{K_{nn}}) \tag{4.23}$$

$$\Psi_p = \left(\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)\det|\boldsymbol{K_{nn}}|^{\frac{1}{2}}}\right)^{1-t}. \tag{4.24}$$

Noteworthily, the form of the gradient and the Hessian are comparable to the ones that are obtained in the GP case (Rasmussen & Williams (2005)).

4.3 Marginal Likelihood and Posterior Approximation

For the approximation of the marginal likelihood, the t-relaxation is applied to the marginal likelihood integral, that is [4]:

$$p(\boldsymbol{y}|\boldsymbol{X}) \approx_t \int p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X}) d\boldsymbol{f}. \tag{4.31}$$

In contrast to the Laplace Approximation, the representation of the integrand is not changed by using a combination of the ordinary logarithm and exponential function, but by their t-counterparts:

$$= \int \exp_t \left(\varphi_t(\boldsymbol{f})\right) d\boldsymbol{f}, \tag{4.32}$$

Taylor expand $\varphi_t(\boldsymbol{f})$ at the posterior mode $\hat{\boldsymbol{f}}$:

$$p(\boldsymbol{y}|\boldsymbol{X}) \approx \int \exp_t \left(\varphi_t(\hat{\boldsymbol{f}}) + \frac{1}{2}\left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)^T \boldsymbol{H} \left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)\right) d\boldsymbol{f}, \tag{4.33}$$

where

$$\boldsymbol{H} = -\boldsymbol{W} - \boldsymbol{D}^{-1}. \tag{4.34}$$

[4] For t close to 1, we can invoke the Dominated Convergence Theorem (DCT, Schilling (2017)) to derive the t-relaxation for the integral by changing the order of integral and limit, that is:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X}) &= \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f} && (4.25)\\ &= \int \lim_{t\to 1} p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f} && (4.26)\\ &= \lim_{t\to 1} \int p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f} && (4.27)\\ &\approx_t \int p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f}. && (4.28) \end{aligned}$$

DCT is applicable here, because, as t goes to 1, $p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})$ approaches $p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})$, that is, for t close to 1, $p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})$ can be bounded for all $\boldsymbol{f}$ by $\epsilon * p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})$ for some finite ϵ greater than 1. Assuming that $\int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f} < \infty$, the DCT condition is fulfilled by:

$$\begin{aligned} \int p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f} &< \epsilon \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f} && (4.29)\\ &< \infty && (4.30) \end{aligned}$$

Utilizing the definition of $\exp_t$ (equation 2.73), the following is obtained:

$$p(\boldsymbol{y}|\boldsymbol{X}) \approx \int \left[1 + (1-t)\varphi_t(\hat{\boldsymbol{f}}) + \frac{1-t}{2}\left(\boldsymbol{f}-\hat{\boldsymbol{f}}\right)^T \boldsymbol{H}\left(\boldsymbol{f}-\hat{\boldsymbol{f}}\right)\right]^{\frac{1}{1-t}} d\boldsymbol{f} \tag{4.35}$$

$$= \underbrace{\left[1 + (1-t)\varphi_t(\hat{\boldsymbol{f}})\right]^{\frac{1}{1-t}}}_{:=C} \times \int \left[1 + \underbrace{\frac{1-t}{2\left[1 + (1-t)\varphi_t(\hat{\boldsymbol{f}})\right]}}_{:=-\tau^{-1}} \left(\boldsymbol{f}-\hat{\boldsymbol{f}}\right)^T \underbrace{\boldsymbol{H}}_{:=-A^{-1}} \left(\boldsymbol{f}-\hat{\boldsymbol{f}}\right)\right]^{\frac{1}{1-t}} d\boldsymbol{f} \tag{4.36}$$

$$= C^{\frac{1}{1-t}} \int \left[1 + \left(\boldsymbol{f}-\hat{\boldsymbol{f}}\right)^T \left(\nu\frac{\tau}{\nu}\boldsymbol{A}\right)^{-1} \left(\boldsymbol{f}-\hat{\boldsymbol{f}}\right)\right]^{\frac{1}{1-t}} d\boldsymbol{f}. \tag{4.37}$$

The expression under the integral sign is equal to the unnormalized density of a multivariate Student-t distribution. From this it follows that the solution to the integral is the reciprocal of the normalizing constant of the corresponding Student-t distribution.

$$p(\boldsymbol{y}|\boldsymbol{X}) \approx C^{\frac{1}{1-t}} \left[\frac{(\pi\tau)^{\frac{n}{2}}\Gamma(\frac{\nu}{2}) \det|\boldsymbol{A}|^{\frac{1}{2}}}{\Gamma\left(\frac{\nu+n}{2}\right)}\right]^{\frac{1}{1-t}} = q(\boldsymbol{y}). \tag{4.38}$$

Based on the posterior mode $\hat{\boldsymbol{f}}$ and the normalizing constant, the posterior approximation can be specified as a Student-t distribution[5]

$$q(\boldsymbol{f}) \sim \mathcal{MVT}\left(\nu, \hat{\boldsymbol{f}}, \frac{\tau}{\nu}(\boldsymbol{D}^{-1} + \boldsymbol{W})^{-1}\right).$$

It is important to note, that there is a close relationship between this posterior approximation and the one obtained in 3.13 for the ordinary Laplace approximation. As the degrees of freedom ν go to infinity the distribution in 3.13 is almost

[5] This result is obtained by using Bayes' Law in conjunction with Laplace approximation, i.e. $p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})} \approx \frac{\left[1+(\boldsymbol{f}-\hat{\boldsymbol{f}})^T(\nu\frac{\tau}{\nu}\boldsymbol{A})^{-1}(\boldsymbol{f}-\hat{\boldsymbol{f}})\right]^{\frac{1}{1-t}}}{q(\boldsymbol{y})}$, which is equivalent to the given Student-t distribution (note that $\boldsymbol{A} = (\boldsymbol{D}^{-1} + \boldsymbol{W})^{-1}$)

recovered, except for the differences in the $\boldsymbol{D}^{-1}$ variable. Interestingly, in the limit, the t-Laplace approximation converges to a Gaussian with similar parameters[6] to the result of the Laplace approximation for the Gaussian process binary classification task (Rasmussen & Williams (2005)). To obtain these convergence results, it is helpful to see that:

$$t = \frac{2}{\nu + n} + 1, \tag{4.39}$$

which allows to express τ in terms of the degrees of freedom:

$$\begin{aligned} \tau &= \frac{2\left[1 - \frac{2}{\nu+n}\varphi_t(\hat{\boldsymbol{f}})\right]}{\frac{2}{\nu+n}} && (4.40) \\ &= \nu + n - 2\varphi_t(\hat{\boldsymbol{f}}). && (4.41) \end{aligned}$$

Consequently, the posterior approximation can be rewritten as:

$$q(\boldsymbol{f}) \sim \mathcal{MVT}\left(\nu, \hat{\boldsymbol{f}}, \left(1 + \frac{n}{\nu} - \frac{2\varphi_t(\hat{\boldsymbol{f}})}{\nu}\right)\left(\boldsymbol{D}^{-1} + \boldsymbol{W}\right)^{-1}\right).$$

As ν goes to infinity, this converges to a multivariate normal distribution:

$$\lim_{\nu \to \infty} q(\boldsymbol{f}) \sim \mathcal{MVN}\left(\hat{\boldsymbol{f}}, \left(\boldsymbol{D}^{-1} + \boldsymbol{W}\right)^{-1}\right).$$

4.4 Predictions

The expectation $\mathbb{E}\left[y_* | \boldsymbol{X}, \boldsymbol{y}, x_*\right]$ for newly observed x_* is generally intractable for the TP case [7]. That is, solving the integral requires Monte Carlo or numerical quadrature approaches. However, if the approximate posterior mean for $\boldsymbol{f}_*$ suffices, then there is a closed form for this expectation under the t-Laplace approximation.

[6] Again, the difference is in the variable $\boldsymbol{D}^{-1}$ which is the inverse kernel matrix $\boldsymbol{K_{nn}}$ in the GP case and the scaled version of $\boldsymbol{K}_{\boldsymbol{nn}}^{-1}$ for the t-Laplace approximation.

[7] And for most GP cases.

$$\begin{aligned}
\mathbb{E}_q\left[\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},x_*\right] &= \int \mathbb{E}\left[\boldsymbol{f}_*|\boldsymbol{f},\boldsymbol{X},x_*\right] q(\boldsymbol{f})d\boldsymbol{f} && (4.42)\\
&= \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\,\mathbb{E}_q\left[\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}\right] && (4.43)\\
&= \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\hat{\boldsymbol{f}} && (4.44)\\
&= \frac{(t-1)\nu}{2\Psi_p}\boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\frac{\partial \log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial \boldsymbol{f}}. && (4.45)
\end{aligned}$$

Where the last equality follows from the fact that the gradient of $\varphi(\boldsymbol{f})$ with respect to $\boldsymbol{f}$ is equal to 0 at the posterior mode $\hat{\boldsymbol{f}}$. That is,

$$\hat{\boldsymbol{f}} = \frac{t-1}{2\Psi_p}(\nu\boldsymbol{K}_{\boldsymbol{nn}})\frac{\partial \log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial \boldsymbol{f}}. \quad (4.46)$$

Moreover, the approximate variance of the posterior predictive can also be computed. Using the law of total variance (Wasserman (2010)), we obtain:

$$\begin{aligned}
\mathbb{V}_q\left[\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},x_*\right] &= \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X})}\left[\mathbb{V}\left[\boldsymbol{f}_*|\boldsymbol{f},x_*\right]\right] + \mathbb{V}_{q(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X})}\left[\mathbb{E}\left[\boldsymbol{f}_*|\boldsymbol{f},x_*\right]\right] && (4.47)\\
&= \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X})}\left[\frac{\nu+n}{\nu+n-2}\frac{\nu+\boldsymbol{f}^T\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{f}}{\nu+n}\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}}\right] + \\
&\quad \mathbb{V}_{q(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X})}\left[\boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{f}\right] && (4.48)\\
&= \frac{\nu + Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\,\mathbb{E}_q\left[\boldsymbol{f}\boldsymbol{f}^T\right]\right\}}{\nu+n-2}\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}} + \\
&\quad \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\,\mathbb{V}_{q(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X})}\left[\boldsymbol{f}\right]\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*} && (4.49)\\
&= \frac{\nu + Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\left[\frac{\nu}{\nu-2}\frac{\tau}{\nu}(\boldsymbol{D}^{-1}+\boldsymbol{W})^{-1} + \hat{\boldsymbol{f}}\hat{\boldsymbol{f}}^T\right]\right\}}{\nu+n-2}\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}} + \\
&\quad \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\left(\frac{\nu}{\nu-2}\frac{\tau}{\nu}(\boldsymbol{D}^{-1}+\boldsymbol{W})^{-1}\right)\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*} && (4.50)\\
&= \underbrace{\frac{\nu + \frac{\tau}{\nu-2}Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1}(\boldsymbol{D}^{-1}+\boldsymbol{W})^{-1}\right\} + \hat{\boldsymbol{f}}^T\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\hat{\boldsymbol{f}}}{\nu+n-2}}_{:=\phi}\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}} + \\
&\quad \frac{\tau}{\nu-2}\boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}(\boldsymbol{D}^{-1}+\boldsymbol{W})^{-1}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*}, && (4.51)
\end{aligned}$$

expanding $\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}}$ as

$$\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}} = \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}_*} - \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*},$$

gives

$$\begin{aligned}\mathbb{V}_q\left[\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},x_*\right] &= \phi\left(\boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}_*}-\boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*}\right)+ \\ &\quad \frac{\tau}{\nu-2}\boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}(\boldsymbol{D}^{-1}+\boldsymbol{W})^{-1}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*} \end{aligned} \tag{4.52}$$

$$\begin{aligned} &= \phi\boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}_*}- \\ &\quad \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\left(\phi\boldsymbol{K}_{\boldsymbol{nn}}-\frac{\tau}{\nu-2}(\boldsymbol{D}^{-1}+\boldsymbol{W})^{-1}\right)\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*}. \end{aligned} \tag{4.53}$$

Interestingly, while the predictive posterior mean for the t-Laplace approximation is identical to the normal Laplace approximation, the posterior predictive variance differs substantially.

As a major drawback to the GP case, it is not possible to derive a closed-form solution for the posterior predictive distribution, that is, the integral:

$$q(\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},x_*)=\int p(\boldsymbol{f}_*|\boldsymbol{f},\boldsymbol{X},x_*)q(\boldsymbol{f})d\boldsymbol{f},$$

is intractable.

4.5 Newton Method for Posterior Mode

As the variable $\boldsymbol{f}$ can be high-dimensional, finding the posterior mode for the t-Laplace approximation can be challenging. Fortunately, the equations presented so far can be translated into an efficient Newton optimization routine in a similar way as done in the GP case (Rasmussen & Williams (2005)).

The Newton update step is given by:

$$\boldsymbol{f}^{n+1}=\boldsymbol{f}^{n}-\left(\frac{\partial\varphi_t(\boldsymbol{f}^n)}{\partial\boldsymbol{f}\partial\boldsymbol{f}^T}\right)^{-1}\frac{\partial\varphi_t(\boldsymbol{f}^n)}{\partial\boldsymbol{f}} \tag{4.54}$$

$$=\boldsymbol{f}^n+\left(\boldsymbol{W}+\boldsymbol{D}^{-1}\right)^{-1}\left(\frac{\partial\log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial\boldsymbol{f}}-\boldsymbol{D}^{-1}\boldsymbol{f}^n\right) \tag{4.55}$$

$$=\left(\boldsymbol{W}+\boldsymbol{D}^{-1}\right)^{-1}\left(\frac{\partial\log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial\boldsymbol{f}}+\boldsymbol{W}\boldsymbol{f}^n\right). \tag{4.56}$$

Analogously to the GP case, numerically stable variants of this update equation can also be derived for the TP-based models (Rasmussen & Williams (2005)).

4.6 Optimization of Kernel Parameters

In addition to optimizing the function values for the t-Laplace approximation, we also need to tune the hyperparameters of the kernel $\boldsymbol{K_{nn}}$. For this, we need the derivatives of the approximate log marginal likelihood with respect to the hyperparameters. We start by stating some important identities from Petersen & Pedersen (2012):

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \det |\boldsymbol{K_{nn}}| = \det |\boldsymbol{K_{nn}}| Tr \left\{ \boldsymbol{K_{nn}^{-1}} \frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K_{nn}} \right\} \tag{4.57}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \log \det |\boldsymbol{K_{nn}}| = Tr \left\{ \boldsymbol{K_{nn}^{-1}} \frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K_{nn}} \right\} \tag{4.58}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K_{nn}^{-1}} = -\boldsymbol{K_{nn}^{-1}} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K_{nn}} \right) \boldsymbol{K_{nn}^{-1}}. \tag{4.59}$$

Additionally, the following identities will also be used:

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{D}_i^{-1} = \frac{\partial}{\partial \boldsymbol{\theta}_i} \frac{2\Psi_p}{t-1} (\nu \boldsymbol{K_{nn}})^{-1} \tag{4.60}$$

$$= \frac{2}{t-1} (\nu \boldsymbol{K_{nn}})^{-1} \frac{\partial}{\partial \boldsymbol{\theta}_i} \Psi_p + \frac{2\Psi_p}{(t-1)\nu} \frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K_{nn}^{-1}} \tag{4.61}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \tau = \frac{2 \frac{\partial}{\partial \boldsymbol{\theta}_i} C}{t-1} \tag{4.62}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} C = \frac{\partial}{\partial \boldsymbol{\theta}_i} \left[1 + (1-t) \varphi_t(\hat{\boldsymbol{f}}) \right] \tag{4.63}$$

$$= \frac{1}{1-t} \frac{\partial}{\partial \boldsymbol{\theta}_i} \Psi_p + \frac{\hat{\boldsymbol{f}}^T (\nu \boldsymbol{K_{nn}})^{-1} \hat{\boldsymbol{f}}}{1-t} \frac{\partial}{\partial \boldsymbol{\theta}_i} \Psi_p + \frac{\Psi_p}{\nu} \hat{\boldsymbol{f}}^T \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K_{nn}^{-1}} \right) \hat{\boldsymbol{f}}, \tag{4.64}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_i}\Psi_p = \frac{\partial}{\partial \boldsymbol{\theta}_i}\left(\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)\det|\boldsymbol{K_{nn}}|^{\frac{1}{2}}}\right)^{1-t} \tag{4.65}$$

$$= (1-t)\left(\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)\det|\boldsymbol{K_{nn}}|^{\frac{1}{2}}}\right)^{-t}\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)}\frac{\partial}{\partial \boldsymbol{\theta}_i}\det|\boldsymbol{K_{nn}}|^{-\frac{1}{2}} \tag{4.66}$$

$$\begin{aligned}
&= \frac{1-t}{2}\left(\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)\det|\boldsymbol{K_{nn}}|^{\frac{1}{2}}}\right)^{-t}\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right)}\times \\
&\quad \det|\boldsymbol{K_{nn}}|^{-\frac{3}{2}}\frac{\partial}{\partial \boldsymbol{\theta}_i}\det|\boldsymbol{K_{nn}}|.
\end{aligned} \tag{4.67}$$

The gradient of the approximate log marginal likelihood with respect to the hyperparameters of the kernel is given by:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}_i}\log q(\boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{\theta}_i}\frac{1}{1-t}\Bigg[&\log C + \frac{n}{2}\log\pi + \frac{n}{2}\log\tau + \\
&\log\Gamma\left(\frac{\nu}{2}\right) + \frac{1}{2}\log\det|\boldsymbol{A}| - \log\Gamma\left(\frac{\nu+n}{2}\right)\Bigg]
\end{aligned} \tag{4.68}$$

$$\begin{aligned}
= \frac{1}{1-t}\Bigg[&\frac{1}{C}\frac{\partial}{\partial \boldsymbol{\theta}_i}C + \frac{n}{2\tau}\frac{\partial}{\partial \boldsymbol{\theta}_i}\tau - \\
&\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\theta}_i}\log\det\left|(\boldsymbol{W}+\boldsymbol{D}^{-1})\right|\Bigg]
\end{aligned} \tag{4.69}$$

$$\begin{aligned}
= \frac{1}{1-t}\Bigg[&\frac{1}{C}\frac{\partial}{\partial \boldsymbol{\theta}_i}C + \frac{n}{2\tau}\frac{\partial}{\partial \boldsymbol{\theta}_i}\tau - \\
&\frac{1}{2}Tr\left\{(\boldsymbol{W}+\boldsymbol{D}^{-1})^{-1}\left(\frac{\partial}{\partial \boldsymbol{\theta}_i}\boldsymbol{D}^{-1}\right)\right\}\Bigg].
\end{aligned} \tag{4.70}$$

While this gradient is complex, it does not capture the full relationship between the approximate log marginal likelihood and the kernel parameters. To complete the picture, we need to take into consideration the changes in the optimal function values $\hat{\boldsymbol{f}}$ with respect to the hyperparameters, i.e. we are interested in

$$\frac{\partial \log q(\boldsymbol{y})}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}_i},$$

for $\boldsymbol{f} = \hat{\boldsymbol{f}}$.

Firstly, it is important to notice that:

$$\frac{\partial}{\partial \boldsymbol{f}}C = \frac{\partial}{\partial \boldsymbol{f}}\tau = 0. \tag{4.71}$$

This is due to the fact that we are at the posterior maximum, therefore $\frac{\partial \varphi_t(\boldsymbol{f})}{\partial \boldsymbol{f}} = 0$ at $\hat{\boldsymbol{f}}$.

Secondly, $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}_i}$ can be obtained from equation 4.46:

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}_i} = \frac{\partial}{\partial \boldsymbol{\theta}_i} \frac{t-1}{2\Psi_p}(\nu \boldsymbol{K_{nn}}) \frac{\partial \log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial \boldsymbol{f}} \tag{4.72}$$

$$\begin{aligned} &= -\frac{t-1}{2\Psi_p^2}\left(\frac{\partial}{\partial \boldsymbol{\theta}_i}\Psi_p\right)(\nu \boldsymbol{K_{nn}})\frac{\partial \log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial \boldsymbol{f}} + \\ &\quad \frac{t-1}{2\Psi_p}(\nu \frac{\partial}{\partial \boldsymbol{\theta}_i}\boldsymbol{K_{nn}})\frac{\partial \log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial \boldsymbol{f}} + \\ &\quad \frac{t-1}{2\Psi_p}(\nu \boldsymbol{K_{nn}})\underbrace{\frac{\partial \log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial \boldsymbol{f}\boldsymbol{f}^T}}_{:=-\boldsymbol{W}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}_i}. \end{aligned} \tag{4.73}$$

This can be further simplified by solving for $\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}_i}$:

$$\begin{aligned} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}_i} &= \left(\boldsymbol{I} + \frac{t-1}{2\Psi_p}(\nu \boldsymbol{K_{nn}})\boldsymbol{W}\right)^{-1} \times \\ &\left[-\frac{t-1}{2\Psi_p^2}\left(\frac{\partial}{\partial \boldsymbol{\theta}_i}\Psi_p\right)(\nu \boldsymbol{K_{nn}})\frac{\partial \log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial \boldsymbol{f}} + \frac{t-1}{2\Psi_p}(\nu \frac{\partial}{\partial \boldsymbol{\theta}_i}\boldsymbol{K_{nn}})\frac{\partial \log_t p(\boldsymbol{y}|\boldsymbol{f})}{\partial \boldsymbol{f}}\right]. \end{aligned} \tag{4.74}$$

$$\tag{4.75}$$

Lastly, we need to find the derivative of the approximate log marginal likelihood with respect to the function values evaluated at the posterior mode:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{f}_i}\log q(\boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{f}_i}\frac{1}{1-t}\Bigg[&\log C + \frac{n}{2}\log\pi + \frac{n}{2}\log\tau + \\ &\log\Gamma\left(\frac{\nu}{2}\right) + \frac{1}{2}\log\det|\boldsymbol{A}| - \log\Gamma\left(\frac{\nu+n}{2}\right)\Bigg] \end{aligned} \tag{4.76}$$

$$\begin{aligned} &= \frac{1}{1-t}\Bigg[\frac{1}{C}\frac{\partial}{\partial \boldsymbol{f}_i}C + \frac{n}{2\tau}\frac{\partial}{\partial \boldsymbol{f}_i}\tau - \\ &\quad \frac{1}{2}\frac{\partial}{\partial \boldsymbol{f}_i}\log\det\left|(\boldsymbol{W}+\boldsymbol{D}^{-1})\right|\Bigg] \end{aligned} \tag{4.77}$$

$$= -\frac{1}{2(1-t)}Tr\left\{(\boldsymbol{W}+\boldsymbol{D}^{-1})^{-1}\left(\frac{\partial}{\partial \boldsymbol{f}_i}\boldsymbol{W}\right)\right\}, \tag{4.78}$$

where

$$\frac{\partial}{\partial \boldsymbol{f}_i}\boldsymbol{W} = -\frac{\partial^3}{\partial^3 \boldsymbol{f}}\log_t p(\boldsymbol{y}|\boldsymbol{f}). \tag{4.79}$$

Putting all the individual parts together, we obtain the total derivative with respect to the kernel hyperparameters:

$$\frac{d}{d\boldsymbol{\theta}_i} \log q(\boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{\theta}_i} \log q(\boldsymbol{y}) + \left(\left. \frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{f}_i} \right|_{\boldsymbol{f}_i = \hat{\boldsymbol{f}}_{1..n}} \right)^T \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}_i}, \tag{4.80}$$

where we use

$$\left. \frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{f}_i} \right|_{\boldsymbol{f}_i = \hat{\boldsymbol{f}}_{1..n}},$$

to indicate that $\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{f}_i}$ needs to be evaluated for each component of the $\hat{\boldsymbol{f}}$ vector to form the gradient vector $\frac{\partial q(\boldsymbol{y})}{\partial \boldsymbol{f}}$.

Chapter 5

Variational Student-t Approximations

In this chapter, two Student-t versions of the variational Gaussian approximation (e.g.Opper & Archambeau (2009)) are developed.

5.1 Variational Bound on Marginal Likelihood

As a first step, the integral for the log marginal likelhood needs to be t-relaxed and the variational distribution $q(\boldsymbol{f})$ is introduced, that is:

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{X}) &= \log \int p(\boldsymbol{y}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{X}) d\boldsymbol{f} && (5.1)\\ &\approx_t \log_t \int p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X}) d\boldsymbol{f} && (5.2)\\ &= \log_t \int q(\boldsymbol{f})^{\frac{1}{t}} \frac{p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})}{q(\boldsymbol{f})^{\frac{1}{t}}} d\boldsymbol{f}. && (5.3)\end{aligned}$$

The $q(\boldsymbol{f})^{\frac{1}{t}}$ has been used for mathematical convenience and additionally because it leads to a less complex lower bound to the t-log marginal likelihood.

