Statistics and Computing (2019) 29:753–773
https://doi.org/10.1007/s11222-018-9836-0

Laplace approximation and natural gradient for Gaussian process regression with heteroscedastic student-*t* model

Marcelo Hartmann[1] · Jarno Vanhatalo[2]

Received: 27 April 2018 / Accepted: 12 October 2018 / Published online: 17 October 2018

Abstract
We propose the Laplace method to derive approximate inference for Gaussian process (GP) regression in the location and scale parameters of the student-*t* probabilistic model. This allows both mean and variance of data to vary as a function of covariates with the attractive feature that the student-*t* model has been widely used as a useful tool for robustifying data analysis. The challenge in the approximate inference for the model, lies in the analytical intractability of the posterior distribution and the lack of concavity of the log-likelihood function. We present the natural gradient adaptation for the estimation process which primarily relies on the property that the student-*t* model naturally has orthogonal parametrization. Due to this particular property of the model the Laplace approximation becomes significantly more robust than the traditional approach using Newton's methods. We also introduce an alternative Laplace approximation by using model's Fisher information matrix. According to experiments this alternative approximation provides very similar posterior approximations and predictive performance to the traditional Laplace approximation with model's Hessian matrix. However, the proposed Laplace–Fisher approximation is faster and more stable to calculate compared to the traditional Laplace approximation. We also compare both of these Laplace approximations with the Markov chain Monte Carlo (MCMC) method. We discuss how our approach can, in general, improve the inference algorithm in cases where the probabilistic model assumed for the data is not log-concave.

Keywords Student-*t* model · Laplace approximation · Heteroscedastic noise · Gaussian processes · Location-scale regression · Fisher information matrix · Natural gradient · Riemannian metric · Approximate inference

1 Introduction

Numerous applications in statistics and machine learning communities are fraught with datasets where some data points appear to strongly deviate from the bulk of the remaining. Usually those points are referred to outliers and in many cases the presence of outliers can drastically change the final result of data analysis (Atkinson and Riani 2000). It is known that, if the probabilistic model for the data is not robust, in the sense of reducing outlier influence, inference for the probabilistic model parameters can be strongly biased and consequently prediction power is reduced (Finetti 1961; West 1984; Atkinson and Riani 2000).

The student-*t* model (Gosset 1908) is a well known three parameter heavy-tailed probabilistic model with the outlier-prone property (robustness) in the sense of Dawid (1973) and O'Hagan (1979). That is, the effect of a group of observations that deviates from the rest of its bulk becomes negligible as that group of observations approaches infinity. The degree of robustness of the model is directly related to the degrees-of-freedom parameter (shape parameter) ν. The smaller the values of ν, the more robust the model is in the presence of outliers (O'Hagan 1979; Fonseca et al. 2008).

Due to the particular outlier-prone property of the student-*t* model, much research has been focused on regression models (linear and non-linear) where the error term is assumed to be distributed according to the student-*t* model.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11222-018-9836-0) contains supplementary material, which is available to authorized users.

✉ Marcelo Hartmann
marcelo.hartmann@helsinki.fi

Jarno Vanhatalo
jarno.vanhatalo@helsinki.fi

[1] Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

[2] Department of Mathematics and Statistics and Organismal and Evolutionary Biology Research Program, University of Helsinki, Helsinki, Finland

Lange et al. (1989), Geweke (1993) and Fernandez and Steel (1999), consider multivariate linear regression models where the error distribution is assumed to follow the student-t probabilistic model. They highlight important aspects such as goodness of fit and inferential difficulties in both Bayesian and non-Bayesian approaches. Tipping and Lawrence (2005) apply variational approximation to the posterior distribution of the regression parameters. Fonseca et al. (2008) obtain the Fisher information matrix and the Jeffrey's prior distribution (Jeffreys 1998) for the vector of parameters in the multivariate regression model with the student-t model. The study of Wang and Yang (2016) is similar to that of Fonseca et al. (2008), but they focus on the reference prior (Bernardo 1979) for the vector of parameters and prove that the posterior distribution for all parameters in the model is improper.