As a second step, the concavity of the t-log for densities (see Appendix B.1) can be used to derive a variational lower bound for the t-log marginal likelihood:

$$\log_t p(\boldsymbol{y}) \geq \int q(\boldsymbol{f})^{\frac{1}{t}} \log_t \frac{p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})}{q(\boldsymbol{f})^{\frac{1}{t}}} d\boldsymbol{f}. \tag{5.4}$$

From Borges (2004), we know that:

$$\log_t\left(\frac{x}{y}\right) = y^{t-1}\left[\log_t(x) - \log_t(y)\right],$$

which gives for the bound on the t-log marginal likelihood:

$$\begin{aligned}\log_t p(y|X) \geq &\int q(\boldsymbol{f})^{\frac{1}{t}} q(\boldsymbol{f})^{\frac{1}{t}(t-1)} \\ &\times \left[\log_t p(y|\boldsymbol{f}) + \log_t p(\boldsymbol{f}|X) - \log_t q(\boldsymbol{f})^{\frac{1}{t}}\right] d\boldsymbol{f}\end{aligned} \tag{5.5}$$

$$\begin{aligned}= &\underbrace{\int q(\boldsymbol{f}) \log_t p(y|\boldsymbol{f}) d\boldsymbol{f}}_{:=A} + \underbrace{\int q(\boldsymbol{f}) \log_t p(\boldsymbol{f}|X) d\boldsymbol{f}}_{:=B} \\ &- \underbrace{\int q(\boldsymbol{f}) \log_t q(\boldsymbol{f})^{\frac{1}{t}} d\boldsymbol{f}}_{:=C}.\end{aligned} \tag{5.6}$$

In general, A requires numerical integration, whereas B and C have closed form solutions for the TP, as long as the degrees of freedom are the same for the variational Student-t distribution and the prior distribution $p(\boldsymbol{f}|\boldsymbol{X})$, that is:

$$p(\boldsymbol{f}|\boldsymbol{X}) \sim \mathcal{MVT}(\nu, \mathbf{0}, \boldsymbol{K}) \tag{5.7}$$

$$q(\boldsymbol{f}) \sim \mathcal{MVT}(\nu, \boldsymbol{m}, \boldsymbol{V}). \tag{5.8}$$

Utilizing the definition of the t-logarithm and using the properties of the trace (Petersen & Pedersen (2012)), we can simplify B to obtain a closed-form solution as follows:

$$B = \frac{1}{1-t}\int q(\boldsymbol{f})\left[\Psi_p + \Psi_p \boldsymbol{f}^T(\nu\boldsymbol{K})^{-1}\boldsymbol{f}\right] d\boldsymbol{f} - \frac{1}{1-t} \tag{5.9}$$

$$= \frac{\Psi_p}{1-t} + \frac{\Psi_p}{1-t} Tr\left\{(\nu\boldsymbol{K})^{-1}\mathbb{E}_q\left[\boldsymbol{f}\boldsymbol{f}^T\right]\right\} - \frac{1}{1-t} \tag{5.10}$$

$$\begin{aligned}= &\frac{\Psi_p}{1-t} + \frac{\Psi_p \nu}{(1-t)(\nu-2)} Tr\left\{(\nu\boldsymbol{K})^{-1}\boldsymbol{V}\right\} + \\ &\frac{\Psi_p}{1-t}\boldsymbol{m}^T(\nu\boldsymbol{K})^{-1}\boldsymbol{m} - \frac{1}{1-t}\end{aligned} \tag{5.11}$$

$$= \frac{\Psi_p}{(1-t)}\left(1 + \frac{1}{\nu-2} Tr\left\{\boldsymbol{K}^{-1}\boldsymbol{V}\right\} + \frac{1}{\nu}\boldsymbol{m}^T\boldsymbol{K}^{-1}\boldsymbol{m}\right) - \frac{1}{1-t}. \tag{5.12}$$

C can be simplified in a similar manner, that is:

$$C = \frac{1}{1-t} \int q(\boldsymbol{f}) \left[q(\boldsymbol{f})^{\frac{1-t}{t}} - 1 \right] d\boldsymbol{f} \tag{5.13}$$

$$= \frac{1}{1-t} \int q(\boldsymbol{f}) \left[\left(\Psi_q + \Psi_q (\boldsymbol{f} - \boldsymbol{m})^T (\nu \boldsymbol{V})^{-1} (\boldsymbol{f} - \boldsymbol{m}) \right)^{\frac{1}{t}} - 1 \right] d\boldsymbol{f} \tag{5.14}$$

$$= \frac{\Psi_q^{\frac{1}{t(1-t)}}}{1-t} \int \left[1 + (\boldsymbol{f} - \boldsymbol{m})^T (\nu \boldsymbol{V})^{-1} (\boldsymbol{f} - \boldsymbol{m}) \right]^{\frac{1}{t(1-t)}} d\boldsymbol{f} - \frac{1}{1-t}, \tag{5.15}$$

with

$$\frac{1}{t(1-t)} = -\frac{\rho + n}{2} \Rightarrow \rho = \frac{2}{t(t-1)} - n, \tag{5.16}$$

we can simplify this expression further:

$$C = \frac{\Psi_q^{\frac{1}{t(1-t)}}}{1-t} \int \left[1 + (\boldsymbol{f} - \boldsymbol{m})^T \left(\rho \frac{\nu}{\rho} \boldsymbol{V} \right)^{-1} (\boldsymbol{f} - \boldsymbol{m}) \right]^{-\frac{\rho+n}{2}} d\boldsymbol{f} - \frac{1}{1-t}. \tag{5.17}$$

The integrand is the unnormalized density of a multivariate Student-t distribution, with ρ degrees of freedom:

$$\mathcal{MVT}(\rho, \boldsymbol{m}, \frac{\nu}{\rho} \boldsymbol{V}).$$

Consequently, the result of the integral is the reciprocal of the normalizing constant:

$$= \frac{\Psi_q^{\frac{1}{t(1-t)}}}{1-t} \frac{\Gamma(\frac{\rho}{2}) \rho^{\frac{n}{2}} \pi^{\frac{n}{2}} \det \left| \frac{\nu}{\rho} \boldsymbol{V} \right|^{\frac{1}{2}}}{\Gamma(\frac{\rho+n}{2})} - \frac{1}{1-t} \tag{5.18}$$

$$= \log_t \left(D^{\frac{1}{1-t}} \det |\boldsymbol{V}|^{-\frac{1}{2t}} \right), \tag{5.19}$$

where

$$D = \frac{\Gamma\left(\frac{\nu+n}{2}\right)^{\frac{1}{t}} \Gamma\left(\frac{\rho}{2}\right) (\nu \pi)^{\frac{n(t-1)}{2t}}}{\Gamma\left(\frac{\nu}{2}\right)^{\frac{1}{t}} \Gamma\left(\frac{\rho+n}{2}\right)}.$$

Concerning the integration of A, due to the high-dimensional[1] nature of the problem, it is not practically feasible to solve it in its current form. In the GP case,

[1] For numerical integration, more than 5 dimensions are already critical (e.g. Burden & Faires (1997)).

the properties of the logarithm are used to simplify the n-dimensional integration problem into n 1-dimensional problems. For the TP case, this cannot be used, as the t-logarithm does not decompose a product into a simple sum over the t-logarithms of the product's elements. For this reason, the t-relaxation is applied to the observation model in order to transform the ordinary product into a t-product:

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_i \otimes_t p(\boldsymbol{y}_i|\boldsymbol{f}_i). \tag{5.20}$$

Consequently, the rules of the t-logarithm can be used to obtain:

$$\begin{aligned} A &= \int q(\boldsymbol{f}) \log_t \prod_i \otimes_t p(y_i|f_i) d\boldsymbol{f} && (5.21) \\ &= \int q(\boldsymbol{f}) \sum_i \log_t p(y_i|f_i) d\boldsymbol{f} && (5.22) \\ &= \sum_i \int \mathcal{T}(\boldsymbol{f}_i; \nu, 0, 1) \log_t p(\boldsymbol{y}_i|\boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}} f_i) d\boldsymbol{f}_i. && (5.23) \end{aligned}$$

Where the last step follows from the marginalization properties of the multivariate Student-t distribution and the relationship of the standard Student-t distribution to Student-t distributions with another mean and dispersion parameter (Kotz & Nadarajah (2004)). The n 1-dimensional problems are readily solved numerically via Monte Carlo or quadrature based methods (Burden & Faires (1997)).

Putting all these steps together, we obtain as an approximate variational lower bound for the evidence:

$$\begin{aligned} \log_t p(\boldsymbol{y}|\boldsymbol{X}) \gtrsim_t \sum_i \int & \mathcal{T}(\boldsymbol{f}_i; \nu, 0, 1) \log_t p(\boldsymbol{y}_i|\boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}} \boldsymbol{f}_i) d\boldsymbol{f}_i + \\ & \frac{\Psi_p}{(1-t)(\nu-2)} Tr \left\{ \boldsymbol{K}^{-1} \boldsymbol{V} \right\} + \log_t \mathcal{MVT}(\boldsymbol{m}; \nu, \boldsymbol{0}, \boldsymbol{K}) \\ & - \log_t \left(D^{\frac{1}{1-t}} \det |\boldsymbol{V}|^{-\frac{1}{2t}} \right), \end{aligned} \tag{5.24}$$

where $\gtrsim_t$ means that the left hand side is lower bounded by the right hand side under a t-relaxation, that is, it signals that the right hand side has been obtained by utilizing a t-relaxation. This bound is the evidence lower bound (ELBO) for generalized Student-t Process regression problems. We will refer to this method as VTP1.

Interestingly, in the limit as t goes to 1 or, equivalently, as ν approaches infinity, the variational bound for the GP case Nickisch & Rasmussen (2008) is almost recovered. This can be seen by looking at the different terms of 5.24 separately. While the convergence of the multiple integrals and of $\log_t \mathcal{MVT}(\boldsymbol{m}; \nu, \boldsymbol{0}, \boldsymbol{K})$ to their Gaussian counterparts is straightforward, the other two terms are more involved. For the term $\frac{\Psi_p}{(1-t)(\nu-2)} Tr\left\{\boldsymbol{K}^{-1}\boldsymbol{V}\right\}$, we need $\frac{\Psi_p}{(1-t)(\nu-2)}$ to approach $-\frac{1}{2}$. This can be done as follows:

$$\lim_{t\to 1} \frac{\Psi_p}{(1-t)(\nu-2)} = \lim_{t\to 1} \left(\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right) \det|\boldsymbol{K_{nn}}|^{\frac{1}{2}}} \right)^{1-t} \frac{1}{(1-t)(\nu-2)} \tag{5.25}$$

$$= \lim_{t\to 1} \left(\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{\frac{n}{2}}\Gamma\left(\frac{\nu}{2}\right) \det|\boldsymbol{K_{nn}}|^{\frac{1}{2}}} \right)^{1-t} \times \lim_{t\to 1} \frac{1}{(1-t)(\nu-2)} \tag{5.26}$$

$$= 1 \times \lim_{t\to 1} \frac{1}{(1-t)(n - \frac{2}{1-t} - 2)} \tag{5.27}$$

$$\approx 1 \times \lim_{t\to 1} \frac{1}{(1-t)(-\frac{2}{1-t})} \tag{5.28}$$

$$= -\frac{1}{2}, \tag{5.29}$$

where the third line follows from $\nu = -\frac{2}{1-t} - n$ and the forth line follows from the fact that, as t goes to 1, $n-2$ is negligible compared to $-\frac{2}{1-t}$. The last term, $\log_t \left(D^{\frac{1}{1-t}} \det|\boldsymbol{V}|^{-\frac{1}{2t}} \right)$, is the troublesome one. While it is close to what we would expect, we were neither able to show that the term converges to $\log\left(\det|\boldsymbol{V}|^{\frac{1}{2}}\right)$ nor that the subterm $D^{\frac{1}{1-t}}$ converges to 1, as t tends to 1.

5.2 Alternative, Variational Bound on Marginal Likelihood

It is also possible to derive a variational lower bound on the marginal likelihood without exponentiating the variational density by $\frac{1}{t}$. Using a proper variational distribution, we obtain for the first step:

$$\begin{aligned}
\log_t p(\boldsymbol{y}|\boldsymbol{X}) &= \log_t \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f} && (5.30)\\
&\approx_t \log_t \int p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})d\boldsymbol{f} && (5.31)\\
&= \log_t \int q(\boldsymbol{f})\frac{p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})}{q(\boldsymbol{f})}d\boldsymbol{f} && (5.32)\\
&\geq \int q(\boldsymbol{f})\log_t \frac{p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})}{q(\boldsymbol{f})}d\boldsymbol{f} && (5.33)\\
&= \int q(\boldsymbol{f})\log_t \frac{p(\boldsymbol{y}|\boldsymbol{f}) \otimes_t p(\boldsymbol{f}|\boldsymbol{X})}{q(\boldsymbol{f})}d\boldsymbol{f} && (5.34)\\
&= \int q(\boldsymbol{f})^t \left[\log_t p(\boldsymbol{y}|\boldsymbol{f}) + \log_t p(\boldsymbol{f}|\boldsymbol{X}) - \log_t q(\boldsymbol{f})\right] d\boldsymbol{f}. && (5.35)
\end{aligned}$$

In the statistical physics community, the $q(\boldsymbol{f})^t$ is known as the unnormalized escort distribution (Naudts (2004))[2] . When a Student-t distribution is used as a variational distribution, $q(\boldsymbol{f})$, the corresponding escort distribution, $\widetilde{q}(\boldsymbol{f})$, is Student-t (Ding et al. (2011)) with:

$$\widetilde{q}(\boldsymbol{f}) \sim \mathcal{MVT}\left(\boldsymbol{f}; \nu+2, \boldsymbol{m}, \frac{\nu}{\nu+2}\boldsymbol{V}\right)$$

.

From this it follows that the expression can be rewritten as:

$$\begin{aligned}
\log p(\boldsymbol{y}|\boldsymbol{X}) \gtrsim_t & \left(\int q(\boldsymbol{f})^t d\boldsymbol{f}\right) \times \\
& \int \widetilde{q}(\boldsymbol{f}) \left[\log_t p(\boldsymbol{y}|\boldsymbol{f}) + \log_t p(\boldsymbol{f}|\boldsymbol{X}) - \log_t q(\boldsymbol{f})\right] d\boldsymbol{f} && (5.36)\\
= & \left(\int q(\boldsymbol{f})^t d\boldsymbol{f}\right) \times \\
& \left[\int \widetilde{q}(\boldsymbol{f}) \log_t p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} - D_t(q(\boldsymbol{f}) \parallel p(\boldsymbol{f}|\boldsymbol{X}))\right]. && (5.37)
\end{aligned}$$

There are two ways to further simplify this expression. Firstly, we can utilize the closed-form solution for the t-divergence for two multivariate Student-t distributions

[2] Escort distributions are of the form $\widetilde{p}(x) = \frac{p(x)^t}{\int p(x)^t dx}$ They are a one-parameter deformation of the original distribution $p(x)$ Naudts (2004)). Their most prominent appearance is in non-extensive statistical mechanics (Tsallis (1988), Tsallis & Brigatti (2004)). Escort distributions are used to compute expectations with respect to heavy-tailed distributions that are not defined otherwise (Amari (2016), Barbaresco & Nielsen (2017)). Instead of computing the expectation with respect to the heavy-tailed distribution, the escort version of the heavy-tailed distribution is used to compute the so called escort expectation.

(Ding et al. (2011)). Secondly, the n-dimensional integral with respect to the escort distribution can be expressed as n 1-dimensional integrals as shown for the previous variational bound.

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{X}) \gtrsim_t & \frac{\Gamma\left(\frac{\nu+2}{2}\right)\left((\nu+2)\pi\right)^{\frac{n}{2}}\left(\frac{\nu}{\nu+2}\right)^{\frac{n}{2}}\det|\boldsymbol{V}|^{\frac{1}{2}}}{\Gamma\left(\frac{\nu+2+n}{2}\right)}\times \\ & \Bigg[\sum_i \int \mathcal{T}(\boldsymbol{f}_i;\nu+2,0,1)\log_t p(\boldsymbol{y}_i|\boldsymbol{m}_i+\sqrt{\frac{\nu}{\nu+2}\boldsymbol{V}_{ii}\boldsymbol{f}_i})d\boldsymbol{f}_i - \\ & \Bigg(\frac{\Psi_q}{1-t}(1+\nu^{-1})-\frac{\Psi_p}{(1-t)\nu}Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{V}\right\}- \\ & \frac{\Psi_p}{(1-t)\nu}\boldsymbol{m}^T\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{m}-\frac{\Psi_p}{1-t}\Bigg)\Bigg].\end{aligned} \tag{5.38}$$

We will refer to this method as VTP2.

5.3 Approximate Posterior

For the variational Gaussian approximation, the variational distribution is the approximate posterior. This can be seen from the fact that the Kullback-Leibler divergence is minimized when posterior and variational distribution are identical (see section 2.3.2).

However, for VTP1, we are not optimizing a Kullback-Leibler divergence dirctly, but a t-relaxed version of it.

Moreover, the density of our variational distribution is exponentiated by $\frac{1}{t}$, that is, we are optimizing our relaxed divergence with respect to an improper distribution which is not equal to the variational distribution.

Nonetheless, in the limit of $t \rightarrow 1$, a Kullback-Leibler divergence is implicitly minimized when our derived ELBO is maximized. We are therefore assuming that for all practical purposes, t is close enough to 1 to justify using the variational distribution as approximate posterior, that is:

$$p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{y}) \approx \mathcal{T}(\boldsymbol{f};\nu;\boldsymbol{m},\boldsymbol{V}) = q(\boldsymbol{f}). \tag{5.39}$$

The same reasoning applies also to VTP2. In contrast to the Gaussian case, we cannot directly show that maximizing the variational bound is minimizing the discrepancy between variational distribution and posterior. Nevertheless, in the limit of $t \to 1$, the standard ELBO is obtained from the intermediate equation 5.36. Therefore, we argue that the variational distribution given in 5.39 is also an approximation of the posterior for the alternative Student-t variational approximation.

5.4 Predictions

Predictions based on the variational distribution $q(\boldsymbol{f})$ can be obtained in a similar manner to the t-Laplace approximation[3]. That is, if only an approximate posterior mean for f_* suffices, then there is a closed form for this expectation:

$$\begin{aligned}
\mathbb{E}_q\left[\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},x_*\right] &= \int \mathbb{E}\left[\boldsymbol{f}_*|\boldsymbol{f},\boldsymbol{X},x_*\right] q(\boldsymbol{f})d\boldsymbol{f} && (5.40)\\
&= \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\,\mathbb{E}_q\left[\boldsymbol{f}\right] && (5.41)\\
&= \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{m}. && (5.42)
\end{aligned}$$

Also the variance of the posterior predictive distribution can be derived in a similar manner to the t-Laplace approximation:

$$\begin{aligned}
\mathbb{V}_q\left[\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},x_*\right] &= \mathbb{E}_{q(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X})}\left[\mathbb{V}\left[\boldsymbol{f}_*|\boldsymbol{f},x_*\right]\right] + \mathbb{V}_{q(\boldsymbol{f})}\left[\mathbb{E}\left[\boldsymbol{f}_*|\boldsymbol{f},x_*\right]\right] && (5.43)\\
&= \frac{\nu + Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\,\mathbb{E}_{q(\boldsymbol{f})}\left[\boldsymbol{f}\boldsymbol{f}^T\right]\right\}}{\nu+n-2}\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}} + \\
&\quad \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\,\mathbb{V}_{q(\boldsymbol{f}|\boldsymbol{y},\boldsymbol{X})}\left[\boldsymbol{f}\right]\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*} && (5.44)\\
&= \underbrace{\frac{\nu + Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\left[\frac{\nu}{\nu-2}\boldsymbol{V} + \boldsymbol{m}\boldsymbol{m}^T\right]\right\}}{\nu+n-2}}_{:=\phi}\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}} + \\
&\quad \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\left(\frac{\nu}{\nu-2}\boldsymbol{V}\right)\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*}. && (5.45)
\end{aligned}$$

expanding $\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}}$ as

$$\boldsymbol{K}_{\boldsymbol{f}_*|\boldsymbol{f}} = \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}_*} - \boldsymbol{K}_{\boldsymbol{n}_*\boldsymbol{n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn}_*},$$

[3] The equations for the predictions apply to both variational bounds

gives

$$\begin{aligned}\mathbb{V}_q\left[\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},x_*\right] = \phi\boldsymbol{K}_{\boldsymbol{n_*n_*}} - \\ \boldsymbol{K}_{\boldsymbol{n_*n}}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\left(\phi\boldsymbol{K}_{\boldsymbol{nn}} - \frac{\nu}{\nu-2}\boldsymbol{V}\right)\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{K}_{\boldsymbol{nn_*}}.\end{aligned} \tag{5.46}$$

If the expectation is required for the target variable y_* itself, we can use a numerical integration scheme with the the posterior predictive distribution for $\boldsymbol{f}_*$. Analogously to the t-Laplace approximation, the variational Student-t approximation does not have a closed-form solution for the predictive posterior distribution.

5.5 Optimization of Variational Parameters

In this section, we demonstrate how the gradients for the variational parameters can be derived.

The critical part in deriving the gradients is the intractable integral in the ELBOs. For our derivation, we will follow the route that Nickisch & Rasmussen (2008) and Sheth et al. (2015) used to obtain gradients for the GP case. For the variational Gaussian approximation, there is an alternative way via Fourier analysis to obtain gradients for the integral (see Opper & Archambeau (2009)).

In order to compute the gradients, we need an additional identity from Petersen & Pedersen (2012):

$$\frac{\partial}{\partial\boldsymbol{V}}\det|\boldsymbol{V}| = \det|\boldsymbol{V}|\boldsymbol{V}^{-1}. \tag{5.47}$$

This identity in conjunction with the ones used for the t-Laplace approximation, results in:

$$\frac{\partial ELBO_{VTP1}}{\partial\boldsymbol{m}} = \frac{\partial A}{\partial\boldsymbol{m}} + \frac{2\Psi_p}{1-t}(\nu\boldsymbol{K}_{\boldsymbol{nn}})^{-1}\boldsymbol{m} \tag{5.48}$$

$$\frac{\partial ELBO_{VTP1}}{\partial\boldsymbol{V}} = \frac{\partial A}{\partial\boldsymbol{V}} + \frac{\Psi_p}{(1-t)(\nu-2)}\boldsymbol{K}_{\boldsymbol{nn}}^{-1} - D\left(\frac{t-1}{2t}\right)\det|\boldsymbol{V}|^{\frac{t-1}{2t}}\boldsymbol{V}^{-1}. \tag{5.49}$$

On the one hand, the gradient of the intractable integral with respect to the location parameter of the variational Student-t distribution is given by:

$$\frac{\partial A}{\partial \boldsymbol{m}} = \frac{\partial}{\partial \boldsymbol{m}} \sum_i \int \mathcal{T}(\boldsymbol{f}_i; \nu, 0, 1) \log_t p(\boldsymbol{y}_i | \boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}} \boldsymbol{f}_i) d\boldsymbol{f}_i \tag{5.50}$$

$$= \sum_i \int \mathcal{T}(\boldsymbol{f}_i; \nu, 0, 1) \frac{\partial \log_t p(\boldsymbol{y}_i | \boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \frac{\partial \boldsymbol{l}_i}{\partial \boldsymbol{m}} d\boldsymbol{f}_i \tag{5.51}$$

$$= \mathbb{E}_{\mathcal{T}(\boldsymbol{f}_i;\nu,0,1)} \left[\frac{\partial \log_t p(\boldsymbol{y}_i | \boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \right] \Bigg|_{i=1..n}, \tag{5.52}$$

where

$$\boldsymbol{l}_i = \boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}} \boldsymbol{f}_i \tag{5.53}$$

$$\frac{\partial \boldsymbol{l}_j}{\partial \boldsymbol{m}_i} = \begin{cases} 1, & \text{if } j = i \\ 0, & \text{otherwise} \end{cases} \tag{5.54}$$

Equation 5.52 states that n 1-dimensional integrals need to be solved numerically in order to compute the gradient of the intractable integral with respect to the mean parameter of the variational distribution.

On the other hand, the gradient with respect to the dispersion parameter of the variational Student-t distribution is given by:

$$\frac{\partial A}{\partial \boldsymbol{V}} = \frac{\partial}{\partial \boldsymbol{V}} \sum_i \int \mathcal{T}(\boldsymbol{f}_i; \nu, 0, 1) \log_t p(\boldsymbol{y}_i | \boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}} \boldsymbol{f}_i) d\boldsymbol{f}_i \tag{5.55}$$

$$= \sum_i \int \mathcal{T}(\boldsymbol{f}_i; \nu, 0, 1) \frac{\partial \log_t p(\boldsymbol{y}_i | \boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \frac{\partial \boldsymbol{l}_i}{\partial \boldsymbol{V}} d\boldsymbol{f}_i \tag{5.56}$$

$$= \frac{1}{2\sqrt{\boldsymbol{V}_{ii}}} \mathbb{E}_{\mathcal{T}(\boldsymbol{f}_i;\nu,0,1)} \left[\boldsymbol{f}_i \frac{\partial \log_t p(\boldsymbol{y}_i | \boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \right] \Bigg|_{i=1..n}, \tag{5.57}$$

where

$$\frac{\partial \boldsymbol{l}_j}{\partial \boldsymbol{V}_{ij}} = \begin{cases} \frac{\boldsymbol{f}_i}{2\sqrt{\boldsymbol{V}_{ii}}}, & \text{if } j = i \\ 0, & \text{otherwise} \end{cases} \tag{5.58}$$

From equation 5.57, we can see that also for the gradient of the dispersion parameter, only n 1-dimensional numerical integrals are necessary. In contrast to the mean parameter, the evaluated integrals are not represented as vector, but as diagonal of a $n \times n$ matrix.

Based on the derivations for VTP1, the gradients for VTP2 can be computed readily:

$$
\begin{aligned}
\frac{\partial ELBO_{VTP2}}{\partial \boldsymbol{m}} &= G \det |\boldsymbol{V}|^{\frac{1}{2}} \left[\frac{\partial A}{\partial \boldsymbol{m}} + \frac{2\Psi_p}{(1-t)\nu} \boldsymbol{m}^T \boldsymbol{K}_{\boldsymbol{nn}}^{-1} \right]. && (5.59)\\
\frac{\partial ELBO_{VTP2}}{\partial \boldsymbol{V}} &= G \det |\boldsymbol{V}|^{\frac{1}{2}} \boldsymbol{V}^{-1} \left[\frac{\partial A}{\partial \boldsymbol{V}} - \frac{1}{1-t}(1+\nu^{-1}) \frac{\partial}{\partial \boldsymbol{V}} \Psi_q + \frac{\Psi_p}{(1-t)\nu} \boldsymbol{K}_{\boldsymbol{nn}}^{-1} \right], && (5.60)
\end{aligned}
$$

with

$$
\frac{\partial}{\partial \boldsymbol{V}} \Psi_q = \frac{t-1}{2} \left(\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{(\pi\nu)^{\frac{n}{2}} \Gamma\left(\frac{\nu}{2}\right) \det |\boldsymbol{V}|^{\frac{1}{2}}} \right)^{1-t} \boldsymbol{V}^{-1} \tag{5.61}
$$

$$
G = \frac{\Gamma\left(\frac{\nu+2}{2}\right) ((\nu+2)\pi)^{\frac{n}{2}} \left(\frac{\nu}{\nu+2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{\nu+2+n}{2}\right)}. \tag{5.62}
$$

The gradients of the integrals for the second ELBO are given by:

$$
\frac{\partial A}{\partial \boldsymbol{m}} = \mathbb{E}_{\mathcal{T}(\boldsymbol{f}_i;\nu+2,0,1)} \left[\frac{\partial \log_t p(\boldsymbol{y}_i|\boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \right] \Bigg|_{i=1..n} \tag{5.63}
$$

$$
\frac{\partial A}{\partial \boldsymbol{V}} = \frac{\nu}{2(\nu-2)\sqrt{\frac{\nu}{\nu+2}\boldsymbol{V}_{ii}}} \mathbb{E}_{\mathcal{T}(\boldsymbol{f}_i;\nu+2,0,1)} \left[\boldsymbol{f}_i \frac{\partial \log_t p(\boldsymbol{y}_i|\boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \right] \Bigg|_{i=1..n} \tag{5.64}
$$

$$
\boldsymbol{l}_i = \boldsymbol{m}_i + \sqrt{\frac{\nu}{\nu+2}\boldsymbol{V}_{ii}} \boldsymbol{f}_i. \tag{5.65}
$$

These gradients can now be used with gradient-based optimization procedures to optimize the different ELBOs.

5.6 Optimization of Kernel Parameters

In contrast to the t-Laplace approximation, we need to optimize the kernel parameters with respect to the ELBOs of the two variational methods.

For the derivation, we will require some of the results and identities established in 4.6 for the t-Laplace approximation, namely equation 4.59 and 4.65. Equipped with these results, the partial derivatives with respect to the hyperparameters are derived straightforwardly:

$$
\begin{aligned}
\frac{\partial ELBO_{VTP1}}{\partial \boldsymbol{\theta}_i} = \frac{\partial}{\partial \boldsymbol{\theta}_i} \Bigg[& \sum_i \int \mathcal{T}(\boldsymbol{f}_i; \nu, 0, 1) \log_t p(\boldsymbol{y}_i | \boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}} \boldsymbol{f}_i) df_i + \\
& \frac{\Psi_p}{(1-t)(\nu-2)} Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1} \boldsymbol{V}\right\} + \log_t \mathcal{MVT}(\boldsymbol{m}; \nu, \boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{nn}}) \\
& - \log_t \left(D^{\frac{1}{1-t}} \det |\boldsymbol{V}|^{\frac{t-1}{2t}} \right) \Bigg]
\end{aligned}
\tag{5.66}
$$

$$
\begin{aligned}
= & \frac{1}{(1-t)(\nu-2)} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \Psi_p \right) Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1} \boldsymbol{V}\right\} - \\
& \frac{\Psi_p}{(1-t)(\nu-2)} Tr\left\{ \boldsymbol{V} \boldsymbol{K}_{\boldsymbol{nn}}^{-1} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{\boldsymbol{nn}} \right) \boldsymbol{K}_{\boldsymbol{nn}}^{-1} \right\} + \\
& \frac{1}{(1-t)} \frac{\partial}{\partial \boldsymbol{\theta}_i} \Psi_p + \frac{1}{(1-t)} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \Psi_p \right) \boldsymbol{m}^T \left(\nu \boldsymbol{K}_{\boldsymbol{nn}} \right)^{-1} \boldsymbol{m} \\
& - \frac{\Psi_p}{(1-t)\nu} \boldsymbol{m}^T \boldsymbol{K}_{\boldsymbol{nn}}^{-1} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{\boldsymbol{nn}} \right) \boldsymbol{K}_{\boldsymbol{nn}}^{-1} \boldsymbol{m}.
\end{aligned}
\tag{5.67}
$$

and for VTP2:

$$\begin{aligned}\frac{\partial ELBO_{VTP2}}{\partial \boldsymbol{\theta}_i} &= \frac{\Gamma\left(\frac{\nu+2}{2}\right)((\nu+2)\pi)^{\frac{n}{2}}\left(\frac{\nu}{\nu+2}\right)^{\frac{n}{2}}\det|\boldsymbol{V}|^{\frac{1}{2}}}{\Gamma\left(\frac{\nu+2+n}{2}\right)} \times \frac{\partial}{\partial\boldsymbol{\theta}_i} \\ &\left[\sum_i \int \mathcal{T}(\boldsymbol{f}_i;\nu+2,0,1)\log_t p(\boldsymbol{y}_i|\boldsymbol{m}_i+\sqrt{\frac{\nu}{\nu+2}\boldsymbol{V}_{ii}}\boldsymbol{f}_i)d\boldsymbol{f}_i - \right. \\ &\left(\frac{\Psi_q}{1-t}(1+\nu^{-1}) - \frac{\Psi_p}{(1-t)\nu}Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{V}\right\} - \right. \\ &\left.\left.\frac{\Psi_p}{(1-t)\nu}\boldsymbol{m}^T\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{m} - \frac{\Psi_p}{1-t}\right)\right]\end{aligned} \tag{5.68}$$

$$\begin{aligned} &= \frac{\Gamma\left(\frac{\nu+2}{2}\right)((\nu+2)\pi)^{\frac{n}{2}}\left(\frac{\nu}{\nu+2}\right)^{\frac{n}{2}}\det|\boldsymbol{V}|^{\frac{1}{2}}}{\Gamma\left(\frac{\nu+2+n}{2}\right)} \times \\ &\left[\frac{1}{(1-t)\nu}\left(\frac{\partial}{\partial\boldsymbol{\theta}_i}\Psi_p\right)Tr\left\{\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{V}\right\} - \right. \\ &\frac{\Psi_p}{(1-t)\nu}Tr\left\{\boldsymbol{V}\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\left(\frac{\partial}{\partial\boldsymbol{\theta}_i}\boldsymbol{K}_{\boldsymbol{nn}}\right)\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\right\} + \\ &\frac{1}{(1-t)}\frac{\partial}{\partial\boldsymbol{\theta}_i}\Psi_p + \frac{1}{(1-t)\nu}\left(\frac{\partial}{\partial\boldsymbol{\theta}_i}\Psi_p\right)\boldsymbol{m}^T\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{m} \\ &\left. - \frac{\Psi_p}{(1-t)\nu}\boldsymbol{m}^T\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\left(\frac{\partial}{\partial\boldsymbol{\theta}_i}\boldsymbol{K}_{\boldsymbol{nn}}\right)\boldsymbol{K}_{\boldsymbol{nn}}^{-1}\boldsymbol{m}\right].\end{aligned} \tag{5.69}$$

Analogously to the t-Laplace case, the change of the variational parameters with respect to the hyperparameters needs to be taken into consideration to obtain the total derivative. However, Nickisch & Rasmussen (2008) report good results without taking these contributions from the variational parameters into consideration. As the gradients are already complex, we follow their approach and do not consider these implicit terms in the gradient calculation[4].

5.6.1 Comparison of Bounds

In this section, we briefly compare the different evidence lower bounds for VGP (from Nickisch & Rasmussen (2008)), VTP1, and VTP2. Equations 5.70, 5.71, and 5.72 show the evidence lower bounds for VGP, VTP1, and VTP2, respectively.

On first inspection, it is apparent that the bounds for VGP and VTP1 look rather similar, we have also shown in section 5.1 that, as t goes to 1, many elements of

[4] That is, terms such as $\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{V}$ and $\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{m}$ are missing from the gradient computation

the VGP bound are recovered from the VTP1. Nevertheless, we were not able to completely recover the VGP bound, due to the last term, $\log_t \left(D^{\frac{1}{1-t}} \det |\boldsymbol{V}|^{-\frac{1}{2t}} \right)$.

In contrast, the ELBO for the VTP2 method does not have such a close resemblance to the other two bounds. There are some common features, that is, the sum over the integrals, the trace operator over $\boldsymbol{K}^{-1}\boldsymbol{V}$, or the term $\boldsymbol{m}^T\boldsymbol{K}^{-1}\boldsymbol{m}$, which is hidden in the $\log \mathcal{MVN}$ and $\log_t \mathcal{MVT}$ terms of VGP and VTP1, respectively. However, neither VGP nor VTP1 rely on the escort distribution of $q(\boldsymbol{f})^t$ to derive their bounds, therefore none of them has a term resembling the leading factor (i.e. the fraction before the ×-operator) of VTP2.

Furthermore, another interesting difference between the bounds is that the VGP is an exact less-than-or-equal relationship, while the other methods rely on t-relaxation.

$$\begin{aligned} ELBO_{VGP} \geq & \sum_i \int \mathcal{N}(\boldsymbol{f}_i; 0, 1) \log p(\boldsymbol{y}_i | \boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}} \boldsymbol{f}_i) d\boldsymbol{f}_i - \\ & \frac{1}{2} Tr \left\{ \boldsymbol{K}^{-1} \boldsymbol{V} \right\} + \log \mathcal{MVN}(\boldsymbol{m}; \boldsymbol{0}, \boldsymbol{K}) \\ & + \log \left(\det |\boldsymbol{V}|^{-\frac{1}{2}} \right) \end{aligned} \tag{5.70}$$

$$\begin{aligned} ELBO_{VTP1} \gtrsim_t & \sum_i \int \mathcal{T}(\boldsymbol{f}_i; \nu, 0, 1) \log_t p(\boldsymbol{y}_i | \boldsymbol{m}_i + \sqrt{\boldsymbol{V}_{ii}} \boldsymbol{f}_i) d\boldsymbol{f}_i + \\ & \frac{\Psi_p}{(1-t)(\nu-2)} Tr \left\{ \boldsymbol{K}^{-1} \boldsymbol{V} \right\} + \log_t \mathcal{MVT}(\boldsymbol{m}; \nu, \boldsymbol{0}, \boldsymbol{K}) \\ & - \log_t \left(D^{\frac{1}{1-t}} \det |\boldsymbol{V}|^{-\frac{1}{2t}} \right) \end{aligned} \tag{5.71}$$

$$\begin{aligned} ELBO_{VTP2} \gtrsim_t & \frac{\Gamma\left(\frac{\nu+2}{2}\right) ((\nu+2)\pi)^{\frac{n}{2}} \left(\frac{\nu}{\nu+2}\right)^{\frac{n}{2}} \det |\boldsymbol{V}|^{\frac{1}{2}}}{\Gamma\left(\frac{\nu+2+n}{2}\right)} \times \\ & \Bigg[\sum_i \int \mathcal{T}(\boldsymbol{f}_i; \nu+2, 0, 1) \log_t p(\boldsymbol{y}_i | \boldsymbol{m}_i + \sqrt{\frac{\nu}{\nu+2} \boldsymbol{V}_{ii} \boldsymbol{f}_i}) d\boldsymbol{f}_i - \\ & \Bigg(\frac{\Psi_q}{1-t}(1+\nu^{-1}) - \frac{\Psi_p}{(1-t)\nu} Tr \left\{ \boldsymbol{K}^{-1} \boldsymbol{V} \right\} - \\ & \frac{\Psi_p}{(1-t)\nu} \boldsymbol{m}^T \boldsymbol{K}^{-1} \boldsymbol{m} - \frac{\Psi_p}{1-t} \Bigg) \Bigg] \end{aligned} \tag{5.72}$$

Chapter 6

Variational Sparse Inducing Point Methods

The purpose of this chapter is to demonstrate how the idea of variational, sparse inducing points is extended to TP-based models. Two different bounds will be covered. The first one will be derived from lower bouding the marginal likelihood directly, while the second one will be based on the t-relaxation of intermediate results for the GP case.

6.1 Scalable, Variational Bound for Marginal Likelihood

The variational lower bound derived in this section is a conceptual generalization of the approach introduced by Hensman, Matthews & Ghahramani (2015).

In a first step, we establish a lower bound for the t-logarithm of the conditional likelihood of the observations $\boldsymbol{y}$ with respect to the function values $\boldsymbol{u}$, which are evaluated at the inducing points $\boldsymbol{z}$ (suppressing the dependence on the data).

$$\begin{aligned}\log_t p(\boldsymbol{y}|\boldsymbol{u}) &= \log_t \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})d\boldsymbol{f} && (6.1)\\ &\geq \int p(\boldsymbol{f}|\boldsymbol{u}) \log_t p(\boldsymbol{y}|\boldsymbol{f})d\boldsymbol{f} && (6.2)\\ &= \mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})}\left[\log_t p(\boldsymbol{y}|\boldsymbol{f})\right]. && (6.3)\end{aligned}$$

In the next step, we lower bound the t-relaxed log marginal likelihood in a similar way as with the variational Student-t approximation. However, in contrast to the variational Student-t approximation, the marginalization is done over the joint distribution of the observations and the function values at the inducing points, that is:

$$
\begin{aligned}
\log_t p(\boldsymbol{y}) &= \log_t \int p(\boldsymbol{y}|\boldsymbol{u}) \otimes p(\boldsymbol{u}) d\boldsymbol{u} \qquad (6.4)\\
&= \log_t \int q(\boldsymbol{u})^{\frac{1}{t}} \frac{p(\boldsymbol{y}|\boldsymbol{u}) \otimes p(\boldsymbol{u})}{q(\boldsymbol{u})^{\frac{1}{t}}} d\boldsymbol{u} \qquad (6.5)\\
&\geq \int q(\boldsymbol{u})^{\frac{1}{t}} \log_t \frac{p(y|\boldsymbol{u}) \otimes p(\boldsymbol{u})}{q(\boldsymbol{u})^{\frac{1}{t}}} d\boldsymbol{u} \qquad (6.6)\\
&= \int q(\boldsymbol{u}) \log_t p(\boldsymbol{y}|\boldsymbol{u}) d\boldsymbol{u} + \\
&\quad \int q(\boldsymbol{u})(\log_t p(\boldsymbol{u}) - log_t q(\boldsymbol{u})^{\frac{1}{t}}) d\boldsymbol{u} \qquad (6.7)\\
&\geq \underbrace{\int q(\boldsymbol{u}) \log_t p(\boldsymbol{y}|\boldsymbol{u}) d\boldsymbol{u}}_{:=A} + \\
&\quad \underbrace{\int q(\boldsymbol{u}) \log_t p(\boldsymbol{u}) d\boldsymbol{u}}_{:=B} - \underbrace{\int q(\boldsymbol{u}) \log_t q(\boldsymbol{u})^{\frac{1}{t}} d\boldsymbol{u}}_{:=C}. \qquad (6.8)
\end{aligned}
$$

B and C have comparable terms in the variational Student-t approximation, with the important difference that they are now with respect to the function values at the inducing points. However, A is more difficult, as we do not know $p(\boldsymbol{y}|\boldsymbol{u})$. Nonetheless, we can use the lower bound 6.3 to further simplify A:

$$
\begin{aligned}
\log_t p(\boldsymbol{y}) &\geq \int q(\boldsymbol{u})\, \mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})} \left[\log_t p(\boldsymbol{y}|\boldsymbol{f})\right] d\boldsymbol{u} + \\
&\quad \int q(\boldsymbol{u})(\log_t p(\boldsymbol{u}) - \log_t q(\boldsymbol{u})^{\frac{1}{t}}) d\boldsymbol{u} \qquad (6.9)\\
&= \int_{\mathbb{U}} q(\boldsymbol{u}) \left(\int_{\mathbb{F}} p(\boldsymbol{f}|\boldsymbol{u}) \log_t p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} \right) d\boldsymbol{u} + \\
&\quad \int q(\boldsymbol{u})(\log_t p(\boldsymbol{u}) - \log_t q(\boldsymbol{u})^{\frac{1}{t}}) d\boldsymbol{u}. \qquad (6.10)
\end{aligned}
$$

Using the Fubini-Tonelli theorem (Schilling (2017)) to change the order of integration and the fact that $q(\boldsymbol{f}) = \int p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u}) d\boldsymbol{u}$, we obtain:

$$\begin{aligned}\log_t p(\boldsymbol{y}) \geq & \int q(\boldsymbol{f}) \log_t p(y|\boldsymbol{f}) d\boldsymbol{f} + \\ & \int q(\boldsymbol{u})(\log_t p(\boldsymbol{u}) - \log_t q(\boldsymbol{u})^{\frac{1}{t}}) d\boldsymbol{u} \end{aligned} \tag{6.11}$$

$$\begin{aligned}= & \int q(\boldsymbol{f}) \log_t p(y|\boldsymbol{f}) d\boldsymbol{f} + \\ & \int q(\boldsymbol{u}) \log_t p(\boldsymbol{u}) d\boldsymbol{u} - \int q(\boldsymbol{u}) \log_t q(\boldsymbol{u})^{\frac{1}{t}} d\boldsymbol{u}. \end{aligned} \tag{6.12}$$

The first integral with respect to $\boldsymbol{f}$ is different from the variational Student-t approximation case. This is due to the variational distribution $q(\boldsymbol{f})$ being the result of marginalizing over the induced function values in the joint distribution $q(\boldsymbol{f}, \boldsymbol{u})$. Due to the marginalization being intractable, we are using the following variational distribution as an approximation:

$$q(\boldsymbol{f}) \sim \mathcal{MVT}\left(\boldsymbol{f}; \nu, \boldsymbol{\alpha}, \boldsymbol{B}\right), \tag{6.13}$$

with

$$\boldsymbol{\alpha} = \boldsymbol{K_{nm}} \boldsymbol{K_{mm}^{-1}} \boldsymbol{m} \tag{6.14}$$

$$\boldsymbol{B} = \boldsymbol{K_{nn}} + \boldsymbol{K_{nm}} \boldsymbol{K_{mm}^{-1}} \left(\boldsymbol{V} - \boldsymbol{K_{mm}}\right) \boldsymbol{K_{mm}^{-1}} \boldsymbol{K_{mn}}. \tag{6.15}$$

In appendix B.2, more details regarding the derivations of this approximation can be found. It is important to notice that this marginalization is reminiscent of the problem of deriving the posterior predictive distribution (for one of the methods so far). Comparing the results, we can see that the approximation is centered at the correct mean, however, the variance of our approximate distribution is substantially different to the ones obtained for the posterior predictive distributions.

Combining these results with the results from the previous section, a new, approximate ELBO is obtained:

$$
\begin{aligned}
\log_t p(\boldsymbol{y}|\boldsymbol{X}) \gtrsim_t \sum_i \int & \mathcal{T}(\boldsymbol{f}_i; \nu, 0, 1) \log_t p(\boldsymbol{y}_i|\boldsymbol{\alpha}_i + \sqrt{\boldsymbol{B}_{ii}}\boldsymbol{f}_i) d\boldsymbol{f}_i + \\
& \frac{\Psi_{p_u}}{(1-t)(\nu-2)} Tr\left\{ \boldsymbol{K}_{\boldsymbol{mm}}^{-1}\boldsymbol{V} \right\} + \log_t \mathcal{MVT}(\boldsymbol{\alpha}; \nu, \boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{mm}}) \\
& - \log_t \left(D^{\frac{1}{1-t}} \det |\boldsymbol{V}|^{-\frac{1}{2t}} \right),
\end{aligned}
\tag{6.16}
$$

where

$$
D = \frac{\Gamma\left(\frac{\nu+m}{2}\right)^{\frac{1}{t}} \Gamma\left(\frac{\rho}{2}\right) (\nu\pi)^{\frac{m(t-1)}{2t}}}{\Gamma\left(\frac{\nu}{2}\right)^{\frac{1}{t}} \Gamma\left(\frac{\rho+m}{2}\right)}
\tag{6.17}
$$

$$
\rho = \frac{2}{(t(t-1))} - m
\tag{6.18}
$$

$$
\Psi_{p_u} = \left(\frac{\Gamma\left(\frac{\nu+m}{2}\right)}{(\pi\nu)^{\frac{m}{2}} \Gamma\left(\frac{\nu}{2}\right) \det |\boldsymbol{K}_{\boldsymbol{mm}}|^{\frac{1}{2}}} \right)^{1-t}.
\tag{6.19}
$$

It is noteworthy, that this bound has a run time complexity of $\mathcal{O}(nm^2)$, similar to its Gaussian counterparts. We will refer to this method as sVTP1.

6.2 Alternative, Variational Bound on the Marginal Likelihood

The variational bounds presented so far relied on t-relaxation of the log marginal likelihood. However, starting the derivations from the original, marginal likelihood is not a necessity. In this section, an alternative, scalable variational bound that is derived by t-relaxing an intermediate result from Hensman, Matthews & Ghahramani (2015) is presented[1]. In Hensman et. al. the following bound on the marginal likelihood is established:

[1] The reader might wonder why the VTP2 method is not extended. The reason for this is given in appendix B.3

$$\log p(\boldsymbol{y}|\boldsymbol{X}) \geq \mathbb{E}_{q(\boldsymbol{f})}\left[\log p(\boldsymbol{y}|\boldsymbol{f})\right] - KL(q(\boldsymbol{u})) \parallel p(\boldsymbol{u}))), \tag{6.20}$$

t-relaxing this bound gives:

$$\log_t p(\boldsymbol{y}|\boldsymbol{X}) \gtrsim_t \mathbb{E}_{q(\boldsymbol{f})}\left[\log_t p(\boldsymbol{y}|\boldsymbol{f})\right] - D_t(q(\boldsymbol{u})) \parallel p(\boldsymbol{u}))), \tag{6.21}$$

as $q(\boldsymbol{u})$ and $p(\boldsymbol{u})$ are both multivariate Student-t distributions, the closed-form solution for the t-Divergence from Ding et al. (2011) can be used to produce an alternative variational lower bound:

$$\begin{aligned}\log_t p(\boldsymbol{y}|\boldsymbol{X}) \gtrsim_t \mathbb{E}_{q(\boldsymbol{f})}\left[\log_t p(\boldsymbol{y}|\boldsymbol{f})\right] - \Bigg[& \frac{\Psi_q}{1-t}(1+\nu^{-1}) - \\ & \frac{\Psi_p}{(1-t)\nu} Tr\left\{\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\boldsymbol{V}\right\} - \frac{\Psi_p}{(1-t)\nu}\boldsymbol{m}^T\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\boldsymbol{m} - \frac{\Psi_p}{1-t}\Bigg],\end{aligned} \tag{6.22}$$

For $q(\boldsymbol{f})$, the variational distribution described in equation 6.13 is used.

Analogously to sVTP1, this bound has a run time complexity of $\mathcal{O}(nm^2)$. We will refer to this method as sVTP2.

6.3 Approximate Posterior

Following Hensman, Matthews & Ghahramani (2015) and our arguments made for the approximate posterior of the variational Student-t approximation, the approximate posterior is given by:

$$p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{X}, \boldsymbol{y}) \approx q(\boldsymbol{f}, \boldsymbol{u}), \tag{6.23}$$

where the approximate marginal distribution for $\boldsymbol{f}$ is given in equation 6.13 and the one for $\boldsymbol{u}$ is distributed according to $\mathcal{T}(\nu, \boldsymbol{m}, \boldsymbol{V})$.

6.4 Predictions

Using the approximate posterior and the property of the induced function values as sufficient statistics (following Titsias (2009) and Hensman, Matthews & Ghahramani (2015)), we can obtain an approximate (intractable) posterior predictive distribution:

$$\begin{aligned}
p(\boldsymbol{f}_*|X, y, x_*) &= \int_{\mathbb{U}} \int_{\mathbb{F}} p(\boldsymbol{f}_*|\boldsymbol{f}, \boldsymbol{u}, X, x_*) p(\boldsymbol{f}, \boldsymbol{u}) d\boldsymbol{f} d\boldsymbol{u} && (6.24)\\
&\approx \int_{\mathbb{U}} \int_{\mathbb{F}} p(\boldsymbol{f}_*|\boldsymbol{f}, \boldsymbol{u}, X, y) p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u}) d\boldsymbol{f} d\boldsymbol{u} && (6.25)\\
&= \int_{\mathbb{U}} \int_{\mathbb{F}} p(\boldsymbol{f}_*|\boldsymbol{u}, x_*) p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u}) d\boldsymbol{f} d\boldsymbol{u} && (6.26)\\
&= \int_{\mathbb{U}} p(\boldsymbol{f}_*|\boldsymbol{u}, x_*) q(\boldsymbol{u}) d\boldsymbol{u} && (6.27)\\
&= q(\boldsymbol{f}_*). && (6.28)
\end{aligned}$$

Finally, the posterior predictive mean and variance for the variational sparse inducing point methods is given by (derivation is similar to the variational Student-t approximation):

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{f}_*)}\left[\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, x_*\right] &= \boldsymbol{K_{n_*m}} \boldsymbol{K_{mm}^{-1}} \boldsymbol{m} && (6.29)\\
\mathbb{V}_{q(\boldsymbol{f}_*)}\left[\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, x_*\right] &= \phi \boldsymbol{K_{n_*n_*}} - \\
&\quad \boldsymbol{K_{n_*m}} \boldsymbol{K_{mm}^{-1}} \left(\phi \boldsymbol{K_{mm}} - \frac{\nu}{\nu - 2} \boldsymbol{V}\right) \boldsymbol{K_{mm}^{-1}} \boldsymbol{K_{nn_*}}, && (6.30)
\end{aligned}$$

where

$$\phi = \frac{\nu + Tr\left\{\boldsymbol{K_{mm}^{-1}}\left[\frac{\nu}{\nu-2}\boldsymbol{V} + \boldsymbol{m}\boldsymbol{m}^T\right]\right\}}{\nu + n - 2}. \qquad (6.31)$$

6.5 Optimization of Variational Parameters

The derivation of the gradients for the variational parameters resembles our work done in section 5.5. The main difference is that the parameters of the integrals, $\boldsymbol{\alpha}$

and $\boldsymbol{B}_{ii}$, are not directly optimized but via the variational parameters $\boldsymbol{m}$ and $\boldsymbol{V}$, that is:

$$\begin{aligned} \frac{\partial ELBO_{sVTP1}}{\partial \boldsymbol{m}} &= \frac{\partial A}{\partial \boldsymbol{\alpha}} \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{m}} + \frac{2\Psi_{p_u}}{1-t} (\nu \boldsymbol{K_{mm}})^{-1} \boldsymbol{m} && (6.32) \\ \frac{\partial ELBO_{sVTP1}}{\partial \boldsymbol{V}} &= \sum_i \frac{\partial A}{\partial \boldsymbol{B}_{ii}} \frac{\partial \boldsymbol{B}_{ii}}{\partial \boldsymbol{V}} + \frac{\Psi_p}{(1-t)(\nu-2)} \boldsymbol{K_{mm}^{-1}} - \\ & \quad D \left(\frac{t-1}{2t} \right) \det |\boldsymbol{V}|^{\frac{t-1}{2t}} \boldsymbol{V}^{-1}. && (6.33) \end{aligned}$$

Moreover, the gradient of the intractable integral with respect to the location parameter of the variational Student-t distribution is given by:

$$\frac{\partial A}{\partial \boldsymbol{\alpha}} \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{m}} = \left(\mathbb{E}_{\mathcal{T}(\boldsymbol{f}_i;\nu,0,1)} \left[\frac{\partial \log_t p(\boldsymbol{y}_i|\boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \right] \Bigg|_{i=1..n} \right)^T \boldsymbol{K_{nm}} \boldsymbol{K_{mm}^{-1}}, \tag{6.34}$$

where

$$\boldsymbol{l}_i = \boldsymbol{\alpha}_i + \sqrt{\boldsymbol{B}_{ii}} \boldsymbol{f}_i. \tag{6.35}$$

Regarding the gradient with respect to the dispersion parameter of the variational Student-t distribution:

$$\frac{\partial A}{\partial \boldsymbol{B}_{ii}} \frac{\partial \boldsymbol{B}_{ii}}{\partial \boldsymbol{V}} = \frac{1}{2\sqrt{\boldsymbol{B}_{ii}}} \mathbb{E}_{\mathcal{T}(\boldsymbol{f}_i;\nu,0,1)} \left[\boldsymbol{f}_i \frac{\partial \log_t p(\boldsymbol{y}_i|\boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \right] \boldsymbol{K_{mm}^{-1}} \boldsymbol{K}_{\boldsymbol{m}i} \boldsymbol{K}_{i\boldsymbol{m}} \boldsymbol{K_{mm}^{-1}} \Bigg|_{i=1..n}, \tag{6.36}$$

where the result for $\frac{\partial \boldsymbol{B}_{ii}}{\partial \boldsymbol{V}}$ is derived from the identity provided in Petersen & Pedersen (2012):

$$\frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{b} = \boldsymbol{b}\boldsymbol{a}^T. \tag{6.37}$$

Also the gradients for sVTP2 are readily available:

$$\begin{aligned} \frac{\partial ELBO_{sVTP2}}{\partial \boldsymbol{m}} &= \frac{\partial A}{\partial \boldsymbol{\alpha}} \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{m}} + \frac{\Psi_p}{(1-t)\nu} \boldsymbol{K_{mm}^{-1}} \boldsymbol{m} && (6.38) \\ \frac{\partial ELBO_{sVTP2}}{\partial \boldsymbol{V}} &= \sum_i \frac{\partial A}{\partial \boldsymbol{B}_{ii}} \frac{\partial \boldsymbol{B}_{ii}}{\partial \boldsymbol{V}} - \frac{1}{1-t} (1+\nu^{-1}) \frac{\partial}{\partial \boldsymbol{V}} \Psi_q + \\ & \quad \frac{\Psi_p}{(1-t)(\nu-2)} \boldsymbol{K_{mm}^{-1}}, && (6.39) \end{aligned}$$

with $\frac{\partial}{\partial \boldsymbol{V}}\Psi_q$ defined in equation 5.61.

The gradients for the integrals of the second inducing point method are given by equations 6.42 and 6.43. That is, there is no difference between the two sparse inducing point methods in terms of gradients for the intractable integrals.

6.6 Optimization of Kernel Parameters and Inducing Points

The gradients for the variational sparse inducing point methods follow mainly from the results obtained for the variational Student-t approximations (see equation 5.67 and 5.69). However, in contrast to the variational Student-t approximation, the sparse methods have an explicit dependence between the intractable integral and the parameters of the kernel based on 6.14 and 6.15. That is:

$$\begin{aligned}\frac{\partial ELBO_{sVTP1}}{\partial \boldsymbol{\theta}_i} = & \sum_j \left(\frac{\partial A}{\partial \boldsymbol{\alpha}_j}\frac{\partial \boldsymbol{\alpha}_j}{\partial \boldsymbol{\theta}_i} + \frac{\partial A}{\partial \boldsymbol{B}_{jj}}\frac{\partial \boldsymbol{B}_{jj}}{\partial \boldsymbol{\theta}_i} \right) + \\ & \frac{1}{(1-t)(\nu-2)}\left(\frac{\partial}{\partial \boldsymbol{\theta}_i}\Psi_{p_u}\right) Tr\left\{\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\boldsymbol{V}\right\} - \\ & \frac{\Psi_{p_u}}{(1-t)(\nu-2)} Tr\left\{\boldsymbol{V}\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\left(\frac{\partial}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}_{\boldsymbol{mm}}\right)\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\right\} + \\ & \frac{1}{(1-t)}\frac{\partial}{\partial \boldsymbol{\theta}_i}\Psi_{p_u} + \frac{1}{(1-t)}\left(\frac{\partial}{\partial \boldsymbol{\theta}_i}\Psi_{p_u}\right)\boldsymbol{m}^T\left(\nu \boldsymbol{K}_{\boldsymbol{mm}}\right)^{-1}\boldsymbol{m} \\ & -\frac{\Psi_{p_u}}{(1-t)\nu}\boldsymbol{m}^T\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\left(\frac{\partial}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}_{\boldsymbol{mm}}\right)\boldsymbol{K}_{\boldsymbol{mm}}^{-1}\boldsymbol{m},\end{aligned} \tag{6.40}$$

and for the alternative ELBO:

$$
\begin{aligned}
\frac{\partial ELBO_{sVTP2}}{\partial \boldsymbol{\theta}_i} = &\sum_j \left(\frac{\partial A}{\partial \boldsymbol{\alpha}_j} \frac{\partial \boldsymbol{\alpha}_j}{\partial \boldsymbol{\theta}_i} + \frac{\partial A}{\partial \boldsymbol{B}_{jj}} \frac{\partial \boldsymbol{B}_{jj}}{\partial \boldsymbol{\theta}_i} \right) + \\
&\frac{1}{(1-t)\nu} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \Psi_{p_u} \right) Tr \left\{ \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \boldsymbol{V} \right\} - \\
&\frac{\Psi_{p_u}}{(1-t)\nu} Tr \left\{ \boldsymbol{V} \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{\boldsymbol{mm}} \right) \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \right\} + \\
&\frac{1}{(1-t)} \frac{\partial}{\partial \boldsymbol{\theta}_i} \Psi_{p_u} + \frac{1}{(1-t)\nu} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \Psi_{p_u} \right) \boldsymbol{m}^T \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \boldsymbol{m} \\
&- \frac{\Psi_{p_u}}{(1-t)\nu} \boldsymbol{m}^T \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{\boldsymbol{mm}} \right) \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \boldsymbol{m}.
\end{aligned}
\tag{6.41}
$$

The integral terms are given by:

$$
\begin{aligned}
\frac{\partial A}{\partial \boldsymbol{\alpha}_j} \frac{\partial \boldsymbol{\alpha}_j}{\partial \boldsymbol{\theta}_i} = &\mathbb{E}_{\mathcal{T}(\boldsymbol{f}_i;\nu,0,1)} \left[\frac{\partial \log_t p(\boldsymbol{y}_i | \boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \right] \times \\
&\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{jm} \right) \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \boldsymbol{m} - \boldsymbol{K}_{jm} \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{\boldsymbol{mm}} \right) \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \boldsymbol{m} \right],
\end{aligned}
\tag{6.42}
$$

and

$$
\begin{aligned}
\frac{\partial A}{\partial \boldsymbol{B}_{jj}} \frac{\partial \boldsymbol{B}_{jj}}{\partial \boldsymbol{\theta}_i} = &\frac{1}{2\sqrt{\boldsymbol{B}_{ii}}} \mathbb{E}_{\mathcal{T}(\boldsymbol{f}_i;\nu,0,1)} \left[\boldsymbol{f}_i \frac{\partial \log_t p(\boldsymbol{y}_i | \boldsymbol{l}_i)}{\partial \boldsymbol{l}_i} \right] \times \\
&\left[\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{jj} - 2 \boldsymbol{K}_{jm} \boldsymbol{K}_{\boldsymbol{mm}}^{-1} (\boldsymbol{V} - \boldsymbol{K}_{\boldsymbol{mm}}) \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{mj} + \right. \\
&2 \boldsymbol{K}_{jm} \boldsymbol{K}_{\boldsymbol{mm}}^{-1} (\boldsymbol{V} - \boldsymbol{K}_{\boldsymbol{mm}}) \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{\boldsymbol{mm}} \right) \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \boldsymbol{K}_{mj} + \\
&\left. \boldsymbol{K}_{jm} \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \left(\frac{\partial}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}_{\boldsymbol{mm}} \right) \boldsymbol{K}_{\boldsymbol{mm}}^{-1} \boldsymbol{K}_{mj} \right].
\end{aligned}
\tag{6.43}
$$

We have again suppressed the implicit dependence between variational parameters and kernel parameters based on the results from Nickisch & Rasmussen (2008). By replacing $\boldsymbol{\theta}_i$ with $\boldsymbol{Z}_{ij}$ in equation 6.40 and 6.41, the gradients for the optimization of the inducing points $\boldsymbol{Z}$ are obtained.