In Gaussian process (GP) regression, the student-t model has been applied with the same aforementioned principles, but instead the focus is on the treatment of the location parameter as an unknown function which follows a Gaussian process prior (Vanhatalo et al. 2009). In this case, the analytical intractability of the posterior distribution with lack of concavity in the log-likelihood function brings difficulties to the estimation process. The early works of Neal (1997) consider the scale-mixture representation (Geweke 1993) which enables more efficient MCMC methods via Gibbs sampling. Vanhatalo et al. (2009) and Jylänki et al. (2011) consider faster approximation methods for the posterior distribution of the Gaussian process, by either considering the Laplace method (Tierney and Kadane 1986; Tierney et al. 1989), variational-Bayes (MacKay 2002; Bishop 2006) or expectation-propagation (EP) (Minka 2001a, b). They point out that, since the log-likelihood function of the student-t model is not log-concave, the posterior distribution of the Gaussian process can possibly present multimodality which makes the implementation of the Laplace method and EP more challenging than with log-concave likelihoods. The variational-Bayes approximation has a stable computational implementation but the approximation underestimates posterior variance (Jylänki et al. 2011). More generally, a detailed analysis carried out by Fernandez and Steel (1999) reveals that parameter inference in both Bayesian and non-Bayesian settings of regression models with student-t errors can be challenging. Firstly because the likelihood can be unbounded for small values of ν and secondly, due to the possibility of multimodality in the likelihood function with certain combinations of the parameters.

This work is developed following the same lines of Vanhatalo et al. (2009). However we use Gaussian process priors to model both the location and the scale parameters of the student-t probabilistic model. This is an important case in which both the mean and variance of the data vary as a function of covariates with the attractive property that the student-t probabilistic model is robust. We focus on Laplace's method to approximate the posterior distribution of the Gaussian process and inferences are also done using it. The difficulty in the estimation process of the parameters of the Laplace approximation, discussed by Vanhatalo et al. (2009) and Jylänki et al. (2011), is circumvented by firstly noting that the location and scale parameters of the student-t model are orthogonal (Cox and Reid 1987; Huzurbazar 1956; Achcar 1994). This particular property of the student-t model will readily allow us to propose an efficient inference algorithm for the Laplace approximation based on the natural gradient of Amari (1998) (also known as the Fisher score algorithm in Statistics).

In this paper, we also propose an alternative Laplace approximation for the posterior distribution of the Gaussian process model. This approximation uses the Fisher information matrix in place of the Hessian matrix of the negative log-likelihood function. Moreover, the alternative Laplace approximation also suggests that the approximate marginal likelihood, which is now based on the Fisher information matrix, offers an alternative way to perform type-II maximum a posteriori (MAP) estimation for the parameters of the probabilistic model and the Gaussian process hyperparameters.

The inference algorithm for estimating the parameters of the Laplace approximation presented here is general. It closely follows the stable implementation of the Laplace approximation for log-concave likelihoods presented by Rasmussen and Williams (2006) with only minor modifications and, hence, generalizes this stable algorithm for general not log-concave likelihoods and multivariate Gaussian process models as well. These general properties are also attractive for other types of models and, hence, we present an example of orthogonal reparametrization for the Weibull probabilistic model and discuss its possible benefits before introducing the GP regression in the heteroscedastic student-t model.

The paper is organized as follows: in Sect. 2 we review some definitions and present one example of orthogonal parametrization for statistical models in the sense of (Jeffreys 1998, p. 207, Sect. 4.31) and Cox and Reid (1987). This concept is needed to introduce an alternative way which can improve inference in Gaussian process models. Section 3 presents the student-t probabilistic model and how the heteroscedastic Gaussian process regression is built. The traditional Laplace approximation with its variant based on the Fisher information matrix is presented in Sect. 4. We also present the approximate marginal likelihood based on the Fisher information in this section. In Sect. 5, we tackle the natural gradient adaptation for finding the parameters of both Laplace approximations. The performance of these approximations and other models are evaluated in Sect. 6, where we examine the quality of these approximations with a simulated example and several real datasets. Section 7 closes the paper with the discussion and concluding remarks.

2 Aspects of orthogonal parametrization for statistical models

This section presents the definition of orthogonal parameters and the equations to find orthogonal parameters for a probabilistic model (Huzurbazar 1956; Cox and Reid 1987). These ideas will be useful later, when we identify that the student-t model directly possesses such a property. One selected example is presented in order to illustrate and clarify concepts of reparametrization in statistical modelling. We end this section by discussing these examples and other aspects of parametric transformations.

During the middle eighties to the end of nineties, a large amount of work in statistics focused in parameter transformation methods for statistical models (Cox and Reid 1987; Achcar and Smith 1990; Achcar 1994; Kass and Slate 1994; MacKay 1998). In both Bayesian and frequentist inference, the performance of numerical procedures and the accuracy of approximation methods (e.g. Laplace approximation) are usually affected by the choice of the parametrization in the probabilistic model. See for example, Cox and Reid (1987), Kass and Slate (1994) and MacKay (1998). In this sense, it is often highly beneficial to identify a new parametrization for a probabilistic model so that the posterior density or the likelihood function are as near as possible to a Gaussian.