Chapter 7

Experiments

In this section, we will firstly compare the t-Laplace approximation and the variational Student-t approximation on simulated data to get an idea of the behaviour of these methods in contrast to their GP counterparts. Then, we will test the methods on real world datasets. Finally, the TP sparse-inducing point methods are compared to their GP counterparts on larger datasets.

7.1 General

7.1.1 Model

For our experiments, we will focus on regression and binary classification models. For the regression, we are interested in the performance and behaviour of the methods for the following models:

$$\boldsymbol{f} \sim \mathcal{GP}(\boldsymbol{0}, \boldsymbol{K}) \tag{7.1}$$

$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{f}_i, \sigma^2), \tag{7.2}$$

and

$$\boldsymbol{f} \sim \mathcal{TP}(\nu, \boldsymbol{0}, \boldsymbol{K}) \tag{7.3}$$

$$\boldsymbol{y}_i \sim \mathcal{T}(\kappa, \boldsymbol{f}_i, \sigma^2). \tag{7.4}$$

We will use these first models to get an initial understanding of the impact of outliers. For the second type of models, the binary classification models, we will use the following two models:

$$\boldsymbol{f} \sim \mathcal{GP}(\mathbf{0}, \boldsymbol{K}_\theta) \tag{7.5}$$

$$\boldsymbol{y}_i \sim \mathrm{Ber}(\mathrm{sigm}(\boldsymbol{f}_i)), \tag{7.6}$$

and

$$\boldsymbol{f} \sim \mathcal{TP}(\nu, \mathbf{0}, \boldsymbol{K}_\theta) \tag{7.7}$$

$$\boldsymbol{y}_i \sim \mathrm{Ber}(\mathrm{sigm}(\boldsymbol{f}_i)). \tag{7.8}$$

The main foucs of our experiments will be on these two models. The reasons for doing so are threefold: Firstly, both problems can be nicely visualized in terms of regression curves and decision boundaries. Secondly, both problems are popular fields of research and there are implementations of algorithms for comparison available. Thirdly, the work of Shah et al. (2013) demonstrates the advantages of TPs for robust regression, while the recent work of Futami et al. (2017) suggests that TP-based models can lead to more robust results for classification problems compared to their GP counterparts. That is, there is already some evidence that TPs can outperform their GP counterparts in these tasks, so it is of particular interest to have efficient inference for these cases available.

7.1.2 Implementation

All the methods for the GTPR were implemented in Python/Tensorflow (Abadi et al. (2015)) with functionality, e.g. kernel implementation, taken from GPflow (Matthews et al. (2017)).

For the t-Laplace approximation, the posterior mode $\hat{\boldsymbol{f}}$ was found with the Newton-method presented in 4.5, whereas the kernel parameters were optimized with Scipy's (Jones et al. (2014)) implementation of the limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization algorithm (Nocedal & Wright (2006)). On the one hand, optimization of the posterior mode was stopped after either 50 optimization iterations had been conducted or there had not been an improvement in the posterior mode for 10 consecutive iterations. On the other hand, the L-BFGS optimization step was terminated after 500 iterations or earlier, if there was no improvement in the approximate marginal likelihood (based on Scipy default values for absolute or relative convergence). This procedure, i.e. finding posterior mode and then optimizing kernel parameters, was repeated for 10 times and the optimized parameters of the last iterations were used for predictions. The results of the t-Lapalce approximation are compared to the results from GPy's (GPy (since 2012)) implementation of Student-t regression and scikit-learns' (Pedregosa et al. (2011)) implementation of Gaussian process classification.

In contrast, the parameters of the variational Student-t approximations were optimized entirely with L-BFGS, whereas the optimization was stopped after 500 iterations or after there had not been an improvement in evidence lower bound (Scipy defaults). The results of the methods are compared with the variational methods provided in GPflow (Matthews et al. (2017)).

The optimization approach for the sparse inducing point methods are based on the work of Titsias (2009) and Hensman, Matthews, Filippone & Ghahramani (2015). That is, the inducing points were initialized with k-means clustering and there was an initial optimization sequence, where only the kernel and variational distribution parameters were tuned (50 iterations). Subsequently, the procedure for the variational Student-t approximations was followed with all the parameters being optimized simultaneously, i.e. including the inducing points.

As a covariance function, the squared-exponential kernel was used for all experiments. This decision was made due to the kernel's high popularity within the kernel methods community (Rasmussen & Williams (2005)).

Gradient computation was done via Tensorflow's automatic differentiation Baydin et al. (2018) and the intractable integrals were approximated via Gauss-Hermite quadrature based on the numpy (Oliphant (2006)) implementation.

The parameter t was chosen in order to use the simplification of the Student-t density described in section 4.1. That is,

$$t = \frac{2}{\nu + n} + 1,$$

where ν are the degrees of freedom of the distribution and n are the number of observations in the dataset. The experiments were conducted for 3, 10, and 25 degrees of freedom. In particular for the Student-t regression, the degrees of freedom for the observation likelihood, κ, were set to the degrees of freedom of the Student-t process prior during the optimization.

7.1.3 MCMC

MCMC is used as ground truth to compare the different models and approximation methods [1]. The regression and binary classification models were implemented in Stan (Carpenter et al. (2016)) to conduct inference by Hamiltonian Monte Carlo. For the hyperparameters of both models we have used truncated, heavy-tailed Cauchy distributions. For all the models, we ran 4 chains with 7000 samples and the first 2000 samples were discarded as warm up/burn in phase. The remaining 5000 samples were thinned by taking only every 20th sample. Convergence was diagnosed with split $\hat{R}$ provided in the Stan output (see Carpenter et al. (2016) or Brooks et al. (2011)).

It is important to notice, that MCMC requires prior distributions for the hyperparameters. These priors are not used for the approximation methods, that is, the approximation approaches approximate a model that is different. However, the priors used are rather uninformative and do not impact the result significantly [2].

7.2 Simulated Data - Regression

7.2.1 Datasets

For the regression model, we test the performance of the methods on two datasets.

On the one hand, data is simulated by sampling from a GP and then perturbated by Gaussian noise. This is the classical GP regression case, for which closed-form solutions are available (Rasmussen & Williams (2005)). To be precise, we are using:

$$\boldsymbol{f} \sim \mathcal{GP}(\boldsymbol{0}, \boldsymbol{K}) \tag{7.9}$$

$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{f}_i, 1), \tag{7.10}$$

where $\boldsymbol{K}$ is based on a squared-exponential kernel with a lengthscale of 0.25 and a variance of 5. The input $\boldsymbol{X}$ was sampled uniformly from the interval between

[1] For the regression example with Gaussian errors, there would be no need for an MCMC approach, however, in order to have a comparable setup over all experiments, we have used MCMC for this problem as well.

[2] We have tested the sensitivity of the result against uniform priors over large intervals containing the true values.

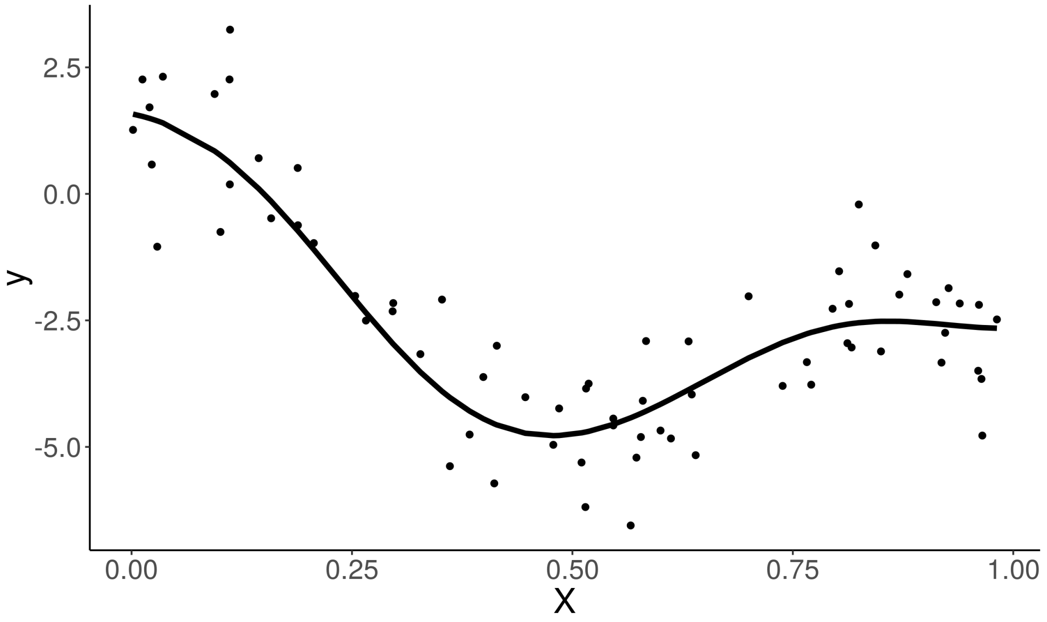

Fig. 7.1: A sample training dataset for the regression experiment with Gaussian error. The line indicates the true function $\boldsymbol{f}$.

0 and 1. Fifty datasets of seventy samples each were generated.

On the other hand, a second batch of datasets is obtained by introducing outliers to the aforementioned datasets with Gaussian noise. These datasets are generated by sampling the number of outliers, r, from the interval 4 to 7 uniformly, then r observations are randomly chosen and multiplied by ten.

Figure 7.1 shows an example dataset with Gaussian noise and Figure 7.2 depicts its perturbation with outliers.

7.2.2 MCMC

In Figure 7.3 we can see the results for the different models when MCMC is used for the example dataset with Gaussian noise presented in the previous section. The first thing to notice is that there is no visible difference in the prediction means for the different GP and TP models. That is, under Gaussian errors, GP and TP give basically the same results. Also the predictive variance is comparable for the GP model and the TP models with five and twenty-five degrees of freedom. Only the model with three degrees of freedom exhibits a wider prediction interval. Moreover, while there are some minor differences in the prediction intervals between the different models, it appears that the prediction intervals are converging to the GP

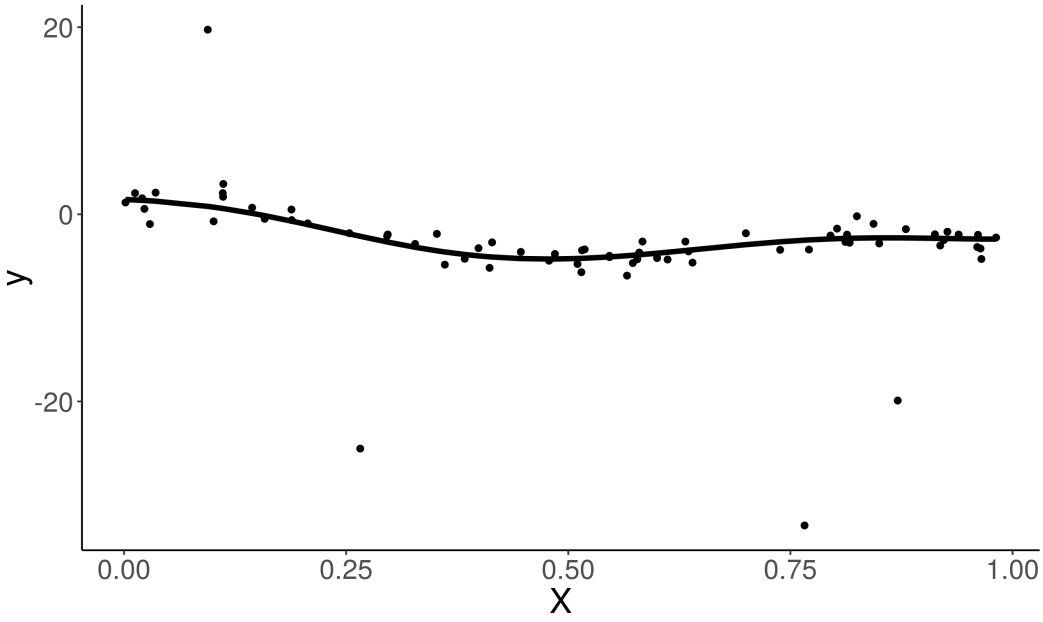

Fig. 7.2: A sample training dataset for the regression experiment with Gaussian observation errors and heavy outliers. The line indicates the true function f.

one, as the degrees of freedom are increased.

The notion that the models are similar for the case of Gaussian noise is also conveyed by Figure 7.12 and 7.13. In these plots we can see the distribution of the kernel parameters that have been obtained from the simulation study. In the left column, we can see the results for the Gaussian noise case. For MCMC, the posterior means of the individual simulations are depicted. For both the variance and the lengthscale parameter, the distributions are comparable. While MCMC seems to be able to capture the lengthscale well, which is seen by the distribution of the results centered around the true value, it appears to produce posterior means for the variance parameter that are too high [3].

In contrast, the introduction of heavy ouliers to the data alters the picture considerably (see Figure 7.4). The GP model disregards most of the curvature of the true f and produces a smooth curve for the prediction means. Additionally, the GP model drastically increases the width of the prediction interval and is now the model with the widest interval. As in the case without outliers, we can see that the TP models are converging to the GP results as the degrees of freedom are increas-

[3] We would like to point out that the log-scale is used for the x-axis of the plot. I.e. the differences between true value and results to the right of it are quite considerable.

ing. However, this time there are still considerable differences, even when the TP model with twenty-five degrees of freedom is used.

For a clearer visualization, we can see in Figure 7.5 a zoomed in version of the plot that focuses on the GP model and the TP model with three degrees of freedom. The TP model captures the true value in its narrow[4] prediction interval quite well. The only exception is around 0.8, where the TP model fails to capture the true value of $\boldsymbol{f}$ in its prediction interval.

The behaviour demonstrated by the models on the example dataset is also supported by our simulation study. Looking at Figure 7.13, we can see that, under outliers, the posterior means for the lengthscale are high compared to the true lengthscale. A consequence of a higher lengthscale is that the posterior mean prediction is more smoothed out. A similar, but not as severe, situation holds for the variance parameter (Figure 7.12), the result distributions for MCMC show stronger focus on higher variance parameters. That is, we would infer a higher uncertainty in the function values, $\boldsymbol{f}$, due to the outliers.

It is noteworthy, that the result distributions for lengthscale and variance are both shifting towards the true value as the degrees of freedom are decreased. A clear indicator that the TP models are more robust to outliers compared to their GP counterparts. However, it is important to state that this performance cannot solely be attributed to the usage of a TP. We would like to emphasize that the GP models are equipped with a Gaussian observation likelihood, while the TP models use a Student-t likelihood. Past research has shown that GP models with Student-t likelihoods can perform well under outliers (e.g. Neal (1997), Jylänki et al. (2011), Vanhatalo et al. (2009)).

Finally, the difference in the observation likelihoods might also be the reason, why we do not see a strong increase in the width of the prediction intervals for the TP models (especially the one with three degrees of freedom). Based on Section 2.2, we would have expected a higher prediction variance under outliers and therefore a wider prediction interval. However, we argue that this effect is mitigated by the Student-t observation likelihood, which can also compensate for outliers partially O'Hagan (1979).

[4] compared to the GP model

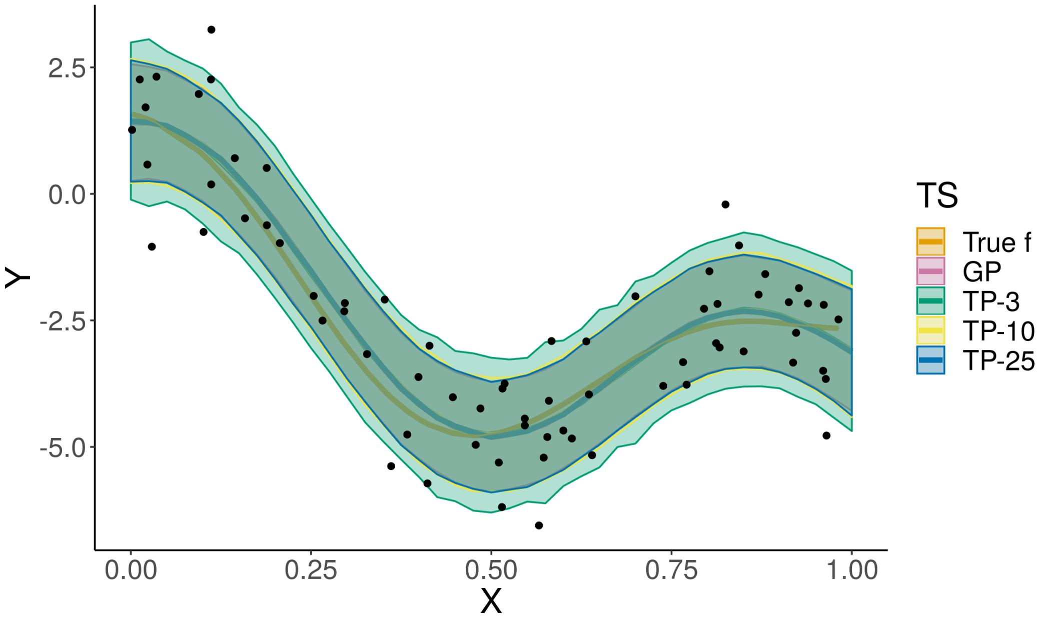

Fig. 7.3: Prediction mean and ± one standard deviation interval for $\boldsymbol{f}$ in the regression problem with Gaussian noise based on MCMC. TP results for 3, 5, and 25 degrees of freedom are presented.

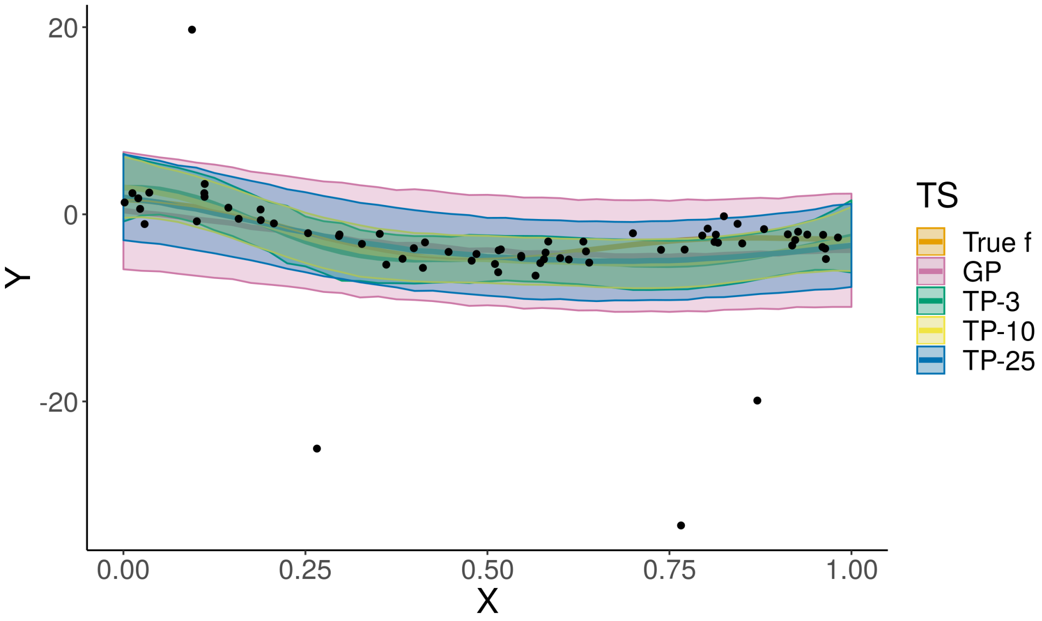

Fig. 7.4: Prediction mean and ± one standard deviation interval for $\boldsymbol{f}$ in the regression problem with Gaussian noise and outliers based on MCMC. TP results for 3, 5, and 25 degrees of freedom are presented.

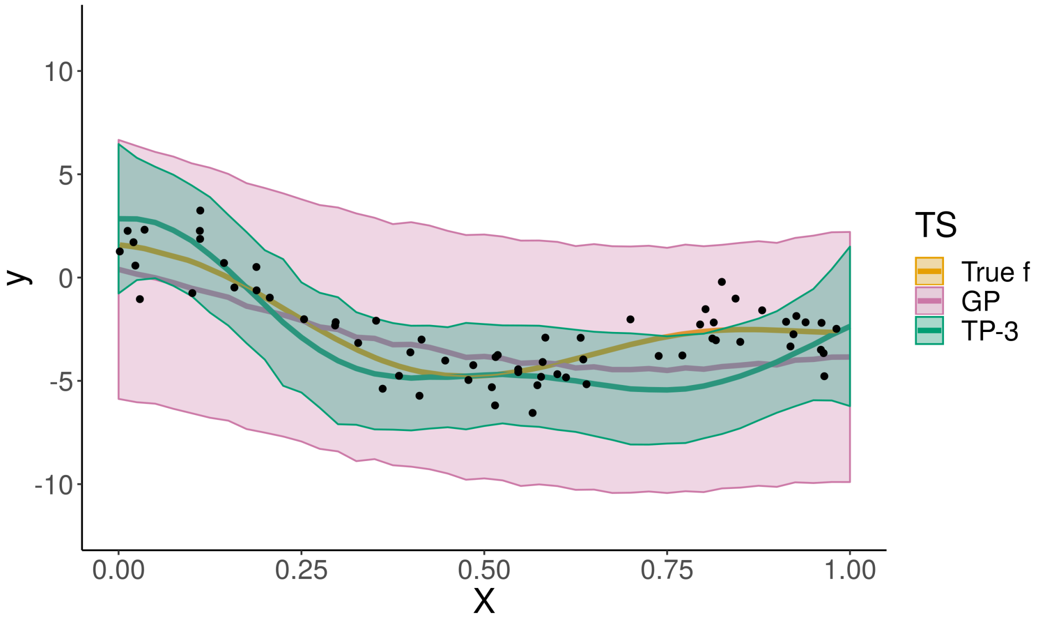

Fig. 7.5: Zoomed in on prediction mean and ± one standard deviation interval for $\boldsymbol{f}$ in the regression problem with Gaussian noise and outliers based on MCMC. Only the results for 3 degrees of freedom are presented.

7.2.3 t-Laplace

Figure 7.6 shows the results for the different models applied to our example dataset with Gaussian noise. Immediately we can see that the t-Laplace approximation has not converged to a sensible result. Both the lengthscale and the variance parameter of the squared-exponential kernel have collapsed to values close to 0, which leads to a straight line centered at 0 as the prediction mean. We have tested whether this result for the t-Laplace approximation is due to bad local minima. However, the suboptimal behaviour is persistent, that is, the outcome for different initial values were the same. Even initiating the t-Laplace approximation with the results of the GP-Laplace approximation did not lead to a more promising result. Unsurprisingly, the problem also occurs in the example with outliers, as well as in the simulation study. Therefore we have removed the results for the t-Laplace approximation from Figures 7.12 and Figure 7.13, in order to make the other results easier to read.

In contrast to the unpleasant outcome of the t-Laplace approximation, the normal Laplace approximation gives results that are comparable to MCMC for the example dataset with Gaussian noise. This is not surprising, as the Laplace approximation approximates the posterior distribution of $\boldsymbol{f}$ with a Gaussian distribution

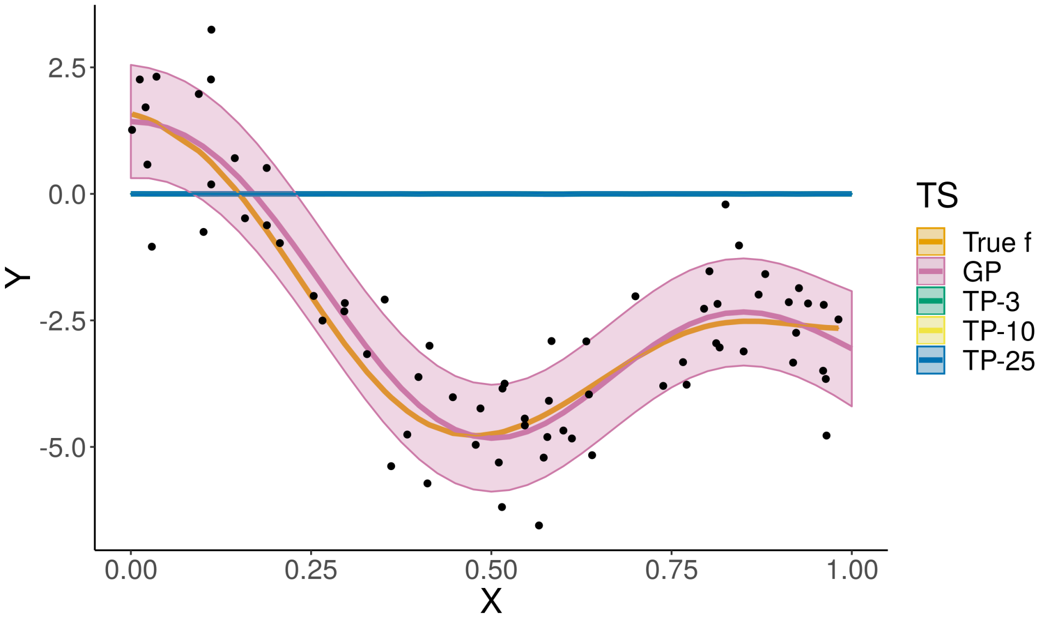

Fig. 7.6: Prediction mean and ± one standard deviation interval for f in the regression problem with Gaussian noise based on Laplace and t-Laplace approximation. TP results for 3, 5, and 25 degrees of freedom are presented.

centered at the posterior mode. For the ordinary regression case with Gaussian noise, the posterior distribution of f is Gaussian, therefore the Laplace approximation can approximate the posterior perfectly (Rasmussen & Williams (2005)). Interestingly, for the case with outliers (Figure 7.7), the prediction means of the Laplace approximation fit the true values better than the result obtained from MCMC. However, the general behaviour of a strong increase in the width of the prediction interval for the GP model persists also for the Laplace approximation.

Fascinatingly, the Laplace approximation captures the true values for the variance and lengthscale in the simulation study better than MCMC (see Figure 7.13). Especially the results for the lengthscale parameter are interesting. Even when outliers are present in the data, the result distribution for the Laplace approximation is centered around the true value and does not exhibit a strong shift to the right like the other methods.

7.2.4 Variational Student-t Approximation

First and foremost, we do not present the results for VTP1 and VTP2 separately, as the outcomes of these methods are visually indistinguishable from each other.

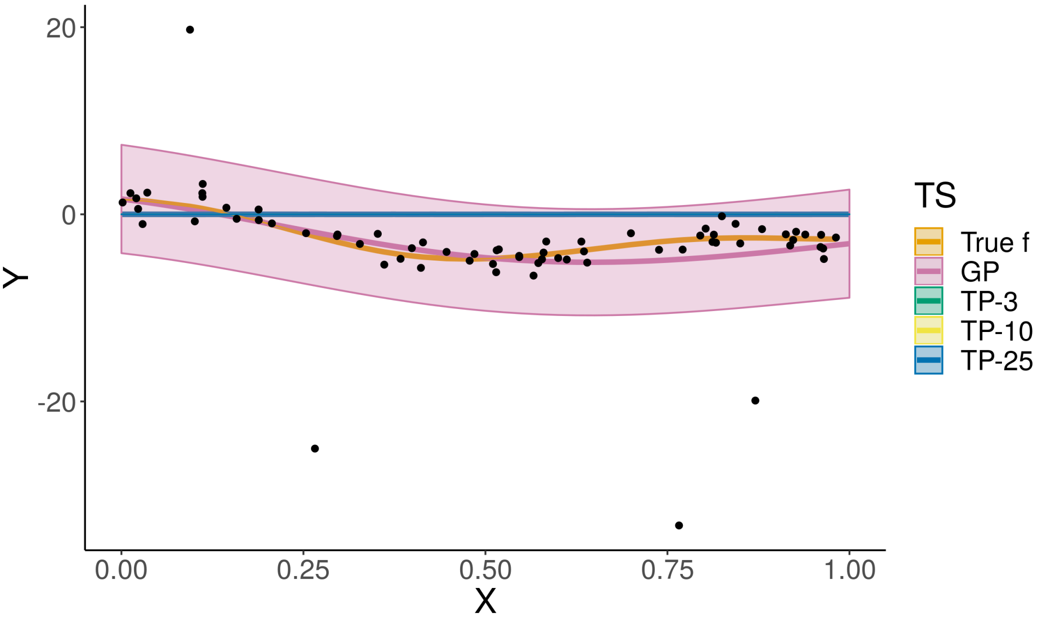

Fig. 7.7: Prediction mean and ± one standard deviation interval for f in the regression problem with Gaussian noise and outliers based on Laplace and t-Laplace approximation. TP results for 3, 5, and 25 degrees of freedom are presented.

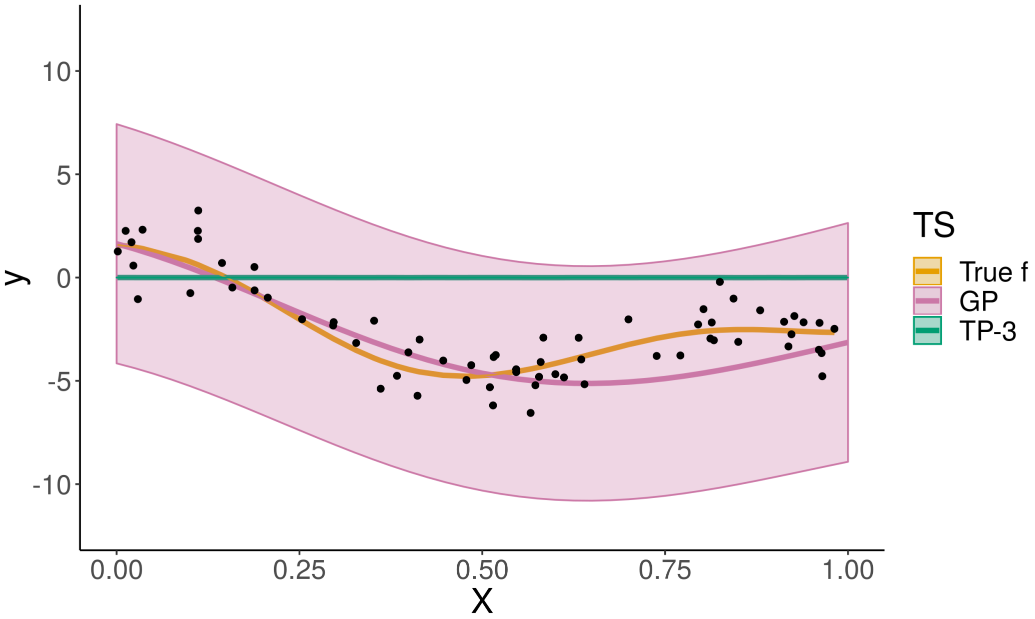

Fig. 7.8: Zoomed in on prediction mean and ± one standard deviation interval for f in the regression problem with Gaussian noise and outliers based on Laplace and t-Laplace approximation. Only the results for 3 degrees of freedom are presented.

E.g. the result distributions for the simulation study (Figure 7.12 and Figure 7.13) are overlapping almost perfectly. Consequently, what can be said about one of the methods, equally applies to the other. Therefore, we are analysing the visual outcome of the two methods together.

In Figure 7.9 we can see the results obtained from VGP and VTP1/2 for the example dataset with Gaussian noise. Similarly to the Laplace approximation, the VGP method can approximate the posterior $p(\boldsymbol{f}|\boldsymbol{y})$ perfectly. In contrast to the MCMC results, the outcomes of the variational methods are marginally more erratic, e.g. the prediction intervals are less smooth. Moreover, while for the MCMC methods, the prediction means were almost identical, there is slightly more variation between the different outcomes. Nevertheless, the pattern that the results of the TP models converge to the results of their GP counterparts persists also under the variational methods.

When it comes to outliers (Figure 7.10 and Figure 7.11), the results of the variational methods are similar to the outcomes obtained from MCMC. On the one hand, the prediction means of the GP model are smoothed out, while the prediction interval increases considerably in width. On the other hand, the TP based models have a narrower prediction interval, which successfully captures the true function values $\boldsymbol{f}$.

For this particular example, VGP and VTP1/2 appear to give good approximations of hte MCMC results, at least based on the prediction means and intervals. However, if we consider the results of the simulation study, we see a more substantial disagreement between the variational methods and MCMC. For VTP1/2, the result distribution for the lengthscale parameter benefits from a lower number of degrees of freedom. Comparably to the MCMC case, the distribution shifts closer to the true value, as the degrees of freedom decrease. However, this does not seem to hold for the variance parameter. Here we cannot see any changes due to a different number of degrees of freedom. Surprisingly, none of the variational methods match the performance of the Laplace approximation in the simulation study, especially for the lengthscale under outliers[5].

[5] This raises the question whether the Laplace approximation is that good, or whether there are some specifics in the GPy (GPy (since 2012)) implementation of the Laplace approximation that makes it more resilient. However, due to the direct competitor of the Laplace approximation, the t-Laplace approximation, not performing at all, we decided against a deeper investigation of the

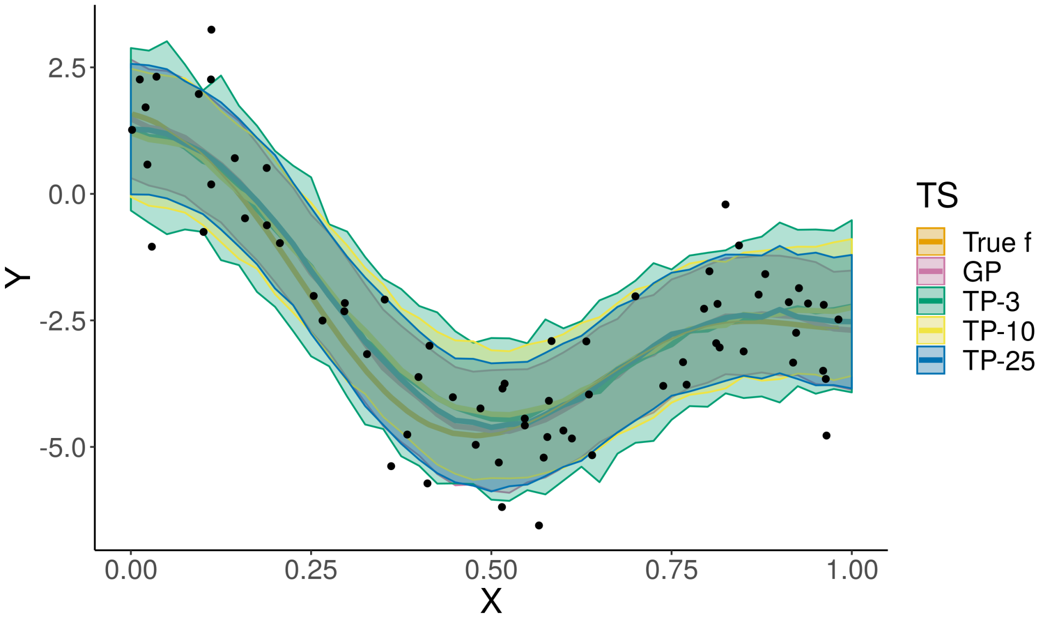

Fig. 7.9: Prediction mean and ± one standard deviation interval for f in the regression problem with Gaussian noise based on VGP and VTP1/2 approximation. TP results for 3, 5, and 25 degrees of freedom are presented.

While the performance of the Student-t approaches on these examples are competitive with their Gaussian counterparts, there is a significant drawback that is not reflected in the figures. Where the variational Gaussian approximation was stable with respect to initial values and optimization method used, the Student-t methods had severe problems to converge to reasonable parameter settings. The results presented here are the best outcomes among several runs. It was not atypical for the two Student-t based approaches to suffer under diverging kernel parameters. For several runs, the variance parameter was blowing up to unreasonable large numbers, while the lengthscale parameter was shrinking to a value close to 0.

7.3 Simulated Data - Classification

7.3.1 Datasets

In order to get an idea of the behaviour of the methods for classification, two setups with simulated data have been investigated. On the one hand, we have a setup with

performance of the Laplace approximation.

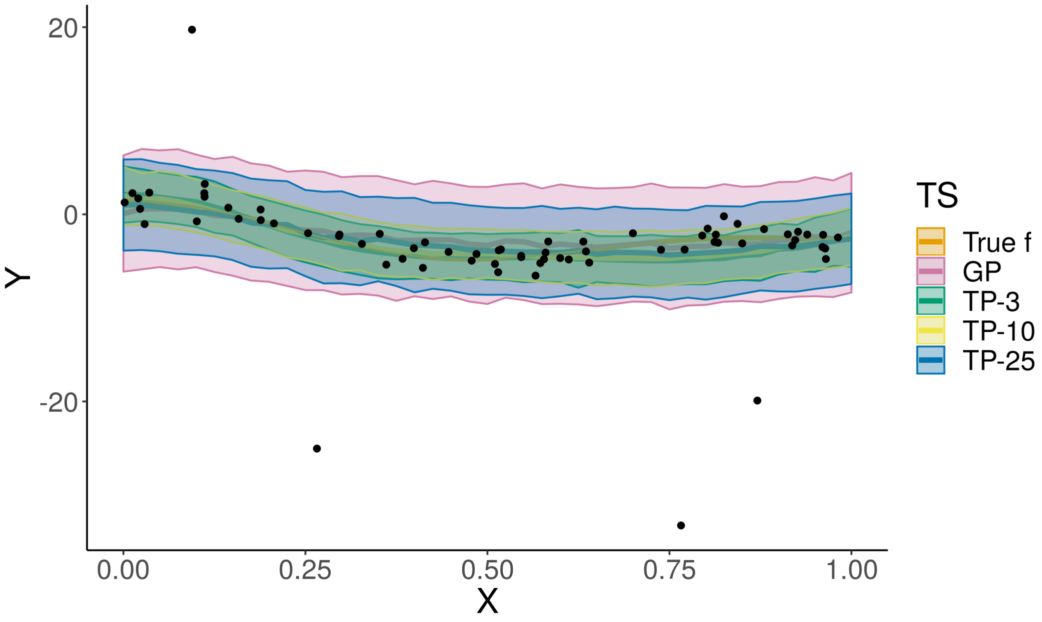

Fig. 7.10: Prediction mean and ± one standard deviation interval for f in the regression problem with Gaussian noise and outliers based on VGP and VTP1/2 approximation approximation. TP results for 3, 5, and 25 degrees of freedom are presented.

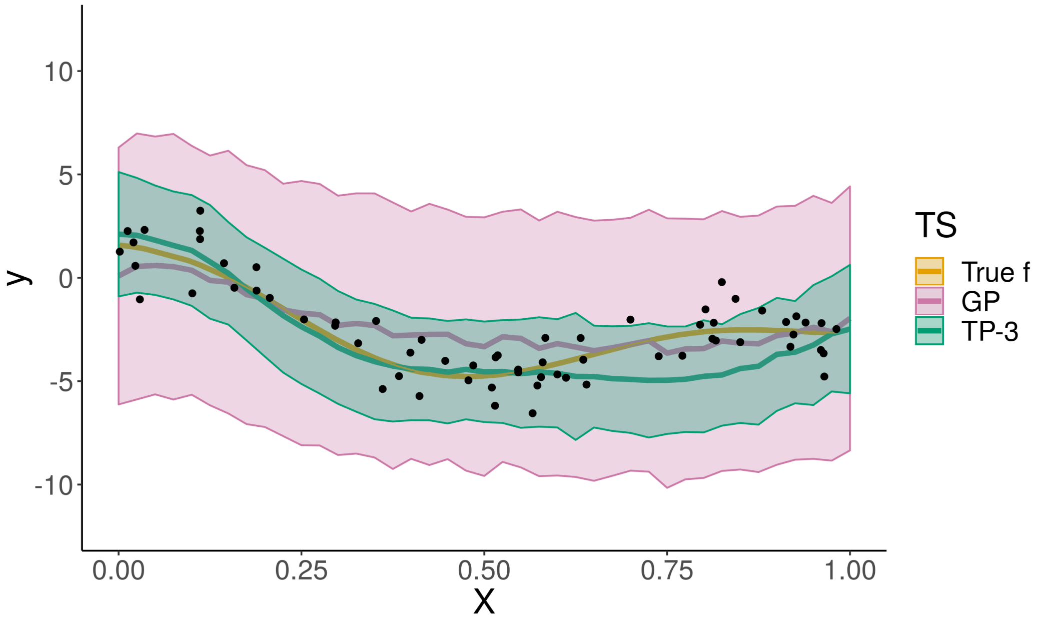

Fig. 7.11: Zoomed in on prediction mean and ± one standard deviation interval for f in the regression problem with Gaussian noise and outliers based on VGP and VTP1/2 approximation approximation. Only the results for 3 degrees of freedom are presented.

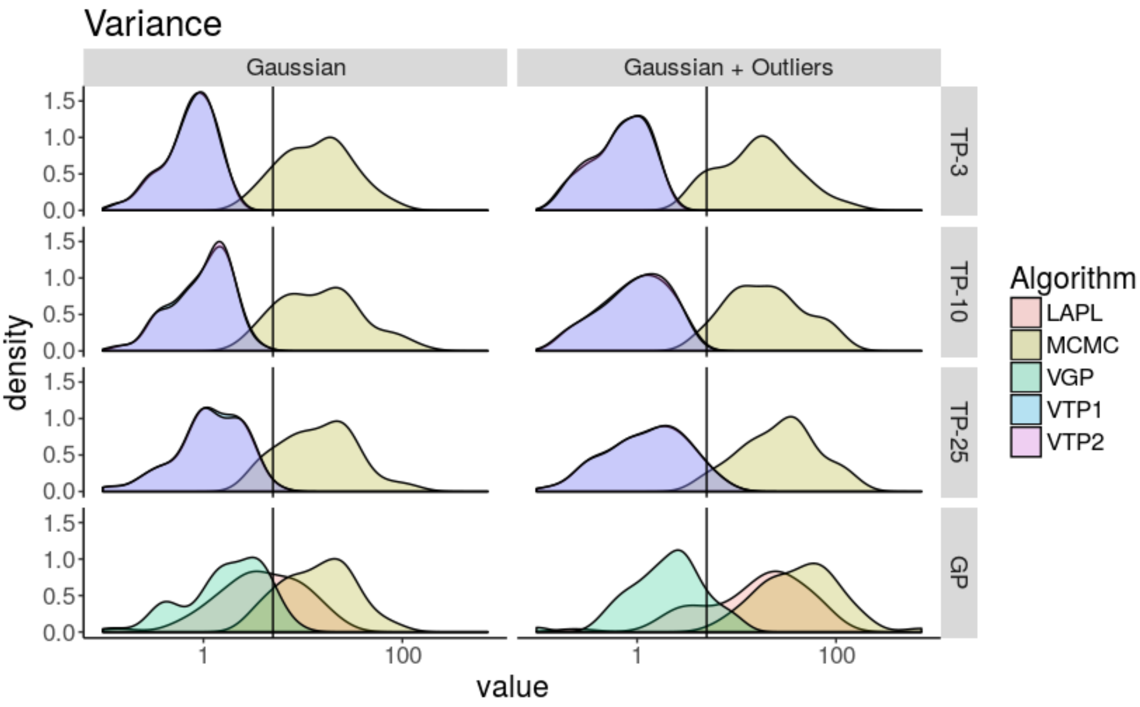

Fig. 7.12: Comparison of the results for the variance parameter for the different models from the regression simulation study. The vertical black line indicates the true value of the variance parameter which has been used to simulate data. None of the methods clearly outperforms the other methods in terms of capturing the true value of the variance. Values for the t-Laplace approximation are missing because the method diverges consistently. Non-converging runs of one of the other methods have also been removed.

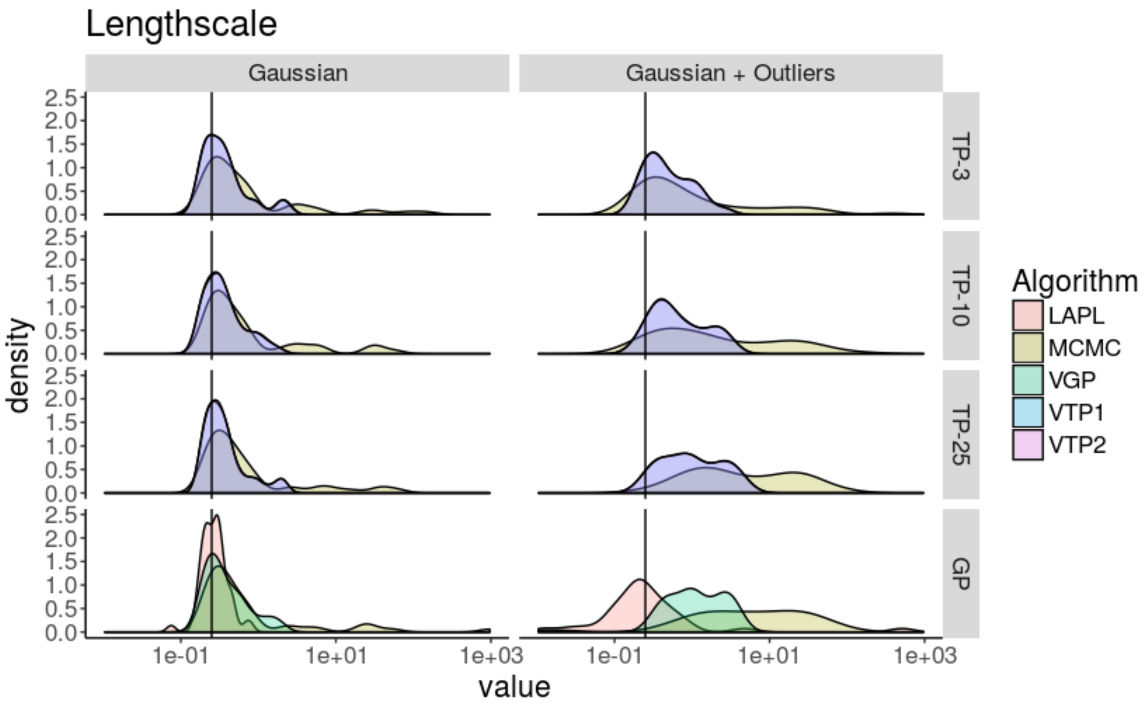

Fig. 7.13: Comparison of the results for the lengthscale parameter for the different models from the regression simulation study. The vertical black line indicates the true value of the variance parameter which has been used to simulate data.The Laplace approximation captures the true value the best over the two scenarios. Interestingly, the TP models are more robust towards outliers, as the degrees of freedom are reduced. Values for the t-Laplace approximation are missing because the method diverges consistently. Non-converging runs of one of the other methods have also been removed.

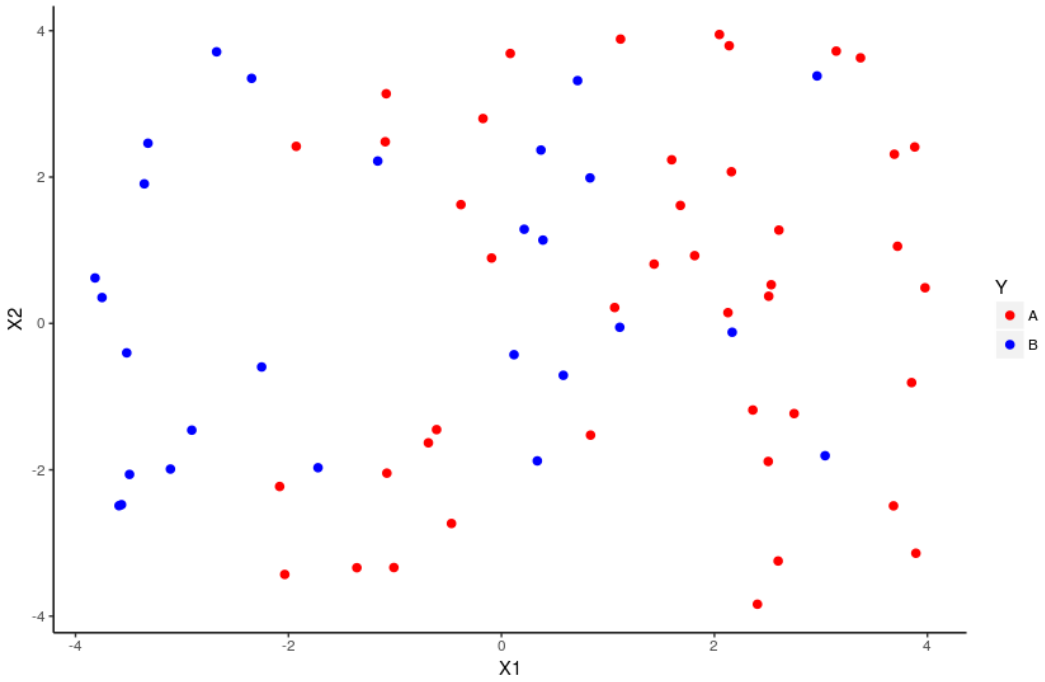

Fig. 7.14: A sample training dataset for the classification experiment without mislabelled data.

simulated data, where the two classes are created by the following procedure:

$$\boldsymbol{f} \sim \mathcal{GP}(\boldsymbol{0}, \boldsymbol{K}) \tag{7.11}$$

$$\boldsymbol{y}_i \sim \text{Ber}(\text{sigm}(\boldsymbol{f}_i)), \tag{7.12}$$

where $\boldsymbol{K}$ is based on a squared-exponential kernel with a lengthscale of 0.25 and a variance of 5. The input $\boldsymbol{X}$ was sampled uniformly from the interval between 0 and 1. Fifty datasets of seventy samples each were generated.

On the other hand, we have the same datasets, but some of the observations are mislabelled. Similar to the simulation study for the regression case, these datasets are generated by sampling the number of outliers, r, from the interval 4 to 7 uniformly, then r observations are randomly chosen and their label is switched.

Figure 7.14 shows an example dataset without mislabelled observations and Figure 7.15 depicts the same dataset with five mislabelled observations.

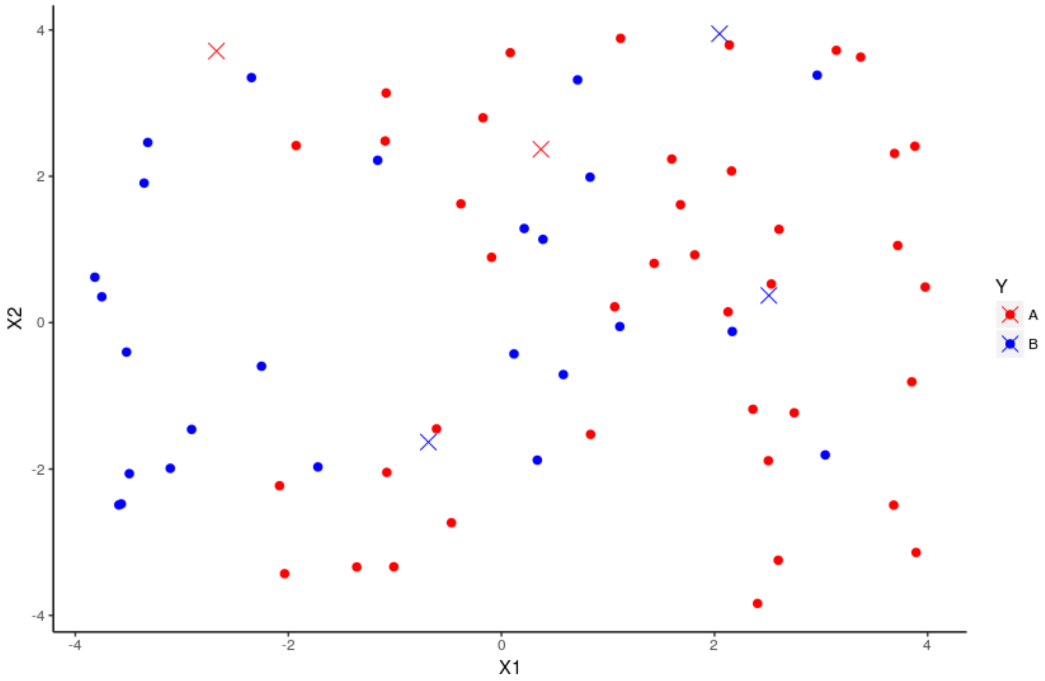

Fig. 7.15: A sample training dataset for the classification experiment with mislabelled observations. The crosses indicated the observations that have been mislabelled.

7.3.2 MCMC

Figure 7.16 presents the posterior predictive probabilities for the simulated data sets. In the left column, we can see the probabilities for the clean case, that is, the unaltered one, while in the right column we can see the impact of the mislabelled observations.

For the case without outliers, we can see a clear difference in the shape of the decision boundary, that is, the boundary that separates the two classes (indicated by the white area between the red and blue areas) (Bishop (2006)). While the TP models promote a smooth boundary, the GP model suggests a decision boundary that extends far into the area that the TP models would assign to class A, the red one. This is due to the GP model favoring a lower posterior mean for the lengthscale compared to the TP models. The lower the lengthscale, the more the high probability areas for a certain class are concentrated around the observations from that class.

When outliers are introduced, the resulting decision boundary and posterior predictive probabilities for the GP model look more like the ones from the TP models.

In contrast, the outcomes of the TP models are rather stable towards the outliers. While there is a shift from the posterior predictive probabilities towards 50%, due to the outliers, the decision boundaries as such, did not move extensively. However, there does not seem to be a relationship between the robustness of the decision boundary and the degrees of freedom. The model with twenty-five degrees of freedom seems to be the least affected, while the model with three degrees of freedom appears to have changed the most.