To improve the Gaussian approximation for the posterior distribution or the likelihood function, different methods have been proposed in the literature. We cite a few of them here. For instance, the concept of orthogonal reparametrization defined by Jeffreys in 1939 (Jeffreys 1998, p. 207, Sect. 4.31) and later investigated by Huzurbazar (1950), Huzurbazar (1956) and Cox and Reid (1987), can improve the "normality" of the likelihood function by choosing a new parametrization such that the Fisher information matrix is diagonal. This means that the likelihood function can be better behaved in the sense that the distribution of the maximum likelihood estimators are closer to a Gaussian density (see Cox and Reid 1987; Kass and Slate 1994, Sect. 3 for diagnostic measures of nonnormality). Another method, as presented by Achcar (1994), proposes a reparametrization such that the Fisher information is constant. In the Bayesian context, this implies a uniform Jeffreys' prior for the parameters (Box and Tiao 1973).

In what follows, we assume a random variable Y with a probability density function $\pi_{Y|\boldsymbol{\alpha}}(y)$, where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_p]^\top \in \mathcal{A} \subseteq \mathbb{R}^p$ is the set of real continuous parameters. We also consider that the regularity conditions hold for the probabilistic model $\pi_{Y|\boldsymbol{\alpha}}(y)$ (see Schervish 2011, Definition 2.78, p. 111).

Definition 1 *(Fisher information matrix)* Given that the regularity conditions hold, the matrix $I(\boldsymbol{\alpha})$ with elements

$$I_{i,j}(\boldsymbol{\alpha}) = \mathbb{E}_{Y|\boldsymbol{\alpha}}\left[-\frac{\partial^2}{\partial\alpha_i\partial\alpha_j}\log\pi_{Y|\boldsymbol{\alpha}}(Y)\right] \tag{1}$$

is called Fisher information matrix.

Note that the matrix $I(\boldsymbol{\alpha})$ is the expected value of the Hessian matrix of the negative log-density function. By definition, this matrix is symmetric and positive-definite. The regularity conditions are necessary so that the Fisher information matrix can be expressed as in the above form. Besides, the inverse of the Fisher information matrix is a covariance matrix which provides the Cramér-Rao lower bound for the class of unbiased estimators (see Schervish 2011, Sects. 2.3 and 5.1.2 for details)[1].

Definition 2 *(Orthogonal parameters)* The set of parameters $\boldsymbol{\alpha}$, in the probabilistic model $\pi_{Y|\boldsymbol{\alpha}}(y)$, are said to be orthogonal if the Fisher information matrix $I(\boldsymbol{\alpha})$ is diagonal, that is,

$$I_{i,j}(\boldsymbol{\alpha}) = 0 \tag{2}$$

for all i, j such that, $i \neq j$. It can also be said that the probabilistic model $\pi_{Y|\boldsymbol{\alpha}}(\cdot)$ possesses orthogonal parametrization.

Equations for finding orthogonal parameters

Consider a probabilistic model $\pi_{Y|\boldsymbol{\alpha}}(y)$ where the regularity conditions hold. Let the new parametrization $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_p]^\top = \boldsymbol{\eta}(\boldsymbol{\alpha})$ be a bijective differentiable map (with differentiable inverse map) of $\boldsymbol{\alpha}$. Rewrite the probabilistic model of Y in the new parametrization as follows,

$$\log\pi_{Y|\boldsymbol{\eta}}(y) = \log\pi_{Y|\boldsymbol{\alpha}(\boldsymbol{\eta})}(y). \tag{3}$$

The second derivatives of (3) w.r.t η_i and η_j leads to

$$\begin{aligned}\frac{\partial^2}{\partial\eta_i\partial\eta_j}\log\pi_{Y|\boldsymbol{\eta}}(y) = &\sum_{k=1}^{p}\sum_{l=1}^{p}\frac{\partial^2\log\pi_{Y|\boldsymbol{\alpha}(\boldsymbol{\eta})}(y)}{\partial\alpha_k\partial\alpha_l}\\ &\times\frac{\partial\alpha_k}{\partial\eta_i}\frac{\partial\alpha_l}{\partial\eta_j}\\ &+\sum_{k=1}^{p}\frac{\partial\log\pi_{Y|\boldsymbol{\alpha}(\boldsymbol{\eta})}(y)}{\partial\alpha_k}\frac{\partial^2\alpha_k}{\partial\eta_i\partial\eta_j}.\end{aligned} \tag{4}$$

[1] Basically, two main properties on the probabilistic model must hold. Fisrt, the expectation of the score function must be zero, that is $\mathbb{E}(\nabla_{\boldsymbol{\alpha}}\log\pi_{Y|\boldsymbol{\alpha}}(Y)) = 0$, where $\nabla_{\boldsymbol{\alpha}}$ is the gradient operator w.r.t $\boldsymbol{\alpha}$. Second, the support of the distribution $\pi_{Y|\boldsymbol{\alpha}}(y)$, denoted by $\mathbb{A} = \{y : \pi_{Y|\boldsymbol{\alpha}}(y) > 0\}$, must not depend on any component of $\boldsymbol{\alpha}$.