As pointed out earlier in the Background Chapter (2.2), one of the interesting aspects of the TP is its posterior predictive variance. However, as Figure 7.17 shows, in a binary classification problem, outliers or mislabelled observations do not necessarily lead to an increase in the posterior predictive variance for the function values $\boldsymbol{f}_*$. While the TP has a higher variance than the GP for the case without mislabelled data, the existence of mislabelled observations leads to a substantial drop in posterior predictive variance. The behaviour can be explained with the long and flat tails of the sigmoid function. In the scenario without outliers, the model wants to assign very high or low posterior predictive probabilities. Due to the s-shape of the sigmoid function with its relatively flat tails, the sigmoid maps function values from a half-unbounded interval onto these high/low probability region. This leads to a high variance in the function values. In contrast, for the perturbed scenario, the model wants to assign posterior probabilities that are closer to 50%. In this regions, the sigmoid function is relatively steep, therefore only a comparatively narrow interval of function values are mapping to these probability regions. Consequently, the variance of the function values $\boldsymbol{f}_*$ is low.

While our example dataset gives some empirical evidence that using a TP can be beneficial in terms of robustness, our simulation study does not show a significant advantage of the TP models when it comes to classification. In Figure 7.20 and Figure 7.21 the results of the simulation study for the variance and lengthscale parameter are presented. None of the distributions seem to be affected strongly by the introduction of outliers. There are some smaller shifts, but, besides of those, the distributions are rather stable. This also supports our conclusion from the example dataset, that the degrees of freedom seem to have no visible impact with respect to robustness, at least within our experimental setup.

All in all, these results are disappointing, however, it is important to consider that the example from Futami et al. (2017) is based on their expectation propagation algorithm. That is, while there might not be a considerable gain from using a TP for a binary classification problem which is solved via MCMC, using Student-t distributions as approximate posterior distributions could still hold substantial value. This will be investigated in the coming sections.

7.3.3 t-Laplace

Figure 7.18 shows the posterior predictive probabilities for the simulated data. We have trained the t-Laplace approximation with three, ten, and twenty-five degrees of freedom. For comparison, the result for the Laplace approximation in the GP case is depicted in the last row.

The most apparent feature of this figure is that the t-Laplace approximation has basically learnt nothing about the two classes. While the ordinary GP-Laplace approximation has resulted in a smooth decision boundary for the case without mislabelled data, the t-Laplace approximation has converged to solutions that predict around 50% class membership for either class. The root cause for these poor results is the same as in the regression case. Neither lengthscale, nor variance have converged to a reasonable value for the t-Laplace approximation. As with the regression case, we have tested different approaches to make the t-Laplace approximation converge to sensible parameter values. However, our attempts were futile and we did not manage to improve the convergence properties of the t-Laplace approximation.

While there is no question about the abysmal quality of the results produced by the t-Laplace approximation, the outcome of the ordinary Laplace approximation is still of interest. For the case without mislabelled observations, the decision boundary produced by the Laplace approximation differs considerably from the one produced by MCMC. This can be seen as an indicator that a Gaussian centered at the posterior mode is not capturing the true shape of the posterior distribution $p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{y})$ adequately. Especially in higher dimensions, this is a well established downside of the Laplace approximation (Nickisch & Rasmussen (2008))[6].

[6] Nevertheless, we would like to stress again, that the models that are approximated by MCMC and Laplace approximation are slightly different. That is, some of the differences could also be due to this marginal difference in model definition.

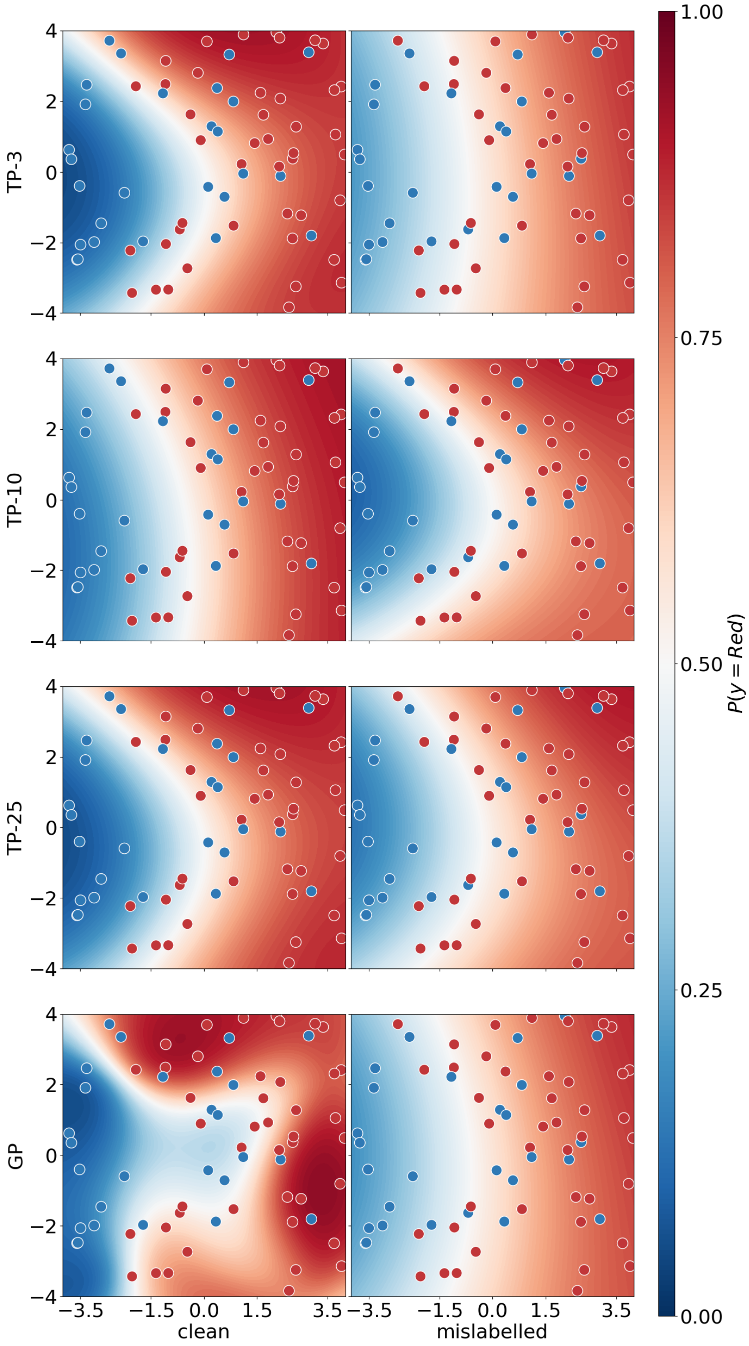

Fig. 7.16: Comparison of posterior predictive probabilities for MCMC for the classification problem. TP results for 3, 5, and 25 degrees of freedom are presented.

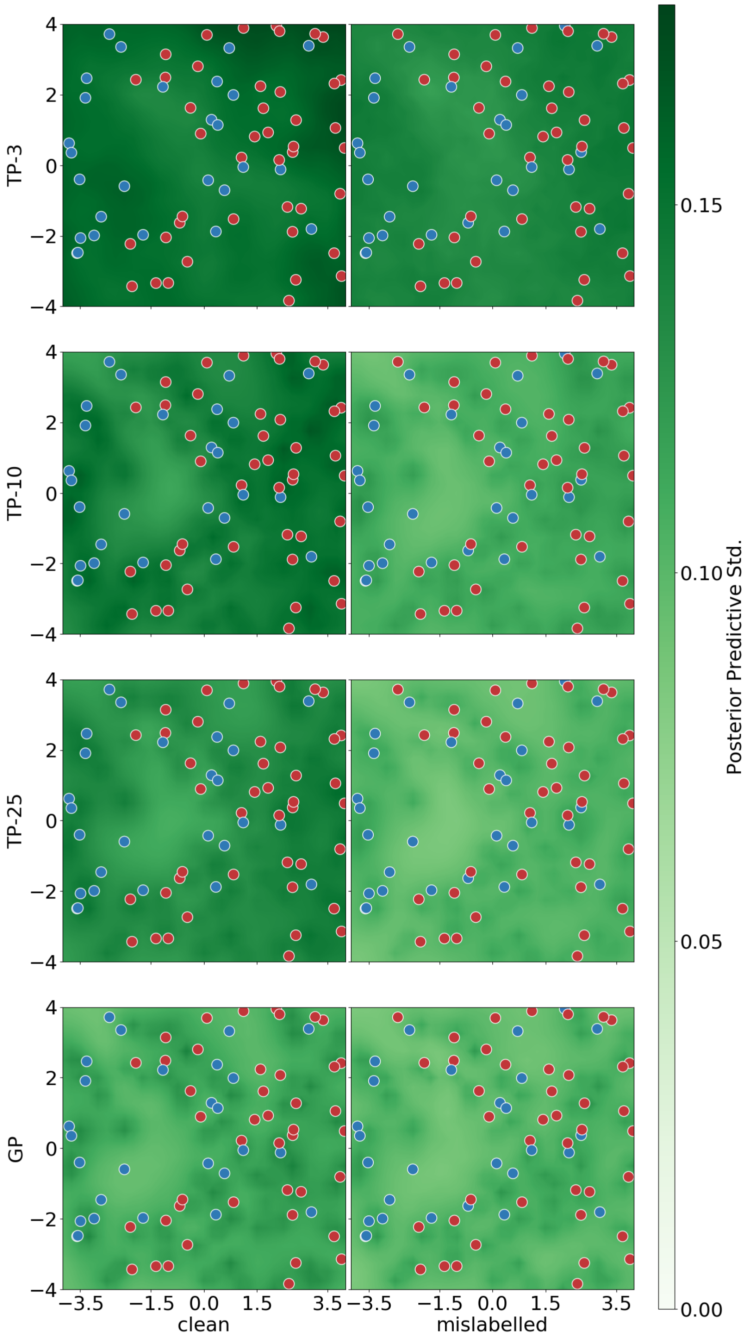

Fig. 7.17: Comparison of posterior predictive standard deviation for MCMC for the classification problem. TP results for 3, 5, and 25 degrees of freedom are presented.

Moreover, we can see that the introduction of the mislabelled observations led to a drastic deterioration of the decision boundary. Similar to the t-Laplace approximation, the Laplace approximation, in this specific case, did not converge to reasonable parameter values. This shows that there is some potential for robust methods to improve on the Laplace approximation.

Finally, taking into consideration the results of the simulation study (Figure 7.20 and Figure 7.21), we can see that the Laplace approximation usually converged to variance parameters that are considerably below the the true value, while the lengthscale was often to be found higher than the actual one. That is, in our experiments, the Laplace approximation had the tendency to underestimate the variance of $\boldsymbol{f}$ and overestimate the impact of individual observations on classifying other observations further away. Similarly to the MCMC case, the introduction of mislabelled observations does not seem to have a strong impact on the result distributions. However, we would like to point out that the Laplace approximation diverged for some of the examples and we have removed these examples from the image. Moreover, the right tail of the result distribution for the lengthscale is more pronounced, indicating that for a larger number of examples, the Laplace approximation led to a smaller lengthscale, which means that the influence of individual observations is more limited to their close neighborhood in the presence of mislabelled data.

7.3.4 Variational Student-t Approximations

Similarly to the regression case, we do not present the results for VTP1 and VTP2 separately, as the outcomes of these methods are again visually indistinguishable from each other (see Figure 7.20 and Figure 7.21). Therefore, we are analysing the visual outcome of the two methods together.

In Figure 7.19 we can see the results for the VGP and the VTP1/2 methods. On the one hand, for the clean data, there does not seem to be any visual difference between the different methods. The decision boundaries appear to be the same. On the other hand, the variational Gaussian approximation is comparatively strongly affected by the mislabelled observations. However, with regards to the outcomes of the experiments for the ordinary regression with outliers, the impact on the GP model is not that severe. Interestingly, similary to the Laplace approximation, the

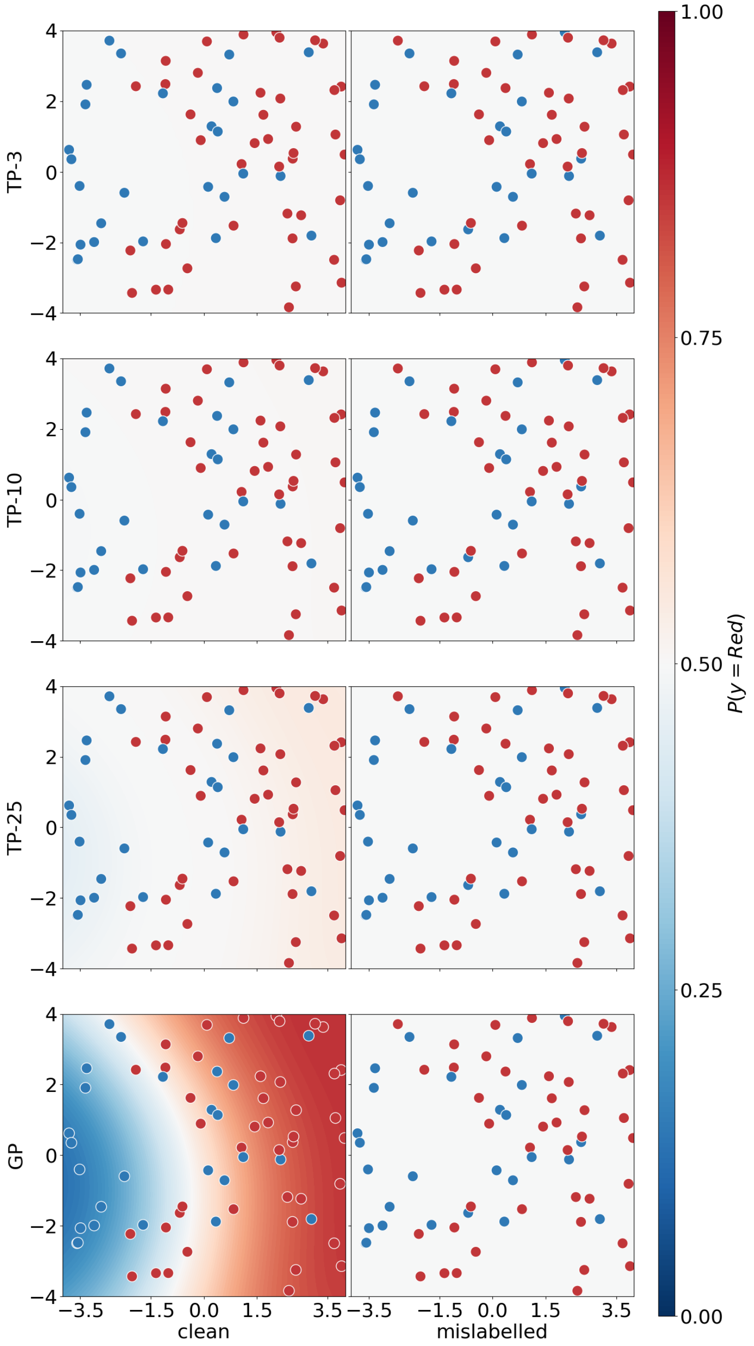

Fig. 7.18: Comparison of posterior predictive probabilities for Laplace and t-Laplace approximation for the classification problem. TP results for 3, 5, and 25 degrees of freedom are presented.

Gaussian approximation cannot recover the decision boundary obtained by MCMC for the clean case.

With respect to the simulation study, the variational methods capture the underlying true lengthscale better than MCMC or Laplace approximation. This holds for the simulated data with and without mislabelled data. In contrast, VGP and VTP1/2 tend to underestimate the variance parameter considerably. Consequently, models based on these parameter values would overestimate the impact individual observations have on distant observations, while providing a too small band in which the function values $\boldsymbol{f}$ are to be expected. Comparing the result distributions, there does not seem to be much of a difference between VGP and VTP1/2, regardless of whether mislabelled data is present or not [7]. That is, the simulation does not provide strong evidence that TPs have substantial benefits over their GP counterparts when it comes to binary classification problem, at least when the kernel parameters are taken into consideration.

However, while the inferred kernel parameters do not seem to differ much between TP and GP models, there is an interesting characteristic of the posterior predictive probabilities (Figure 7.19) obtained from VTP1/2, which is not observed for MCMC. The variational Student-t approximations are assigning less posterior probability mass than their Gaussian counterpart on average. That is, the probability parameter of the Bernoulli observation model is more regularized towards 50% than in the GP case. This is especially pronounced for the perturbed data set, but the effect also persists in the clean data. This can be explained with a reduced need for the variational Student-t distribution to deviate from the prior (which is centered at 0) to accommodate observed data, especially outliers, in the observation space, due to the heavy-tails of the Student-t distribution. That is, the Student-t distribution that is used in VTP1/2 to approximate the posterior of $\boldsymbol{f}$ does not need to shift as much as the Gaussian that is used for VGP.

[7] This behaviour is observed over all the different methods. The introduction of outliers does not seem to affect the estimates for the kernel parameters considerably. This might not be a general characteristic, but more of an artifact of our experimental setup. The setup that we have used to generate the data does not create easily separable classification examples. That is, even for the clean case, it is not uncommon to find individual observations that belong to a different class than other classes in close proximity (e.g. our example dataset). Obviously, introducing to such datasets a small number of mislabelled observations, cannot have an extreme impact, as even for the clean dataset, the model parameters need to reflect that the data is not composed of monolithic, easily separable clusters of different classes

While the performance of the Student-t approaches on these examples are competitive with their Gaussian counterparts, there is a significant drawback that is not reflected in the figures. Where the variational Gaussian approximation was stable with respect to initial values and optimization method used, the Student-t methods had severe problems to converge to reasonable parameter settings. The results presented here are the best outcomes among several runs. It was not atypical for the two Student-t based approaches to suffer under diverging kernel parameters. For several runs, the variance parameter was blowing up to unreasonable large numbers, while the lengthscale parameter was shrinking to a value close to 0. This constellation led to posterior predictive probabilities that are close to 50% almost everywhere.

7.4 Real Data

7.4.1 Datasets

For testing the methods on real data, the Penn Machine Learning Benchmarks (PMLB) were used (Olson et al. (2017)). The dataset contains 94 binary classification problems with up to 49000 observations and up to 1000 features. The problems were split into two groups, problems with less than 500 observations were used to test the variational Student-t approximations and the t-Laplace approximation, whereas the other problems were used to evaluate the variational sparse inducing point methods.

In both cases, each problem dataset was split into a training set (60%) and a test set (40%). The performance of the models were evaluated on the test set in terms of accuracy and the average log-loss per observation, where log-loss is given by Bishop (2006)

$$\text{log-loss} = -(y_{observed} \log(p(y_{predicted})) + (1 - y_{observed}) \log(1 - p(y_{predicted}))).$$

Where $p(y_{predicted})$ indicates the probability of y having the label one assigned by the model. The log-loss has been chosen due to its more fine-grained view on the prediction performance of an algorithm. In contrast to accuracy, log loss takes into account the confidence of the model with respect to its predictions. That is, if a model assigns a high probability to a class, but this prediction is wrong, then a high

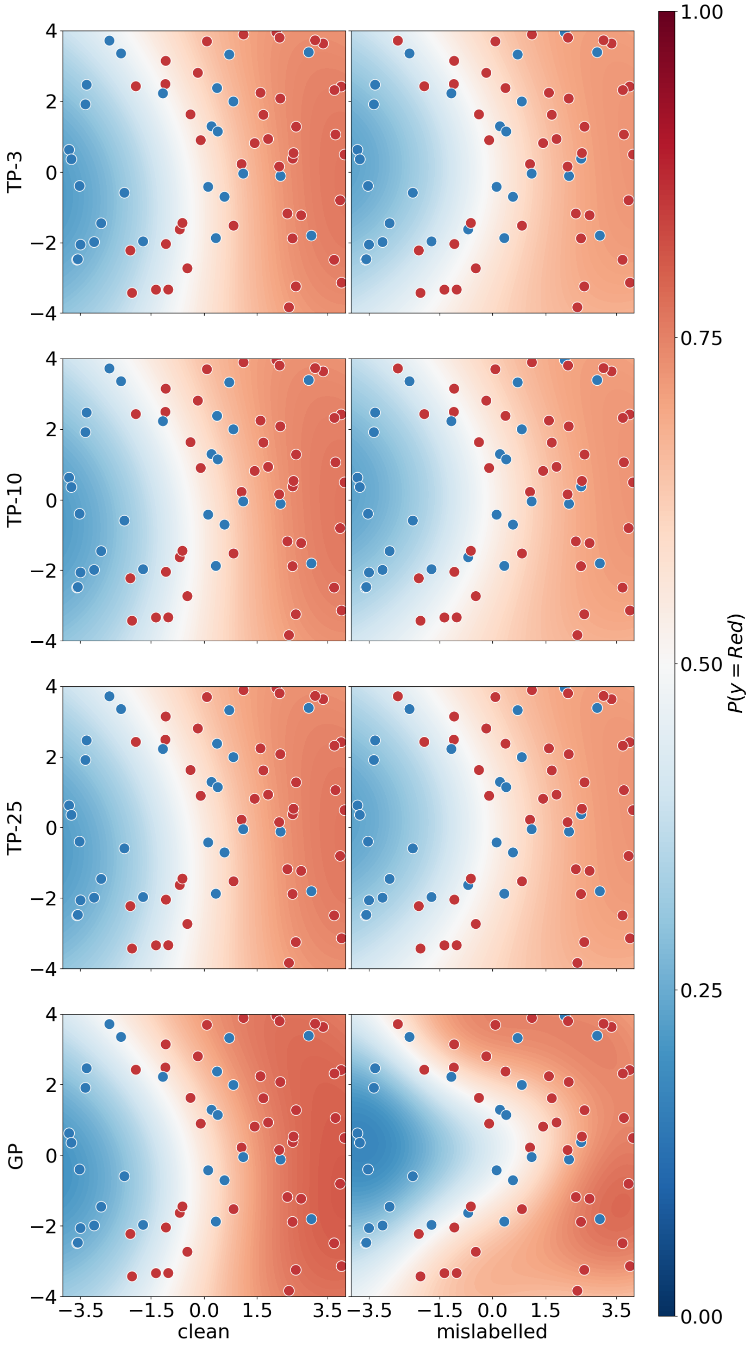

Fig. 7.19: Comparison of posterior predictive probabilities for VGP and VTP1/2 approximation2 for the classification problem. TP results for 3, 5, and 25 degrees of freedom are presented.

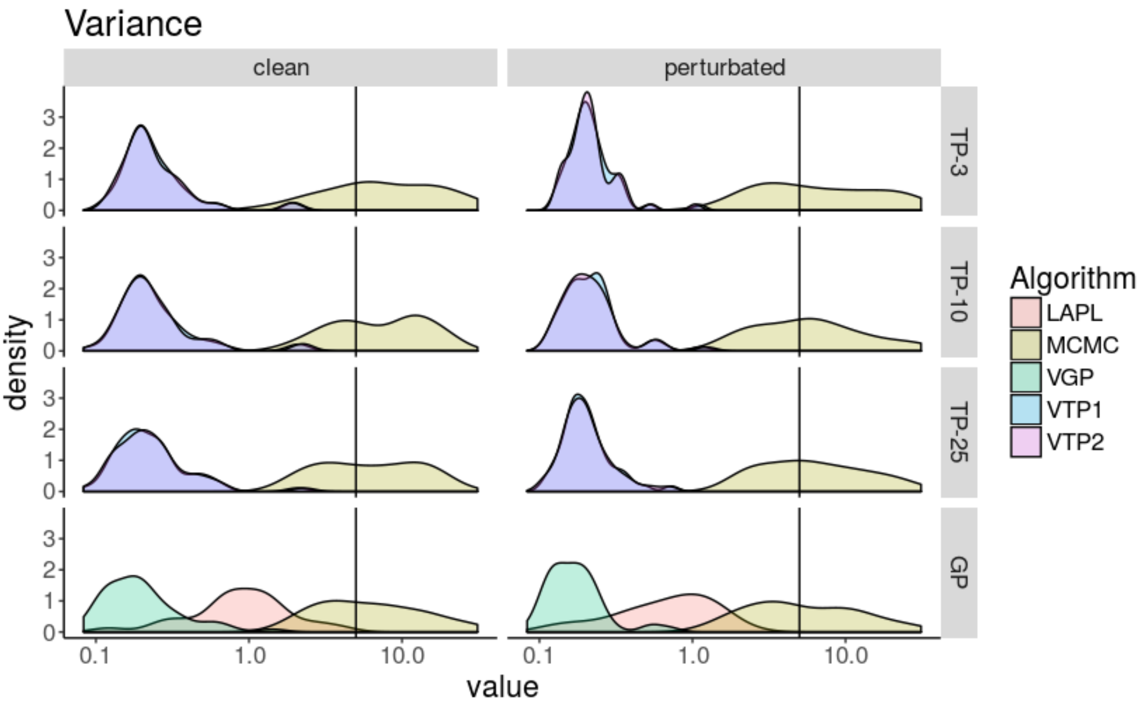

Fig. 7.20: Comparison of the results for the variance parameter for the different models from the classification simulation study. The variational methods result in similar values for the variance parameter. The vertical line indicates the true value that has been used to simulate the datasets. Only MCMC captures the true value adequately. Values for the t-Laplace approximation are missing because the method diverges consistently. Non-converging runs of one of the other methods have also been removed.

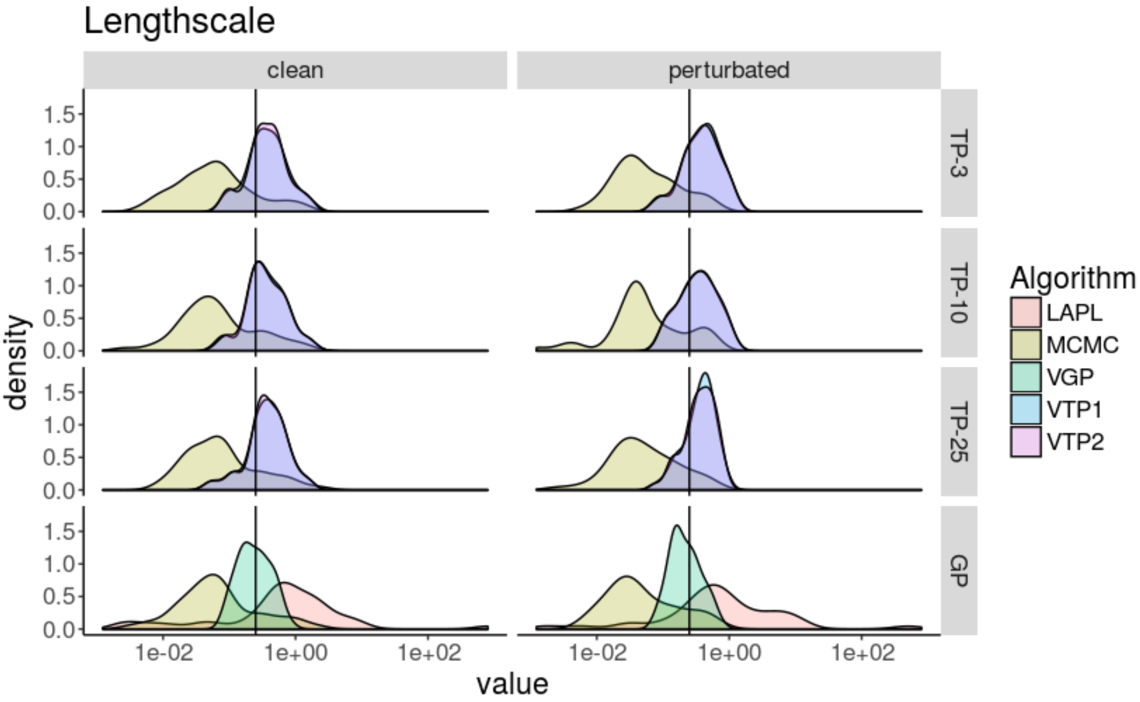

Fig. 7.21: Comparison of the results for the lengthscale parameter for the different models from the classification simulation study. Comparably to the variance parameter case, the variational methods result in similar values. Moreover, the variational methods capture the true value (indicated by the vertical line) the best. Values for the t-Laplace approximation are missing because the method diverges consistently. Non-converging runs of one of the other methods have also been removed.

loss is incurred.

In addition of using the average log loss as a performance indicator, we have also used it as a way to assess whether an algorithm has converged to a sensible parameter setting or not. An average log loss of close to 0.69 is comparable to a classifier that outputs class probability of 50% every time ($\log 0.5 \approx -0.69$), i.e. the classifier flips a coin for assigning a class. Datasets, for which none of the methods was able to achieve an average log loss of less than 0.69, are not presented in this comparison.

7.4.2 Small Data

Table 7.1 depicts the summary statistics of the results for each method on the small datasets in terms of average log loss and accuracy. The results for each individual data set can be found in Appendix C.1. The degrees of freedom used for the Student-t based methods have been determined via cross-validation on the training data set. In contrast to the simulated data sets, we have not tuned any of the methods with respect to the initial values to give a more clear idea about the out-of-box performance of these methods. Importantly, this would put the variational Student-t methods at a disadvantage, as they are suffering from poor convergence properties. However, this is mitigated by the cross-validation which frequently tends to find degrees of freedom for which the method converges to reasonable parameter settings. Noteworthy, for 11 out of the 45 data sets with less than 500 observations none of the methods converged reasonably with regards to our log loss convergence criterion.

Overall, the GP-Laplace approximation is the method with the smallest mean log loss per observation as well as the one with the highest accuracy on the data sets with less than 500 observations. However, if we look at the number of times the method was the best among all the different methods, we can see that in terms of average log loss, the variational Gaussian and the Student-t approximation (VTP1) are more successful, while the two methods are close to the GP-Laplace approximation in terms of accuracy. The strong performance of the method is due to it being less prone to converge to suboptimal solutions. The variational methods have converged to parameter settings with an average log loss worse than 0.69 for several data sets. That is, on these data sets, the variational methods were inferior to a classifier that flips a coin. The higher standard deviation with respect to the performance indicators of the variational methods can also be traced back to this brittleness when it comes to convergence.

In contrast to the good performance of the GP-Laplace approximation, the t-Laplace approximation performs poorly with respect to log loss and accuracy. Aggravatingly, this performance cannot be explained with bad results for some of the data sets. The t-Laplace approximation performs uniformly weakly over the training sets (indicated by the relatively low standard deviation of the performance metrics). However, based on the results for the simulated data, this poor performance is not surprising.

Comparing the variational methods to each other, we see that the two variational Student-t methods perform differently for the real world data sets. On the one hand, the VTP1 algorithm is the best variational method in terms of log loss and a close second to the GP-Laplace approximation for how many times the method was the best method for the accuracy score. On the other hand, the VTP2 algorithm cannot compete with the other methods. Especially with respect to accuracy, the results are disappointing for VTP2, even the t-Laplace approximation scored higher. In contrast, the variational Gaussian approximation performs relatively well and, in particular, dominates the ”log loss - times best” category clearly. However, like the other variational algorithms, it suffers from occasionally converging to suboptimal parameters.

Table 7.1: Summary statistics of the results on the small data sets.

	Avg. Log Loss			Accuracy		
Method	Mean	Std	times best	Mean	Std	times best
LAPL	**0.473**	0.181	7	**0.777**	0.146	**18**
TLAPL	0.657	0.236	2	0.700	0.182	6
VGP	0.543	0.386	**15**	0.733	0.223	16
VTP1	0.508	0.287	10	0.716	0.224	17
VTP2	0.579	0.279	3	0.639	0.236	10

7.4.3 Large Data

Tables 7.2, 7.3, and 7.4 show the summary statistics for the sparse methods with 50, 100, and 200 inducing points respectively. In accordance with the training for the smaller data sets, the degrees of freedom of the Student-t based methods have been picked via cross-validation and besides of this, none of the methods has been tuned to improve performance. The individual results for the data sets can be found in the appendix C.1.

The most noteworthy characteristic of the results is the overall number of successful trainings. While the Penn Machine Learning Benchmarks contain around 50 binary classification problems with more than 500 observations, the variational sparse inducing point methods have converged for less than 50% of them to a result that is, on average, better than flipping a coin. The literature contains references to these convergence problems for inducing point methods Gal et al. (2014). However, besides using kmeans-clustering to initiate the inducing points[8], we have not found any other initialization strategy to mitigate these convergence problems in the literature. Therefore, we argue that our experiments show that the effective initialization of inducing points is still an open research question.

Regarding the performance of the methods, the sparse variational Gaussian approximation is the best method in every category except for the accuracy score when 200 inducing points are used. Nevertheless, it is difficult to generalize these results. All the methods failed to converge on a majority of the data sets. Furthermore, in cases where all the methods converged, the difference between them is frequently not that substantial. Anyway, for all configurations, there are multiple data sets for which the Student-t methods outperform their Gaussian counterpart, indicating that the Student-t approaches can be a viable opportunity to improve on the variational Gaussian approximation results.

Moreover, the results show two additional characteristics that need to be addressed. On the one hand, it might be surprising that the average log loss is not steadily declining with the increasing number of inducing points. However, this can be explained that the underlying data sets are not the same. The methods have converged for more data sets with 200 inducing points than for 100 or 50. That is, the

[8] Which we have done.

statistics of the different tables are only to a limited extend comparable. On the other hand, one might wonder why the Student-t based methods seem to perform almost on par with their Gaussian counterpart in terms of accuracy, even outperforming the Gaussian method for 200 inducing points, while in terms of log loss, the variational Gaussian method clearly wins for all inducing points. This behaviour can be linked back to what we have observed for the simulated data. The variational Student-t approximations assign less probability to classes in general. Consequently, even when one of the Student-t methods has a higher accuracy than the Gaussian method, it might still fall short in terms of log loss, because it assigns lower class probabilities to most of the observations, which can be worse than predicting individual observations wrongly.

Table 7.2: Summary statistics for the sparse methods with 50 inducing points

	Avg. Log Loss			Accuracy		
Method	Mean	Std	times best	Mean	Std	times best
VGP	**0.193**	0.174	**10**	**0.854**	0.234	**13**
VTP1	0.221	0.214	3	0.836	0.242	11
VTP2	0.211	0.193	5	0.842	0.237	11

Table 7.3: Summary statistics for the sparse methods with 100 inducing points

	Avg. Log Loss			Accuracy		
Method	Mean	Std	times best	Mean	Std	times best
VGP	**0.183**	0.156	**8**	**0.856**	0.238	**10**
VTP1	0.213	0.194	6	0.841	0.243	8
VTP2	0.212	0.196	3	0.840	0.243	9

Table 7.4: Summary statistics for the sparse methods with 200 inducing points

	Avg. Log Loss			Accuracy		
Method	Mean	Std	times best	Mean	Std	times best
VGP	**0.214**	0.231	**13**	0.858	0.204	**13**

VTP1	0.265	0.261	3	0.861	0.183	10
VTP2	0.264	0.261	5	**0.862**	0.181	11

Chapter 8

Outlook Future Work

In this chapter, we give an overview of potential future work. The first two sections are dedicated to topics which we are actively pursuing, whereas the last one covers less concrete research projects.

8.1 Reparameterization Trick

One disadvantage of the variational Student-t approaches derived in Chapter 5 is that, in case of a dense covariance matrix $\boldsymbol{V}$, $\mathcal{O}(n^2)$ need to be learnt for the normal variational bound or $\mathcal{O}(m^2)$ in the scalable version.

However, for the GP case, Seeger (1999) and Nickisch & Rasmussen (2008) showed that a bound with $\mathcal{O}(n)$ free parameters can be obtained. This was achieved by calculating the first derivatives with respect to the variational parameters, $\boldsymbol{m}$ and $\boldsymbol{V}$, and then solving for the parameters. Following the same approach for the ELBO of VTP1:

$$\frac{\partial ELBO}{\partial \boldsymbol{m}} = \frac{\partial A}{\partial \boldsymbol{m}} + \frac{2\Psi_p}{1-t}(\nu \boldsymbol{K})^{-1}\boldsymbol{m} = 0 \Rightarrow \frac{\partial A}{\partial \boldsymbol{m}} = \frac{2\Psi_p}{t-1}(\nu \boldsymbol{K})^{-1}\boldsymbol{m} \quad (8.1)$$

$$\frac{\partial ELBO}{\partial \boldsymbol{V}} = \frac{\partial A}{\partial \boldsymbol{V}} + \frac{\Psi_p}{(1-t)(\nu-2)}\boldsymbol{K}^{-1} - D\left(\frac{t-1}{2t}\right)\det|\boldsymbol{V}|^{\frac{t-1}{2t}}\boldsymbol{V}^{-1} = 0. \quad (8.2)$$

In contrast to the GP case, $\frac{\partial ELBO}{\partial \boldsymbol{V}}$ cannot be used to obtain an alternative parameterization for the covariance matrix of the variational distribution. Nevertheless, if it is assumed that

$$\det |\boldsymbol{V}|^{\frac{t-1}{2t}} \approx 1,$$

which can be justified by $\frac{t-1}{2t}$ being close to 0, equation 8.2 can be solved for $\boldsymbol{V}$:

$$\frac{\partial ELBO}{\partial \boldsymbol{V}} = 0 \Rightarrow \boldsymbol{V} = D\left(\frac{t-1}{2t}\right)\left[\underbrace{\frac{\partial A}{\partial \boldsymbol{V}}}_{\boldsymbol{\Lambda}} + \frac{\Psi_p}{(1-t)(\nu-2)}\boldsymbol{K}^{-1}\right]^{-1}. \quad (8.3)$$

The interesting property of this reparameterization is that $\boldsymbol{\Lambda}$ is a diagonal matrix, while the variational parameter $\boldsymbol{V}$ is a dense matrix. As a consequence, instead of $\mathcal{O}(n^2)$ parameters to optimize, the method requires only the optimization of $2n$ parameters.

8.2 Variational Inference for Latent Student-t Process Models

The latent Gaussian process model by Lawrence (2005) is a Bayesian non-linear dimensionality reduction method that uses the GP regression model (with multiple outputs):

$$p(\boldsymbol{Y}|\boldsymbol{X}) = \int p(\boldsymbol{Y}|\boldsymbol{F})p(\boldsymbol{F}|\boldsymbol{X})d\boldsymbol{f}, \quad (8.4)$$

with unobserved $\boldsymbol{X}$. In the original work, the latent inputs $\boldsymbol{X}$ were optimized, as marginalizing out $\boldsymbol{X}$ requires the computation of integrals involving the kernel function with respect to $\boldsymbol{X}$, e.g. $\mathbb{E}_{p(\boldsymbol{X})}\left[\boldsymbol{K_{nn}}\right]$.

Titsias & Lawrence (2010) managed to approximately solve these integrals with variational sparse inducing point methods for certain types of kernels. That is, Titsias et. al. established variational lower bound on $p(\boldsymbol{Y})$ itself.

For the Student-t Process, it might not be possible to derive a lower bound in a similar way. Following Titsias & Lawrence (2010), we can start by t-relaxing and lower bounding $p(\boldsymbol{Y})$:

$$\log_t p(\boldsymbol{Y}) = \log_t \int q(\boldsymbol{X}) \frac{p(\boldsymbol{Y}|\boldsymbol{X}) \otimes_t p(\boldsymbol{X})}{q(\boldsymbol{X})} d\boldsymbol{X} \tag{8.5}$$

$$\geq \int q(\boldsymbol{X})^t \left[\log_t p(\boldsymbol{Y}|\boldsymbol{X}) + \log_t p(\boldsymbol{X}) - log_t q(\boldsymbol{X})\right] d\boldsymbol{X} \tag{8.6}$$

$$\geq \frac{1}{\int q(\boldsymbol{X})^t d\boldsymbol{X}} \left[\int \widetilde{q}(\boldsymbol{X}) \log_t p(\boldsymbol{Y}|\boldsymbol{X}) - D_t(q(\boldsymbol{X}) \parallel p(\boldsymbol{X})) \right]. \tag{8.7}$$

From chapter 6, we have our two ELBOs to lower bound $\log_t p(\boldsymbol{Y}|\boldsymbol{X})$. However, both involve $\boldsymbol{X}$ in the numerical integral via the kernel functions. In Titsias & Lawrence (2010) the integral is not an issue, because they were using a Gaussian observation model, which allows for a bound that does not require $q(\boldsymbol{f})$. The dispersion parameter of this distribution is given by

$$\boldsymbol{K}_{nn} + \boldsymbol{K}_{nm}\boldsymbol{K}_{mm}^{-1}\left(\boldsymbol{V} - \boldsymbol{K}_{mm}\right)\boldsymbol{K}_{mm}^{-1}\boldsymbol{K}_{mn},$$

whereas $\boldsymbol{K}_{nn}$ and $\boldsymbol{K}_{nm}$ both depend on $\boldsymbol{X}$. We have not found a reasonable approximation for the integral so far.

8.3 Miscellaneous

There are several research topics that immediately arise from this thesis, but for which we have not devoted extensive effort so far:

1. The primary issue of the TP methods is their instability with respect to convergence. There needs to be more work done on what factors are affecting the convergence of the methods and how to control them to obtain robust results. Potential starting points for this could be, for example, Sheth et al. (2015), who address convergence properties for the numerical integral of the variational Gaussian approximation, or Challis & Barber (2013), who investigate more general convergence questions of the variational Gaussian approximation.

2. Moreover, the current approach is to choose the degrees of freedom by cross-validation. This is necessary, because with variable degrees of freedom, the methods presented in this thesis get extremely unstable. However, being able

to optimize the degrees directly would avoid the cost of training the models multiple times. One approach could be to regularize the degrees of freedom to avoid large jumps, e.g. with a prior. Alternatively, proximal optimization methods (Parikh et al. (2014)) could also potentially allow for the optimization of the degrees of freedom.

3. As a natural extension of putting a prior onto the degrees of freedom, we could use priors on all the hyperparameters and test our methods in a fully Bayesian setting. This approach is also interesting from a stability point of view, the regularization imposed by the prios might force the parameters to converge to sensible values.

4. Additionally, there are also some potential TP models that could be targeted by the methods presented in this thesis. Our current work focuses on data without special structure. In contrast, GP priors have been used with structured data, such as time series data (e.g. Roberts et al. (2013), Frigola-Alcalde (2016), Wilson & Ghahramani (2010), Wu et al. (2014)) or graph/network data (Lloyd et al. (2012), Yu & Chu (2008)). It would be interesting to see how a TP prior performs and how to utilize our methods in such scenarios.

5. Another potential research direction is the deep Student-t Process. In case that we manage to derive tractable equations for latent Student-t Process models, we would attempt to generalize deep Gaussian Process by Damianou & Lawrence (2012).

6. Finally, there is also some opportunity in extending the algorithm toolbox for (generalized) TP models. A straighforward contribution would be the extension of the expectation propagation method from Futami et al. (2017) to GTPR models. Moreover, there is also the possibility to extend our algorithms to utilize distributed computing environments. For example, Gal et al. (2014) provides a foundation for distributed sparse inducing point methods.

Chapter 9

Conclusion

In this thesis, the problem of efficient and scalable inference for GTPR models has been considered. Using t-relaxation, a tool build on top of q-algebra, we have demonstrated how methods for GGPR models can be adapted to the GTPR case.

To be precise, we have used t-relaxation to derive the t-Laplace approximation, two versions of variational Student-t approximation, and two sparse inducing point methods. All these methods have in common that, in contrast to the methods that have inspired them, they use a multivariate Student-t distribution to approximate the posterior distribution.

The methods have also been tested on simulated and real data. From the experiments with simulated data, there are four important takeaways:

1. The t-Laplace approximation does not work properly. We are not sure whether the approximation is generally poor or there is an error in the implementation/derivation of the method, but in its current form, the t-Laplace approximation cannot be used reasonably.

2. While the regression experiments suggest that the degrees of freedom have a strong impact on the results obtained by the Student-t based approximation methods, the classification experiments draw a different picture, there we have seen only a negligible impact of the degrees of freedom on the results.

3. For individual datasets, the Student-t based methods seem to be less affected by mislabelled observations in terms of distorted decision boundaries. However, based on our simulation study, this does not appear to hold in general.

4. The variational Student-t methods are prone to converge to bad local minima and tuning them properly can take a considerable amount of time.

In contrast, from the real world benchmarks, we have learnt two important lessons:

1. For small data sets, there is no clear best-method-to-use. While the GP-Laplace approximation was the best method in terms of mean average log loss and accuracy, there were still many different data sets, where one of the other methods was better.

2. All the sparse inducing point methods had severe convergence problems.

Conclusively, we argue that the variational Student-t methods have the potential to be used in the future not as a replacement for the variational Gaussian methods, but as a complementary method. While the out-of-the-box performance of the variational Gaussian method is better, the variational Student-t can shine when the effect of outliers or mislabelled data needs to be mitigated and sufficient time is available to tune them properly.

In addition, our experiments would also suggest that the sparse inducing point methods for GTPR models can outperform their Gaussian counterparts occasionally. However, the standard inducing point methods was still more successful in the comparison. Moreover, the convergence problems of all the inducing point methods makes a comparison difficult.

In contrast to the other methods, we do not believe that the t-Laplace approximation is currently a viable alternative for GP models with Laplace approximation.

Appendix A

Important Identities and Properties

A.1 Q-Algebra

$$\log_t(x) = \frac{x^{1-t} - 1}{1-t} \tag{A.1}$$

$$\exp_t(x) = [1 + (1-t)x]^{\frac{1}{1-t}} . \tag{A.2}$$

$$\log_t(xy) = \log_t(x) + \log_t(y) + (1-t)xy. \tag{A.3}$$

$$\log_t\left(\frac{x}{y}\right) = y^{1-t}\left[\log_t(x) - \log_t(y)\right]. \tag{A.4}$$

$$x \otimes_t y = \left[x^{1-t} + y^{1-t} - 1\right]^{\frac{1}{1-t}} . \tag{A.5}$$

$$x \ominus_t y = \left[x^{1-t} - y^{1-t} + 1\right]^{\frac{1}{1-t}} . \tag{A.6}$$

$$\log_t(x \otimes_t y) = \log_t(x) + \log_t(y). \tag{A.7}$$

$$\log_t(x \ominus_t y) = \log_t(x) - \log_t(y). \tag{A.8}$$

Appendix B

Supplementary Derivations

B.1 Concavity of t-logarithm

It can be shown that the t-logarithm is concave for positive values of x and t. Recall that a function f(x) is concave, if its second derivative is negative for all x Wasserman (2010).

$$\begin{aligned}\frac{d^2}{dx^2}\log_t(x) &= \frac{d^2}{dx^2}\frac{x^{1-t}-1}{1-t} && \text{(B.1)}\\ &= \frac{d}{dx}\frac{d}{dx}\frac{x^{1-t}-1}{1-t} && \text{(B.2)}\\ &= \frac{d}{dx}\frac{1}{x^{-t}} && \text{(B.3)}\\ &= (-t)\frac{1}{x^{t+1}}. && \text{(B.4)}\end{aligned}$$

While the first term, $-t$, is negative for all $t > 0$, the second term is positive for all $x > 0$. Consequently, the second derivative of the t-logarithm is negative for all $t > 0$ and $x > 0$ and therefore the t-logarithm is concave within these bounds.

From this it readily follows that the Jensen's inequality holds for the t-logarithm for all $t > 0$ and $x > 0$.

B.2 Derivation of q(f)

First, we start with the joint multivariate Student-t distribution of $p(\boldsymbol{f}, \boldsymbol{u})$ and use the Gaussian scale-mixture representation on it:

$$
\begin{aligned}
p\left(\boldsymbol{f}\right) &= \int_{\mathbb{F}}\int_{\mathbb{R}} p\left(\boldsymbol{f},\boldsymbol{u}|r\right)p(r)drd\boldsymbol{f} && \text{(B.5)}\\
&= \int_{\mathbb{F}}\int_{\mathbb{R}} p\left(\boldsymbol{f},|\boldsymbol{u},r\right)p(\boldsymbol{u}|r)p(r)drd\boldsymbol{f}. && \text{(B.6)}
\end{aligned}
$$

Now we are introducing the variational distribution by substituting $p(\boldsymbol{u}|r)$ with the Gaussian part of the scale-mixture representation of the variational distribution $q(\boldsymbol{u})$, that is:

$$
q\left(\boldsymbol{u}\right) = \int q(\boldsymbol{u}|r)p(r)dr, \tag{B.7}
$$

where we assume that the Gamma density term $p(r)$ is identical to the one used for the joint multivariate Student-t distribution, that is:

$$
q(\boldsymbol{f}) = \int_{\mathbb{F}}\int_{\mathbb{R}} \underbrace{p\left(\boldsymbol{f},|\boldsymbol{u},r\right)q(\boldsymbol{u}|r)}_{:=A}p(r)drd\boldsymbol{f}. \tag{B.8}
$$

It is important to notice, that the integral A is a marginalization over a multivariate Gaussian distribution, that is:

$$
A \sim \mathcal{N}\left(\boldsymbol{f};\boldsymbol{\alpha},\frac{\boldsymbol{B}}{r}\right), \tag{B.9}
$$

with

$$
\begin{aligned}
\boldsymbol{\alpha} &= \boldsymbol{K_{nm}}\boldsymbol{K_{mm}^{-1}}\boldsymbol{m} && \text{(B.10)}\\
\boldsymbol{B} &= \boldsymbol{K_{nn}} + \boldsymbol{K_{nm}}\boldsymbol{K_{mm}^{-1}}\left(\boldsymbol{V}-\boldsymbol{K_{mm}}\right)\boldsymbol{K_{mm}^{-1}}\boldsymbol{K_{mn}}. && \text{(B.11)}
\end{aligned}
$$

Substituting this result back into B.8 gives:

$$
q\left(\boldsymbol{f}\right) = \int_{\mathbb{R}} \mathcal{N}\left(\boldsymbol{f};\boldsymbol{\alpha},\frac{\boldsymbol{B}}{r}\right) Gamma\left(r;\frac{\nu}{2},\frac{\nu}{2}\right)dr. \tag{B.12}
$$

This is the Gaussian scale-mixture representation of a Student-t distribution:

$$
q\left(\boldsymbol{f}\right) \sim \mathcal{T}\left(\nu,\boldsymbol{\alpha},\boldsymbol{B}\right), \tag{B.13}
$$

whereas $\boldsymbol{\alpha}$ and $\boldsymbol{B}$ are the same as in B.10 and B.11 respectively.

B.3 Sparse VTP2 Problem

While the VTP2 method seems to be cleaner in terms of complexity deriving it, compared to the VTP1 method, its major drawback is that the method cannot be readily translated into sparse inducing point method.

From section 6.1, we know that:

$$\begin{aligned} \log_t p(\boldsymbol{y}|\boldsymbol{u}) &= \log_t \int p(\boldsymbol{y}|\boldsymbol{f}) p(\boldsymbol{f}|\boldsymbol{u}) d\boldsymbol{f} && \text{(B.14)} \\ &\geq \int p(\boldsymbol{f}|\boldsymbol{u}) \log_t p(\boldsymbol{y}|\boldsymbol{f}) d\boldsymbol{f} && \text{(B.15)} \\ &= \mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})} \left[\log_t p(\boldsymbol{y}|\boldsymbol{f})\right]. && \text{(B.16)} \end{aligned}$$

Then we can derive a lower bound on $p(\boldsymbol{y})$ based on VTP2:

$$\begin{aligned} \log_t p(\boldsymbol{y}) &= \log_t \int p(\boldsymbol{y}|\boldsymbol{u}) \otimes p(\boldsymbol{u}) d\boldsymbol{u} && \text{(B.17)} \\ &= \log_t \int q(\boldsymbol{u}) \frac{p(\boldsymbol{y}|\boldsymbol{u}) \otimes p(\boldsymbol{u})}{q(\boldsymbol{u})} d\boldsymbol{u} && \text{(B.18)} \\ &= \left(\int q(\boldsymbol{u})^t d\boldsymbol{u}\right) \Bigg[\int \widetilde{q}(\boldsymbol{u}) \log_t p(\boldsymbol{y}|\boldsymbol{u}) d\boldsymbol{u} + \\ &\quad \int \widetilde{q}(\boldsymbol{u}) (\log_t p(\boldsymbol{u}) - \log_t q(\boldsymbol{u})) d\boldsymbol{u} \Bigg] && \text{(B.19)} \\ &\geq \left(\int q(\boldsymbol{u})^t d\boldsymbol{u}\right) \times \Bigg[\int \widetilde{q}(\boldsymbol{u}) \log_t \mathbb{E}_{p(\boldsymbol{f}|\boldsymbol{u})} \left[\log_t p(\boldsymbol{y}|\boldsymbol{f})\right] d\boldsymbol{u} + \\ &\quad \int \widetilde{q}(\boldsymbol{u}) (\log_t p(\boldsymbol{u}) - \log_t q(\boldsymbol{u})) d\boldsymbol{u} \Bigg], && \text{(B.20)} \end{aligned}$$

where $\widetilde{q}(\boldsymbol{u}) = \frac{q(\boldsymbol{u})^t}{\int q(\boldsymbol{u})^t d\boldsymbol{u}}$ is the escort distribution of $q(\boldsymbol{u})$ (see Section 5.2 for more details).

We would now need to approximate the integral $\int p(\boldsymbol{f}|\boldsymbol{u}) \widetilde{q}(\boldsymbol{u})$. However, the problem is that the distributions have different degrees of freedom, we can therefore not argue that they are jointly, multivariate Student-t distributed or invoke our approximation from appendix B.2, which would also require the distributions to have the same degrees of freedom to be justifiable. Therefore, we have not covered a direct extension of the VTP2 algorithm in the sparse inducing point section.

Appendix C

Experiments

C.1 Result Tables

C.2 Small Data

Table C.1: Comparison of average log loss results for the small data sets.

	Avg. Log Loss				
Dataset	LAPL	TLAPL	VGP	VTP	VTP2
heart-h	**0.4612**	0.6922	0.6397	0.6785	0.6873
analcatdata_asbestos	0.6347	0.661	**0.6339**	0.6431	0.6946
analcatdata_boxing1	0.614	0.6144	**0.5720**	0.6421	0.6931
analcatdata_boxing2	0.6097	0.6372	0.6131	**0.6076**	0.6931
analcatdata_creditscore	0.1532	0.6842	0.0711	**0.0589**	**0.0589**
analcatdata_fraud	0.6043	0.6064	**0.5970**	0.6494	0.6494
analcatdata_japansolvent	0.6931	**0.6741**	0.6929	0.6931	0.6931
analcatdata_lawsuit	0.5006	0.656	**0.0642**	0.0919	0.1131
appendicitis	0.3428	1.1515	**0.3408**	0.3437	0.4485
backache	0.3807	0.6931	**0.3802**	0.4221	0.6931
breast-cancer	0.5899	0.5784	**0.5355**	0.7358	0.748
bupa	**0.5770**	0.694	1.0395	0.6831	0.6831

table continues

	Avg. Log Loss				
Dataset	LAPL	TLAPL	VGP	VTP	VTP2
cleve	0.6931	0.6931	**0.6287**	0.6931	0.6931
colic	**0.6516**	0.6897	0.6637	0.67	0.67
corral	0.1975	0.258	0.0437	**0.0301**	0.0305
glass2	0.5417	0.5486	0.5801	**0.5411**	0.6011
haberman	**0.5141**	0.7131	1.2906	0.8244	0.7945
heart-c	0.6931	0.6931	**0.6059**	0.6931	0.6931
hepatitis	**0.4944**	0.6931	1.5821	0.6931	0.6931
horse-colic	0.6916	0.6945	0.7088	**0.6763**	**0.6763**
house-votes-84	0.1071	0.2487	**0.0848**	0.1919	0.542
hungarian	**0.5350**	0.6923	0.5936	0.6931	0.6931
ionosphere	0.2911	0.3799	**0.2422**	0.3927	0.7295
labor	0.3329	0.6924	**0.2330**	0.2859	0.2859
liver-disorder	**0.6246**	0.6937	1.1256	0.6835	0.6852
mux6	0.3958	0.5072	0.1383	**0.0954**	0.0975
postoperative-patient-data	0.5596	**0.5478**	1.3944	1.351	1.397
prnn_crabs	0.1394	0.6615	0.0681	**0.0266**	0.1201
prnn_synth	0.2694	1.691	0.2493	**0.2443**	0.6931
saheart	0.5636	0.6931	**0.5633**	0.6931	0.6931
sonar	0.4446	0.5501	**0.4081**	0.4292	0.6931
spect	0.4861	0.5836	**0.4271**	0.5397	0.7526
spectf	0.6037	0.6485	0.5763	**0.5102**	**0.5102**
vote	0.1086	0.4493	0.1003	**0.0963**	0.1021

Table C.2: Comparison of accuracy results for the small data sets.