Take the expectation $\mathbb{E}_{Y|\boldsymbol{\alpha}}[\cdot]$ with the negative sign in both sides of Eq. (4). Given that the regularity conditions hold, we have

$$I_{i,j}(\boldsymbol{\eta}) = \sum_{k=1}^{p}\sum_{l=1}^{p} \frac{\partial \alpha_k}{\partial \eta_i}\frac{\partial \alpha_l}{\partial \eta_j} I_{k,l}(\boldsymbol{\alpha}(\boldsymbol{\eta})). \tag{5}$$

If we want the parameters $\boldsymbol{\eta}$ to be orthogonal we set,

$$I_{i,j}(\boldsymbol{\eta}) = 0, \tag{6}$$

for $i \neq j$. In order to find such parametrization we need to solve the system of $p(p-1)/2$ first order partial differential equations, with the α_i, $i = 1, \ldots, p$ as dependent variables.

Example Now we investigate one type of orthogonal parametrization for the Weibull model. Then, we compare the Laplace approximation of the posterior densities in the common parametrization and in the orthogonal parametrization. This comparison will use diagnostic measures of nonnormality to indicate how adequate the Laplace's method can be to approximate the true posterior distribution. These measures are provided by (Sect. 3, Kass and Slate 1994) and they are denoted as $\overline{B}^2$ and B from now on[2]. Let $Y|\alpha_1, \alpha_2 \sim \mathcal{W}(\alpha_1, \alpha_2)$ denote a random variable following the Weibull distribution with common parametrization α_1 and α_2. Then the probability density function of Y is given by,

$$\pi_{Y|\boldsymbol{\alpha}}(y) = \alpha_1\alpha_2(\alpha_2 y)^{\alpha_1 - 1}\exp(-(\alpha_2 y)^{\alpha_1}) \tag{7}$$

for $y, \alpha_1, \alpha_2 \in (0, \infty)$. The Fisher information matrix for this model was obtained by Gupta and Kundu (2006) (in their notation, $\alpha_1 = \beta$ and $\alpha_2 = \theta$, see p. 3131). We now consider that, in the new parametrization $[\eta_1, \eta_2]^\top = \boldsymbol{\eta}(\alpha_1, \alpha_2)$ the Fisher information matrix is diagonal. To find this new parametrization, we start with Eq. (6), which gives

$$\begin{aligned} I_{1,2}(\eta_1, \eta_2) &= \frac{\partial \alpha_1}{\partial \eta_1}\frac{\partial \alpha_1}{\partial \eta_2} I_{1,1}(\boldsymbol{\alpha}) + \frac{\partial \alpha_1}{\partial \eta_1}\frac{\partial \alpha_2}{\partial \eta_2} I_{1,2}(\boldsymbol{\alpha}) \\ &\quad + \frac{\partial \alpha_2}{\partial \eta_1}\frac{\partial \alpha_1}{\partial \eta_2} I_{2,1}(\boldsymbol{\alpha}) + \frac{\partial \alpha_2}{\partial \eta_1}\frac{\partial \alpha_2}{\partial \eta_2} I_{2,2}(\boldsymbol{\alpha}) \\ &= 0. \end{aligned} \tag{8}$$

Now, we fix $\alpha_1 = h_1(\eta_1)$ and choose $\alpha_2 = h_2(\eta_1, \eta_2)$, such that η_1 and η_2 are orthogonal parameters (we also could fix $\alpha_2 = h_2(\eta_2)$ and choose $\alpha_1 = h_1(\eta_1, \eta_2)$, such that η_1 and η_2 are orthogonal parameters). We choose $\alpha_1 = \exp(\eta_1)$. Thus, given the elements of the Fisher information matrix in Gupta and Kundu (2006) (p. 3134), Eq. (8) becomes,

$$\exp(\eta_1) I_{1,2}(\boldsymbol{\alpha}) + \frac{\partial \alpha_2}{\partial \eta_1} I_{2,2}(\boldsymbol{\alpha}) = 0 \tag{9}$$

which leads to $c\partial\eta_1 \exp(-\eta_1) = -\partial\alpha_2/\alpha_2$ and whose one solution is

$$c\exp(-\eta_1) + cz(\eta_2) = \ln \alpha_2 \tag{10}$$

where $c = 1 + \psi(1)$ and $\psi(\cdot)$ is the digamma function. $z(\eta_2)$ is our integration constant and we set $z(\eta_2) = \eta_2$. Rearrange Eq. (10) to get

$$\alpha_2 = \exp\left(c\exp(-\eta_1) + c\eta_2\right). \tag{11}$$

Hence the Weibull probabilistic model with orthogonal parameters is given by,

$$\begin{aligned} \pi_{Y|\boldsymbol{\eta}}(y) &= \exp(\eta_1 + ce^{-\eta_1} + c\eta_2) \\ &\quad \times (\exp(ce^{-\eta_1} + c\eta_2)y)^{\exp(\eta_1)-1} \\ &\quad \times \exp\left(-(\exp(ce^{-\eta_1} + c\eta_2)y)^{\exp(\eta_1)}\right). \end{aligned} \tag{12}$$

The parametrization (η_1, η_2) is now unconstrained (on $\mathbb{R}^2$) with diagonal Fisher information matrix and the transformation $[\eta_1, \eta_2]^\top = [\log\alpha_1, (\log\alpha_2)/c - 1/\alpha_1]^\top$ is one-to-one.

In order to compare the Laplace approximation of the posterior densities in the two parametrizations (α_1, α_2) and (η_1, η_2), we simulated data $Y_i \sim \mathcal{W}(7, 1.5)$ with two different sample sizes, $n = 3$ and $n = 15$. Our prior choice for the orthogonal parametrization is $\eta_1, \eta_2 \overset{i.i.d}{\sim} \mathcal{N}(0, 100)$ and the prior for the original parametrization is $\alpha_1, \alpha_2 \overset{i.i.d}{\sim} \mathcal{N}_+(0, 100)$. Where the notation $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian with parameters μ and σ^2 and $\mathcal{N}_+ \, (\mu, \sigma^2)$ denotes a truncated Gaussian distribution on $\mathbb{R}_+$. Figure 1 displays the countour comparisons between the approximate posterior distribution of $\alpha_1, \alpha_2|\,\mathbf{y}$ and $\eta_1, \eta_2|\,\mathbf{y}$ using Laplace's method with sample sizes $n = 3$ and $n = 15$. We note that, the shape of the posterior distribution in parametrization (η_1, η_2) is visually closer to an independent Gaussian density than the shape of the posterior distribution in parametrization (α_1, α_2). Besides, the measures of posterior nonnormality $\overline{B}^2$ and B, indicate that the parametrization (η_1, η_2) leads to an improvement in the normality of the posterior distribution. This is summarized in Table 1. In the example presented above the new parameters for the statistical model improved the Laplace approximation to the true posterior. As pointed out by MacKay (1998), the effect of the reparametrization in probabilistic models can also lead to better approximation for the marginal likelihood. If the posterior density is well approximated by a Gaussian, then the Laplace approximation for the marginal likelihood is also improved. In real-world scenarios, where complex models impose challenges, it is

[2] The smaller the values of $\overline{B}^2$ and B are, the better the Gaussian approximation to the posterior distribution is. Those measures indicates how "close" the log-posterior is to a quadratic form.