	Accuracy				
Dataset	LAPL	TLAPL	VGP	VTP	VTP2
heart-h	**0.7881**	0.661	0.6271	0.678	0.678

table continues

	Accuracy				
Dataset	LAPL	TLAPL	VGP	VTP	VTP2
analcatdata_asbestos	0.5294	0.6765	0.5294	**0.7059**	**0.7059**
analcatdata_boxing1	0.7083	0.6875	**0.7500**	**0.7500**	0.6875
analcatdata_boxing2	0.7358	0.6415	0.717	**0.7925**	0.5472
analcatdata_creditscore	**1.0000**	**1.0000**	0.975	**1.0000**	0.975
analcatdata_fraud	**0.6471**	**0.6471**	**0.6471**	**0.6471**	**0.6471**
analcatdata_japansolvent	**0.5714**	0.5238	0.4762	0.5238	0.5238
analcatdata_lawsuit	**0.9717**	0.9623	**0.9717**	**0.9717**	**0.9717**
appendicitis	**0.8605**	0.7674	**0.8605**	**0.8605**	0.7907
backache	**0.8750**	0.8611	**0.8750**	0.8611	0.125
breast-cancer	**0.7304**	**0.7304**	0.7217	0.3043	0.3043
bupa	**0.7101**	0.3768	0.2609	0.4565	0.4565
cleve	0.5738	0.5984	**0.6885**	0.4508	0.4098
colic	0.6419	**0.6622**	0.6554	0.6486	0.6554
corral	**1.0000**	0.8594	**1.0000**	**1.0000**	**1.0000**
glass2	0.7424	0.7273	0.7273	**0.7879**	0.6818
haberman	**0.7724**	0.6748	0.6829	0.7236	0.7236
heart-c	0.5328	0.5574	**0.6475**	0.4672	0.4672
hepatitis	**0.8065**	0.1935	0.0968	0.1935	0.1935
horse-colic	0.6216	0.6284	0.5676	**0.6486**	**0.6486**
house-votes-84	0.977	0.9195	**0.9885**	0.9828	0.454
hungarian	**0.7542**	0.5339	0.6864	0.4237	0.4237
ionosphere	0.8794	0.8369	**0.9007**	0.6809	0.6241
labor	0.913	0.8696	**0.9565**	0.913	0.913
liver-disorder	**0.6957**	0.3551	0.2246	0.4348	0.4348
mux6	0.9231	0.7692	**0.9615**	**0.9615**	**0.9615**
postoperative-patient-data	**0.7778**	**0.7778**	**0.7778**	**0.7778**	**0.7778**
prnn_crabs	**1.0000**	0.9375	0.9875	0.9875	0.9875
prnn_synth	**0.9000**	0.51	**0.9000**	**0.9000**	0.49

table continues

	Accuracy				
Dataset	LAPL	TLAPL	VGP	VTP	VTP2
saheart	**0.6865**	0.6108	0.6811	0.3568	0.3568
sonar	0.8333	0.7143	0.8214	**0.8452**	0.5
spect	0.8131	0.7757	0.7757	**0.8318**	**0.8318**
spectf	0.5214	**0.8429**	**0.8429**	**0.8429**	**0.8429**
vote	0.954	0.9425	**0.9598**	**0.9598**	**0.9598**

C.3 Large Data

Table C.3: Comparison of average log loss results for the sparse inducing point methods with 50 inducint points.

	Avg. Log Loss		
Dataset	VGP	VTP	VTP2
agaricus-lepiota	**0.0103**	0.0171	0.0175
banana	0.2103	0.2087	**0.2086**
breast-w	0.0808	**0.0778**	0.0782
chess	**0.0482**	0.0622	0.064
flare	0.4245	0.4229	**0.4228**
irish	0.0918	**0.0757**	0.0819
mofn-3-7-10	0.0026	0.0026	**0.0024**
monk1	**0.3720**	0.4898	0.4894
monk2	**0.4172**	0.4272	0.4297
monk3	0.1122	0.1101	**0.1091**
mushroom	**0.0085**	0.0151	0.0148
phoneme	**0.3126**	0.3265	0.3277
spambase	**0.3930**	0.6931	0.5087
threeOf9	**0.0886**	0.0938	0.0961
tic-tac-toe	**0.2937**	0.3611	0.3533

table continues

	Avg. Log Loss		
Dataset	VGP	VTP	VTP2
titanic	0.5183	**0.5124**	0.5133
twonorm	0.0553	0.0553	**0.0552**
xd6	**0.0341**	0.041	0.0423

Table C.4: Comparison of accuracy results for the sparse inducing point methods with 50 inducing points.

	Accuracy		
Dataset	VGP	VTP	VTP2
agaricus-lepiota	**1.0000**	0.9988	0.9985
banana	0.3901	**0.3920**	**0.3920**
breast-w	**0.9750**	**0.9750**	**0.9750**
chess	**0.9867**	0.9805	0.9789
flare	0.8126	**0.8150**	**0.8150**
irish	0.985	**0.9900**	**0.9900**
mofn-3-7-10	**1.0000**	**1.0000**	**1.0000**
monk1	**0.8072**	0.704	0.704
monk2	0.7925	**0.8091**	0.805
monk3	**0.9775**	**0.9775**	**0.9775**
mushroom	**1.0000**	**1.0000**	**1.0000**
phoneme	**0.8529**	0.8511	0.8483
spambase	**0.8256**	0.6181	0.7355
threeOf9	**0.9610**	**0.9610**	**0.9610**
tic-tac-toe	**0.9062**	0.8776	0.8802
titanic	**0.1260**	**0.1260**	**0.1260**
twonorm	0.9801	0.9801	**0.9824**
xd6	**1.0000**	**1.0000**	**1.0000**

table continues

Table C.5: Comparison of average log loss results for the sparse inducing point methods with 100 inducint points.

	Avg. Log Loss		
Dataset	VGP	VTP	VTP2
banana	0.2092	**0.2088**	0.2089
breast-w	0.0809	0.0806	**0.0803**
chess	**0.0418**	0.0577	0.0565
flare	0.4246	**0.4229**	0.423
irish	0.0856	0.0835	**0.0654**
kr-vs-kp	**0.0530**	0.0723	0.0721
mofn-3-7-10	0.0026	**0.0016**	0.0018
monk1	**0.2678**	0.286	0.296
monk2	0.3463	**0.3290**	0.3535
monk3	0.1114	**0.1031**	0.1063
phoneme	**0.3022**	0.3125	0.3123
spambase	**0.3490**	0.6931	0.6931
threeOf9	**0.0771**	0.0856	0.0854
tic-tac-toe	**0.1755**	0.2872	0.2545
titanic	0.5143	**0.5127**	0.514
twonorm	0.0555	0.0554	**0.0553**
xd6	**0.0299**	0.0323	0.0352

Table C.6: Comparison of accuracy results for the sparse inducing point methods with 100 inducing points.

	Accuracy		
Dataset	VGP	VTP	VTP2
banana	0.3915	0.3929	**0.3934**
breast-w	**0.9750**	**0.9750**	**0.9750**

table continues

	Accuracy		
Dataset	VGP	VTP	VTP2
chess	**0.9891**	0.9812	0.982
flare	0.8126	**0.8150**	**0.8150**
irish	0.985	0.985	**0.9900**
kr-vs-kp	**0.9828**	0.9789	0.9797
mofn-3-7-10	**1.0000**	**1.0000**	**1.0000**
monk1	**0.8879**	0.8744	0.852
monk2	0.8257	**0.8382**	0.8299
monk3	0.9775	**0.9820**	**0.9820**
phoneme	0.8566	**0.8580**	0.8575
spambase	**0.8539**	0.6312	0.6339
threeOf9	**0.9707**	0.961	0.961
tic-tac-toe	**0.9505**	0.9167	0.9245
titanic	**0.1260**	**0.1260**	**0.1260**
twonorm	0.9804	0.9818	**0.9824**
xd6	**1.0000**	**1.0000**	**1.0000**

Table C.7: Comparison of average log loss results for the sparse inducing point methods with 200 inducint points.

	Avg. Log Loss		
Dataset	VGP	VTP	VTP2
agaricus-lepiota	**0.0060**	0.0119	0.0121
banana	**0.2089**	0.2119	0.2122
breast-w	**0.0810**	0.0827	0.0824
chess	**0.0387**	0.0514	0.0498
flare	**0.4246**	0.4248	0.4256
GAMETES_Epistasis_2-Way_20atts_0	0.6931	0.6887	**0.6869**
GAMETES_Heterogeneity_20atts_1600_Het_0	0.6931	**0.6421**	0.6466

table continues

	Avg. Log Loss		
Dataset	VGP	VTP	VTP2
GAMETES_Heterogeneity_20atts_1600_Het_0	0.6931	0.6847	**0.6834**
irish	0.0847	0.0662	**0.0592**
kr-vs-kp	**0.0475**	0.0619	0.0606
mofn-3-7-10	0.0028	**0.0014**	**0.0014**
monk1	0.2235	**0.2110**	0.2146
monk2	0.2793	**0.2771**	0.2838
monk3	**0.1109**	0.1173	0.1185
mushroom	**0.0049**	0.0088	0.0096
phoneme	**0.2940**	0.2989	0.2988
spambase	**0.2969**	0.6931	0.6931
threeOf9	**0.0765**	0.094	0.0882
tic-tac-toe	**0.1525**	0.2269	0.213
twonorm	**0.0559**	0.6931	0.6931
xd6	0.0299	0.0295	**0.0294**

Table C.8: Comparison of accuracy results for the sparse inducing point methods with 200 inducing points.

	Accuracy		
Dataset	VGP	VTP	VTP2
agaricus-lepiota	**1.0000**	**1.0000**	**1.0000**
banana	0.3915	0.3929	**0.3934**
breast-w	**0.9750**	0.9679	0.9679
chess	**0.9898**	0.9859	0.9867
flare	**0.8126**	**0.8126**	0.808
GAMETES_Epistasis_2-Way_20atts_0	0.4859	0.5641	**0.5859**
GAMETES_Heterogeneity_20atts_1600_Het_0	0.4859	**0.6438**	0.6406
GAMETES_Heterogeneity_20atts_1600_Het_0	0.4859	0.5719	0.5813

table continues

Dataset	Accuracy		
	VGP	VTP	VTP2
irish	0.985	**0.9900**	**0.9900**
kr-vs-kp	**0.9883**	0.9812	0.9812
mofn-3-7-10	**1.0000**	**1.0000**	**1.0000**
monk1	0.9148	**0.9327**	**0.9327**
monk2	0.8714	**0.8880**	**0.8880**
monk3	**0.9775**	**0.9775**	0.973
mushroom	**1.0000**	**1.0000**	**1.0000**
phoneme	**0.8659**	0.8571	0.8589
spambase	**0.8821**	0.6442	0.6426
threeOf9	**0.9707**	0.9659	0.9659
tic-tac-toe	**0.9583**	0.9375	0.9375
twonorm	0.9791	0.9797	**0.9821**
xd6	**1.0000**	**1.0000**	**1.0000**

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. & Zheng, X. (2015), 'TensorFlow: Large-scale machine learning on heterogeneous systems'. Software available from tensorflow.org.
URL: *http://tensorflow.org/*

Abramowitz, M. & Stegun, I. A. (1964), *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, Vol. 55, Courier Corporation.

Amari, S.-i. (2016), *Information geometry and its applications*, Vol. 194, Springer.

Archambeau, C. & Bach, F. (2011), 'Multiple gaussian process models', *arXiv preprint arXiv:1110.5238* .

Barbaresco, F. & Nielsen, F. (2017), *Differential Geometrical Theory of Statistics*, MDPI AG-Multidisciplinary Digital Publishing Institute.

Barber, D. & Bishop, C. M. (1998), 'Ensemble learning for multi-layer networks', *Advances in neural information processing systems* pp. 395–401.

Bauer, M., van der Wilk, M. & Rasmussen, C. E. (2016), Understanding probabilistic sparse gaussian process approximations, *in* 'Advances in Neural Information Processing Systems', pp. 1525–1533.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A. & Siskind, J. M. (2018), 'Automatic differentiation in machine learning: a survey', *Journal of Marchine Learning Research* **18**, 1–43.

Bishop, C. M. (2006), *Pattern recognition and machine learning*, springer.

Borges, E. P. (2004), 'A possible deformed algebra and calculus inspired in nonextensive thermostatistics', *Physica A: Statistical Mechanics and its Applications* **340**(1-3), 95–101.

Brooks, S., Gelman, A., Jones, G. L. & Meng, X.-L. (2011), *Handbook of markov chain monte carlo*, Chapman and Hall/CRC.

Bui, T. D., Nguyen, C. & Turner, R. E. (2017), Streaming sparse gaussian process approximations, *in* 'Advances in Neural Information Processing Systems', pp. 3299–3307.

Burden, R. L. & Faires, J. D. (1997), 'Numerical analysis, brooks', *Cole Pub* **7**.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P. & Riddell, A. (2016), 'Stan: A probabilistic programming language', *Journal of Statistical Software* **20**.

Challis, E. & Barber, D. (2013), 'Gaussian kullback-leibler approximate inference', *The Journal of Machine Learning Research* **14**(1), 2239–2286.

Cseke, B. & Heskes, T. (2011), 'Approximate marginals in latent gaussian models', *Journal of Machine Learning Research* **12**(Feb), 417–454.

Damianou, A. C. & Lawrence, N. D. (2012), 'Deep gaussian processes', *arXiv preprint arXiv:1211.0358* .

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J. & Blei, D. (2017), Variational inference via chi upper bound minimization, *in* 'Advances in Neural Information Processing Systems', pp. 2732–2741.

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

Ding, N., Qi, Y. & Vishwanathan, S. (2011), t-divergence based approximate inference, *in* 'Advances in Neural Information Processing Systems', pp. 1494–1502.

Ding, N. & Vishwanathan, S. (2010), t-logistic regression, *in* 'Advances in Neural Information Processing Systems', pp. 514–522.

Ding, P. (2016), 'On the conditional distribution of the multivariate t distribution', *The American Statistician* **70**(3), 293–295.

Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987), 'Hybrid monte carlo', *Physics letters B* **195**(2), 216–222.

Fang, K.-T., Kotz, S. & Ng, K. W. (1990), *Symmetric multivariate and related distributions*, Chapman and Hall.

Fog, A. (2008), 'Calculation methods for wallenius' noncentral hypergeometric distribution', *Communications in StatisticsSimulation and Computation®* **37**(2), 258–273.

Frigola-Alcalde, R. (2016), Bayesian time series learning with Gaussian processes, PhD thesis, University of Cambridge.

Futami, F., Sato, I. & Sugiyama, M. (2017), Expectation propagation for t-exponential family using q-algebra, *in* 'Advances in Neural Information Processing Systems', pp. 2245–2254.

Gal, Y., van der Wilk, M. & Rasmussen, C. (2014), Distributed variational inference in sparse gaussian process regression and latent variable models, *in* 'Advances in Neural Information Processing Systems', pp. 3257–3265.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian data analysis*, Chapman and Hall/CRC.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, gibbs distributions, and the bayesian restoration of images', *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

Golub, G. H. & Van Loan, C. F. (2012), *Matrix computations*, Vol. 3, JHU press.

GPy (since 2012), ‘GPy: A gaussian process framework in python’, `http://github.com/SheffieldML/GPy`.

Gradshteyn, I. S. & Ryzhik, I. M. (2014), *Table of integrals, series, and products*, Academic press.

Hastings, W. K. (1970), ‘Monte carlo sampling methods using markov chains and their applications’.

Hensman, J., Matthews, A. G., Filippone, M. & Ghahramani, Z. (2015), Mcmc for variationally sparse gaussian processes, *in* ‘Advances in Neural Information Processing Systems’, pp. 1639–1647.

Hensman, J., Matthews, A. & Ghahramani, Z. (2015), ‘Scalable variational gaussian process classification’.

Herbrich, R., Lawrence, N. D. & Seeger, M. (2003), Fast sparse gaussian process methods: The informative vector machine, *in* ‘Advances in neural information processing systems’, pp. 625–632.

Hernández-Lobato, D., Hernández-Lobato, J. M. & Dupont, P. (2011), Robust multi-class gaussian process classification, *in* ‘Advances in neural information processing systems’, pp. 280–288.

Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernandez-Lobato, D. & Turner, R. (2016), Black-box alpha divergence minimization, *in* ‘International Conference on Machine Learning’, pp. 1511–1520.

Huber, P. J. (2011), *Robust statistics*, Springer.

Jones, E., Oliphant, T. & Peterson, P. (2014), ‘{SciPy}: open source scientific tools for {Python}’.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1999), ‘An introduction to variational methods for graphical models’, *Machine learning* **37**(2), 183–233.

Jylänki, P., Vanhatalo, J. & Vehtari, A. (2011), ‘Robust gaussian process regression with a student-t likelihood’, *Journal of Machine Learning Research* **12**(Nov), 3227–3257.

Khan, E., Mohamed, S. & Murphy, K. P. (2012), Fast bayesian inference for non-conjugate gaussian process regression, *in* 'Advances in Neural Information Processing Systems', pp. 3140–3148.

Kim, H.-C. & Ghahramani, Z. (2008), Outlier robust gaussian process classification, *in* 'Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)', Springer, pp. 896–905.

Kotz, S. & Nadarajah, S. (2004), *Multivariate t-distributions and their applications*, Cambridge University Press.

Kullback, S. (1997), *Information theory and statistics*, Courier Corporation.

L. Wauthier, F. & Jordan, M. (2010), Heavy-tailed process priors for selective shrinkage., pp. 2406–2414.

Laplace, P. S. (1986), 'Memoir on the probability of the causes of events', *Statistical Science* **1**(3), 364–378.

Lawrence, N. (2005), 'Probabilistic non-linear principal component analysis with gaussian process latent variable models', *Journal of machine learning research* **6**(Nov), 1783–1816.

Liu, Q. & Wang, D. (2016), Stein variational gradient descent: A general purpose bayesian inference algorithm, *in* 'Advances In Neural Information Processing Systems', pp. 2378–2386.

Lloyd, C., Gunter, T., Osborne, M. & Roberts, S. (2015), Variational inference for gaussian process modulated poisson processes, *in* 'International Conference on Machine Learning', pp. 1814–1822.

Lloyd, J., Orbanz, P., Ghahramani, Z. & Roy, D. M. (2012), Random function priors for exchangeable arrays with applications to graphs and relational data, *in* 'Advances in Neural Information Processing Systems', pp. 998–1006.

Mair, S. & Brefeld, U. (2018), 'Distributed robust gaussian process regression', *Knowledge and Information Systems* **55**(2), 415–435.
URL: *https://doi.org/10.1007/s10115-017-1084-7*

Matthews, A. G. d. G. (2017), Scalable Gaussian process inference using variational methods, PhD thesis, University of Cambridge.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., Leon-Villagra, P., Ghahramani, Z. & Hensman, J. (2017), 'Gpflow: A gaussian process library using tensorflow', *Journal of Machine Learning Research* **18**(40), 1–6.
URL: *http://jmlr.org/papers/v18/16-537.html*

Mattos, C. L. C., Dai, Z., Damianou, A., Barreto, G. A. & Lawrence, N. D. (2017), 'Deep recurrent gaussian processes for outlier-robust system identification', *Journal of Process Control* **60**, 82 – 94. DYCOPS-CAB 2016.
URL: *http://www.sciencedirect.com/science/article/pii/S0959152417301233*

Maybeck, P. S. (1982), *Stochastic models, estimation, and control*, Vol. 3, Academic press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The journal of chemical physics* **21**(6), 1087–1092.

Minka, T. P. (2001), Expectation propagation for approximate bayesian inference, *in* 'Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence', Morgan Kaufmann Publishers Inc., pp. 362–369.

Močkus, J. (1975), On bayesian methods for seeking the extremum, *in* 'Optimization Techniques IFIP Technical Conference', Springer, pp. 400–404.

Naish-Guzman, A. & Holden, S. (2008), The generalized fitc approximation, *in* 'Advances in Neural Information Processing Systems', pp. 1057–1064.

Naudts, J. (2004), 'Estimators, escort probabilities, and phi-exponential families in statistical physics', *J. Ineq. Pure Appl. Math* **5**(4), 102.

Neal, R. M. (1997), 'Monte carlo implementation of gaussian process models for bayesian regression and classification', *arXiv preprint physics/9701026* .

Nickisch, H. & Rasmussen, C. E. (2008), ‘Approximations for binary gaussian process classification’, *Journal of Machine Learning Research* **9**(Oct), 2035–2078.

Nocedal, J. & Wright, S. (2006), *Numerical optimization*, Springer Science & Business Media.

O’Hagan, A. (1979), ‘On outlier rejection phenomena in bayes inference’, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 358–367.

Oliphant, T. E. (2006), *A guide to NumPy*, Vol. 1, Trelgol Publishing USA.

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J. & Moore, J. H. (2017), ‘Pmlb: a large benchmark suite for machine learning evaluation and comparison’, *BioData Mining* **10**(1), 36.
URL: *https://doi.org/10.1186/s13040-017-0154-4*

Opper, M. & Archambeau, C. (2009), ‘The variational gaussian approximation revisited’, *Neural computation* **21**(3), 786–792.

Opper, M. & Saad, D. (2001), *Advanced mean field methods: Theory and practice*, MIT press.

Parikh, N., Boyd, S. et al. (2014), ‘Proximal algorithms’, *Foundations and Trends® in Optimization* **1**(3), 127–239.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), ‘Scikit-learn: Machine learning in Python’, *Journal of Machine Learning Research* **12**, 2825–2830.

Petersen, K. B. & Pedersen, M. S. (2012), ‘The matrix cookbook’. Version 20121115.
URL: *http://www2.imm.dtu.dk/pubdb/p.php?3274*

Prüher, J., Tronarp, F., Karvonen, T., Särkkä, S. & Straka, O. (2017), Student-t process quadratures for filtering of non-linear systems with heavy-tailed noise, *in* ‘2017 20th International Conference on Information Fusion (Fusion)’, IEEE, pp. 1–8.

Quiñonero-Candela, J. & Rasmussen, C. E. (2005), ‘A unifying view of sparse approximate gaussian process regression’, *The Journal of Machine Learning Research* **6**, 1939–1959.

Ranganath, R., Tran, D., Altosaar, J. & Blei, D. (2016), Operator variational inference, *in* ‘Advances in Neural Information Processing Systems’, pp. 496–504.

Ranjan, R., Huang, B. & Fatehi, A. (2016), ‘Robust gaussian process modeling using em algorithm’, *Journal of Process Control* **42**, 125 – 136.
URL: *http://www.sciencedirect.com/science/article/pii/S0959152416300282*

Rasmussen, C. E., Kuss, M. et al. (2003), Gaussian processes in reinforcement learning., *in* ‘NIPS’, Vol. 4, p. 1.

Rasmussen, C. E. & Williams, C. K. I. (2005), *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N. & Aigrain, S. (2013), ‘Gaussian processes for time-series modelling’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**(1984), 20110550.

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.

Ruxanda, G., Opincariu, S. & Ionescu, S. (2019), ‘Modelling non-stationary financial time series with input warped student t-processes’, *Romanian Journal of Economic Forecasting* **22**(3), 51–61.

Schilling, R. L. (2017), *Measures, integrals and martingales*, Cambridge University Press.

Seeger, M. (1999), Bayesian methods for support vector machines and gaussian processes, Technical report.

Shah, A., Wilson, A. G. & Ghahramani, Z. (2013), Bayesian optimization using student-t processes, *in* 'NIPS Workshop on Bayesian Optimisation'.

Shah, A., Wilson, A. G. & Ghahramani, Z. (2014), 'Student-t processes as alternatives to gaussian processes', *arXiv preprint arXiv:1402.4306* .

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & De Freitas, N. (2015), 'Taking the human out of the loop: A review of bayesian optimization', *Proceedings of the IEEE* **104**(1), 148–175.

Shang, L. & Chan, A. B. (2013), 'On approximate inference for generalized gaussian process models', *arXiv preprint arXiv:1311.6371* .

Sheth, R., Wang, Y. & Khardon, R. (2015), Sparse variational inference for generalized gp models, *in* 'Proceedings of the 32nd International Conference on Machine Learning (ICML-15)', pp. 1302–1311.

Snelson, E. & Ghahramani, Z. (2005), Sparse gaussian processes using pseudo-inputs, *in* 'Advances in neural information processing systems', pp. 1257–1264.

Snoek, J., Larochelle, H. & Adams, R. P. (2012), Practical bayesian optimization of machine learning algorithms, *in* 'Advances in neural information processing systems', pp. 2951–2959.

Solin, A. & Särkkä, S. (2015), State space methods for efficient inference in student-t process regression, *in* 'Artificial Intelligence and Statistics', pp. 885–893.

Tierney, L. & Kadane, J. B. (1986), 'Accurate approximations for posterior moments and marginal densities', *Journal of the american statistical association* **81**(393), 82–86.

Tipping, M. E. & Lawrence, N. D. (2005), 'Variational inference for student-t models: Robust bayesian interpolation and generalised component analysis', *Neurocomputing* **69**(1), 123 – 141. Neural Networks in Signal Processing.
URL: *http://www.sciencedirect.com/science/article/pii/S0925231205001694*

Titsias, M. K. (2009), Variational learning of inducing variables in sparse gaussian processes, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 567–574.

Titsias, M. K. & Lawrence, N. D. (2010), Bayesian gaussian process latent variable model, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 844–851.

Tracey, B. D. & Wolpert, D. (2018), Upgrading from gaussian processes to studentst processes, *in* '2018 AIAA Non-Deterministic Approaches Conference', p. 1659.

Tsallis, C. (1988), 'Possible generalization of boltzmann-gibbs statistics', *Journal of statistical physics* **52**(1-2), 479–487.

Tsallis, C. & Brigatti, E. (2004), 'Nonextensive statistical mechanics: A brief introduction', *Continuum Mechanics and Thermodynamics* **16**(3), 223–235.

van der Herten, J., Couckuyt, I. & Dhaene, T. (2016), 'Hypervolume-based multi-objective bayesian optimization with student-t processes', *arXiv preprint arXiv:1612.00393* .

Vanhatalo, J., Jylänki, P. & Vehtari, A. (2009), Gaussian process regression with student-t likelihood, *in* 'Advances in Neural Information Processing Systems', pp. 1910–1918.

Wainwright, M. J. & Jordan, M. I. (2008), 'Graphical models, exponential families, and variational inference', *Foundations and Trends® in Machine Learning* **1**(1-2), 1–305.

Wasserman, L. (2010), *All of Statistics: A Concise Course in Statistical Inference*, Springer Publishing Company, Incorporated.

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

Williams, C. K. & Seeger, M. (2001), Using the nyström method to speed up kernel machines, *in* 'Advances in neural information processing systems', pp. 682–688.

Wilson, A. G. & Ghahramani, Z. (2010), 'Generalised wishart processes', *arXiv preprint arXiv:1101.0240* .

Wu, Y., Hernández-Lobato, J. M. & Ghahramani, Z. (2014), Gaussian process volatility model, *in* 'Advances in Neural Information Processing Systems', pp. 1044–1052.

Xie, G. & Chen, X. (2017), A heteroscedastic t-process simulation metamodeling approach and its application in inventory control and optimization, *in* '2017 Winter Simulation Conference (WSC)', IEEE, pp. 3242–3253.

Xu, Z., Kersting, K. & von Ritter, L. (2017), Stochastic online anomaly analysis for streaming time series., *in* 'IJCAI', pp. 3189–3195.

Xu, Z., Yan, F. & Qi, Y. (2011), Sparse matrix-variate t process blockmodels, *in* 'Twenty-Fifth AAAI Conference on Artificial Intelligence'.

Yu, K. & Chu, W. (2008), Gaussian process models for link analysis and transfer learning, *in* 'Advances in Neural Information Processing Systems', pp. 1657–1664.

Yu, S., Tresp, V. & Yu, K. (2007), Robust multi-task learning with t-processes, *in* 'Proceedings of the 24th international conference on Machine learning', ACM, pp. 1103–1110.

Zhang, C., Butepage, J., Kjellstrom, H. & Mandt, S. (2018), 'Advances in variational inference', *IEEE transactions on pattern analysis and machine intelligence* .

Zhang, C., Chen, Z., Wang, Z. & Wu, Y. (2018), 'Robust functional anova model with t-process', *arXiv preprint arXiv:1812.07173* .

Zhang, Y. & Yang, Q. (2017), 'A survey on multi-task learning', *arXiv preprint arXiv:1707.08114* .

Zhang, Y. & Yeung, D.-Y. (2010), Multi-task learning using generalized t process, *in* 'Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics', pp. 964–971.