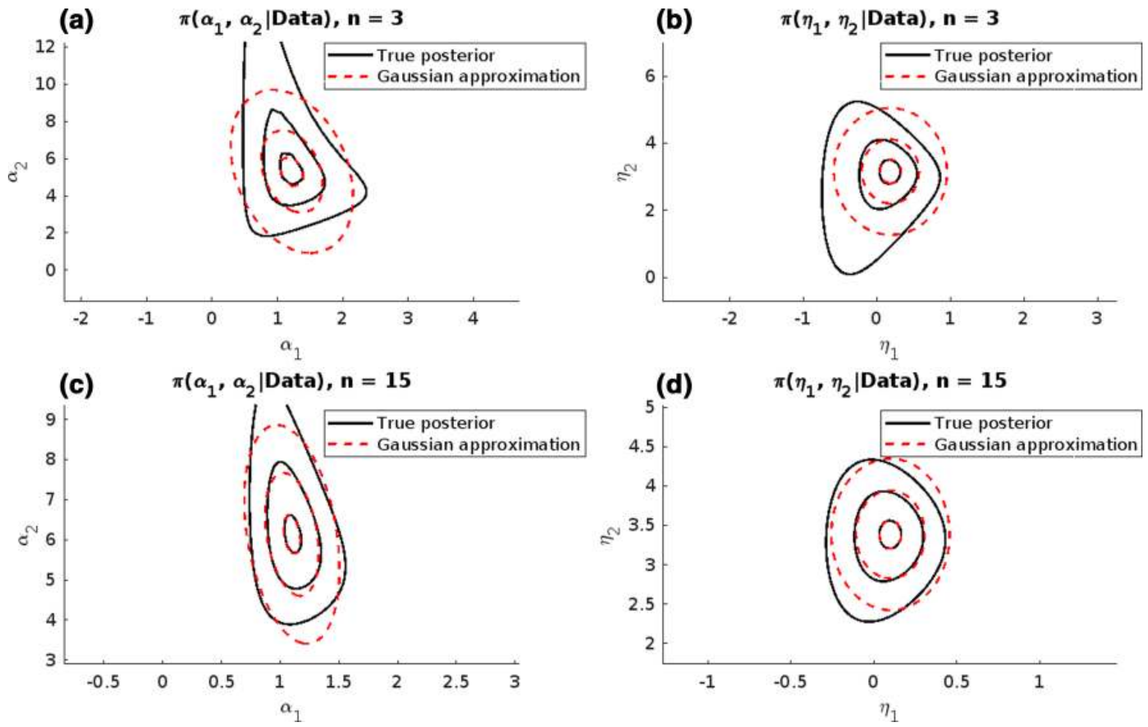

Fig. 1 **a**, **b** shows the approximate posterior distribution of $\alpha_1, \alpha_2 | \mathbf{y}$ and $\eta_1, \eta_2 | \mathbf{y}$ using Laplace's method with $n = 3$. In **c**, **d** we redo the same but with the larger sample size $n = 15$. For the sample size $n = 15$, both approximations are much closer to a Gaussian as seen in **c**, **d**. However, in **d**, the approximate Gaussian is still close to being independent, which does not happen in the approximate posterior for the parametrization (α_1, α_2) in **c**

Table 1 Diagnostics of global assessment of posterior nonnormality for the example with the Weibull model

LP approximation	Sample size	$\overline{B}^2$	B
$\pi_{\text{LP}}(\boldsymbol{\alpha} \mid \mathbf{y})$	$n = 3$	24.06	28.64
	$n = 15$	2.82	1.82
$\pi_{\text{LP}}(\boldsymbol{\eta} \mid \mathbf{y})$	$n = 3$	11.58	3.37
	$n = 15$	2.34	0.66

The values of $\overline{B}^2$ and B for the Laplace approximation of $\pi(\boldsymbol{\eta} \mid \mathbf{y})$ are smaller compared to those values of $\overline{B}^2$ and B for the Laplace approximation of $\pi(\boldsymbol{\alpha} \mid \mathbf{y})$. They indicate better Gaussian approximation to the posterior distribution

surely beneficial to search for a parametrization of the probabilistic model so that approximation methods and numerical procedures can be improved. Hence, the necessity to engineer a complex inference algorithm could be alleviated whereas existing methods could be ameliorated.

3 Gaussian process regression with the heteroscedastic student-*t* model

In this section, we highlight the basic properties of the student-*t* probabilistic model, which are useful to clarify how model building with Gaussian process priors is done. We present how the student-*t* model is parametrized and where the Gaussian process prior is introduced in the parameters of the probabilistic model to build the Gaussian process regression.

3.1 Student-*t* model and basic properties

Let us denote by $Y|\mu, \sigma, \nu \sim \mathcal{S}(\mu, \sigma, \nu)$ a random variable which follows the student-*t* probabilistic model with the location (μ), scale (σ) and degrees-of-freedom (ν) parameters. Then we denote the probability density function of $Y|\mu, \sigma, \nu$ as,

$$\pi(y|\mu, \sigma, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\frac{\nu}{2})\sigma\sqrt{\pi\nu}} \left[1 + \frac{1}{\nu}\frac{(y-\mu)^2}{\sigma^2}\right]^{-\frac{\nu+1}{2}} \tag{13}$$

for $\mu \in \mathbb{R}$, $\sigma > 0$ and $\nu > 0$. The expected value of Y is $\mathbb{E}(Y) = \mu$ which exists only for $\nu > 1$ (otherwise Y is non-integrable). The variance of Y is $\mathbb{V}(Y) = \sigma^2\nu/(\nu-2)$, which only depends on the scale and degrees-of-freedom paramaters. If $\nu \leqslant 2$ then $\mathbb{V}(Y) = \infty$. The degrees-of-freedom parameter controls the "thickness" of the tails in the probability density function (13). The smaller the values of ν, the more robust the student-*t* model is in presence of outliers (O'Hagan 1978). If $\nu \to \infty$ the model (13) converges to a

Gaussian density function with parameters (μ, σ^2) (Fonseca et al. 2008).

The Fisher information matrix for the set of parameters (μ, σ, ν) was obtained by Fonseca et al. (2008) (p. 332, Proof of Theorem 2) and we note that entries (1, 2) and (1, 3) of the Fisher information matrix are zero. Therefore, this means that (μ, σ) and (μ, ν) are pairs of orthogonal parameters. Although the model does not posses full orthogonality, since entry (2, 3) of the Fisher information matrix is non-zero, this particular property of the model will be useful later in Sect. 5. In that section, we tackle a computational implementation to efficiently perform approximate inference with this model.

3.2 Gaussian process regression in the location and scale parameter

Consider the regression model for a set of data $\boldsymbol{Y}^\top = [Y_1 \cdots Y_n] \in \mathbb{R}^n$ that satisfies

$$Y_i = f_1(\mathbf{x}_i) + \varepsilon_i \exp(f_2(\mathbf{x}_i)) \tag{14}$$

for $i = 1, \ldots, n$ where n is the number of observations and $\mathbf{x}_i$ is the i^{th} vector of covariates. Assume that $f_1(\cdot)$ and $f_2(\cdot)$ follow independent zero-mean Gaussian process priors. This implies that $\mathbf{f}_1^\top = [f_1(\mathbf{x}_1) \cdots f_1(\mathbf{x}_n)] \sim \mathcal{N}(\mathbf{0}, K_1)$ and $\mathbf{f}_2^\top = [f_2(\mathbf{x}_1) \cdots f_2(\mathbf{x}_n)] \sim \mathcal{N}(\mathbf{0}, K_2)$. The matrix $\{K_1\}_{i,j} = \text{Cov}(f_1(\mathbf{x}_i), f_1(\mathbf{x}_j)|\gamma_1)$ is the covariance matrix of the process f_1, which depends on a vector of hyperparameters γ_1 and the matrix $\{K_2\}_{i,j} = \text{Cov}(f_2(\mathbf{x}_i), f_2(\mathbf{x}_j)|\gamma_2)$ is the covariance matrix for the process f_2, which depends on a vector of hyperparameters γ_2. Now, let $\varepsilon_i|\nu \overset{i.i.d}{\sim} \mathcal{S}(0, 1, \nu)$. Therefore for each i, the random variable $Y_i|f_1(\mathbf{x}_i), f_2(\mathbf{x}_i), \nu \sim \mathcal{S}(f_1(\mathbf{x}_i), \exp(f_2(\mathbf{x}_i)), \nu)$ has density function given by

$$\pi(y_i|f_1(\mathbf{x}_i), f_2(\mathbf{x}_i), \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\frac{\nu}{2}) \exp(f_2(\mathbf{x}_i))\sqrt{\pi\nu}} \times \left[1 + \frac{1}{\nu}\frac{(y_i - f_1(\mathbf{x}_i))^2}{\exp(2f_2(\mathbf{x}_i))}\right]^{-\frac{\nu+1}{2}}. \tag{15}$$

Let us denote by $\mathbf{y}^\top = [y_1 \cdots y_n]$ the set of measured data, $\mathbf{f}^\top = [\mathbf{f}_1^\top \, \mathbf{f}_2^\top]$ the vector of all the latent function values and $\boldsymbol{\theta}^\top = [\nu \, \gamma_1 \, \gamma_2]$ the collection of all probabilistic model's parameters and covariance functions' hyperparameters. Then, by the Bayes' rule, the conditional posterior distribution for $\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\theta}$ is obtained as

$$\pi(\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\theta}) = \frac{1}{\pi(\mathbf{y} \,|\, \boldsymbol{\theta})} L(\mathbf{y} \,|\, \mathbf{f}, \nu)\mathcal{N}(\mathbf{f}_1 \,|\, \mathbf{0}, K_1) \times \mathcal{N}(\mathbf{f}_2 \,|\, \mathbf{0}, K_2) \tag{16}$$

where

$$L(\mathbf{y} \,|\, \mathbf{f}, \nu) = \prod_{i=1}^{n} \pi(y_i|f_1(\mathbf{x}_i), f_2(\mathbf{x}_i), \nu) \tag{17}$$

is the likelihood function of $\mathbf{f}$ and

$$\pi(\mathbf{y} \,|\, \boldsymbol{\theta}) = \int_{\mathbb{R}^N} L(\mathbf{y} \,|\, \mathbf{f}, \nu)\mathcal{N}(\mathbf{f}_1 \,|\, \mathbf{0}, K_1) \times \mathcal{N}(\mathbf{f}_2 \,|\, \mathbf{0}, K_2)\mathrm{d}\,\mathbf{f} \tag{18}$$

is the marginal likelihood (the normalizing constant). Note that, expression (18) cannot be solved analytically. Besides, posterior expectations and posterior variances are not found in closed-form. The posterior distribution (16) has dimension two times greater than the number of data points ($N = 2n$), which additionally imposes more difficulty in the implementation of any inference algorithm.

4 Approximate inference with the Laplace method

In this section, we present the Laplace method to perform approximate inference. This method is a useful technique for integrals which arises in Bayesian inference (Tierney and Kadane 1986; Tierney et al. 1989; Rue and Martino 2009; Migon et al. 2014). The approximation is analytical and utilizes the Gaussian density function for the approximation. The Gaussian density has desirable analytical properties such as, closed under marginalization and conditioning (Seber and Wild 2003; Seber and Lee 2012). In what follows, we carry out the Laplace approximation for (16) and (18) using a similar approach and notation as in Rasmussen and Williams (2006). We also present the Laplace approximation where the Hessian matrix of the negative log-likelihood function is replaced by its expected value, that is, the Fisher information matrix.

4.1 Laplace approximation

The Laplace approximation is based on the second-order Taylor expansion of $\log \pi(\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\theta})$ around the mode (maximum a posteriori estimate) $\hat{\mathbf{f}} = \arg\max_{\mathbf{f} \in \mathbb{R}^N} \log \pi(\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\theta})$. The method yields a multivariate Gaussian approximation for the conditional posterior distribution (16) given by

$$\pi_{\text{LP}}(\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}\big(\mathbf{f} \,|\hat{\mathbf{f}}, (K^{-1} + \widehat{\mathrm{W}})^{-1}\big). \tag{19}$$

The matrix K is a block diagonal covariance matrix whose blocks are K_1 and K_2, that is, $K = \text{diag}(K_1, K_2)$. We denote $\mathrm{W} = -\nabla\nabla \log L(\mathbf{y} \,|\, \mathbf{f}, \boldsymbol{\theta})$ the Hessian matrix of the negative log-likelihood function with respect to $\mathbf{f}$ and by $\widehat{\mathrm{W}} = \mathrm{W}\,|_{\mathbf{f}=\hat{\mathbf{f}}}$ the matrix W evaluated at $\hat{\mathbf{f}}$. More specifically, the matrix W is a two-diagonal banded matrix whose elements are given in Appendix B.

4.2 Laplace–Fisher approximation

Since the student-t model is regular and possesses orthogonal parametrization with respect to μ and σ, we follow Jeffreys (1998) and Kass and Vaidyanathan (1992) and substitute the nonzero elements of $\widehat{\mathrm{W}}$ by the expected values of W evaluated at $\hat{\mathbf{f}}$ in the traditional Laplace approximation (19)[3]. That it is, we denote the expected value of W by $\mathbb{E}_{\boldsymbol{Y}\,|\,\mathbf{f},\boldsymbol{\theta}}[\mathrm{W}]$ (the Fisher information matrix) and evaluate it at $\hat{\mathbf{f}}$. This is denoted as $\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}]$.

Due to the real-valued random variable $f_2(\mathbf{x}_i)$ in (15), we have to obtain the Fisher information matrix $\mathbb{E}_{\boldsymbol{Y}\,|\,\mathbf{f},\boldsymbol{\theta}}[\mathrm{W}]$ with respect to this specific real-line parametrization. The elements of $\mathbb{E}_{\boldsymbol{Y}\,|\,\mathbf{f},\boldsymbol{\theta}}[\mathrm{W}]$, in this specific parametrization, are given in Appendix B. The Laplace approximation for the conditional posterior distribution (16) is now given by,

$$\pi_{\mathrm{LF}}(\mathbf{f}\,|\,\mathbf{y},\boldsymbol{\theta}) = \mathcal{N}\big(\mathbf{f}\,|\hat{\mathbf{f}}, (K^{-1} + \mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}])^{-1}\big). \qquad (20)$$

In case of the Laplace–Fisher approximation (20), $\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}]$ is diagonal with positive-elements, thus the covariance matrix $(K^{-1}+\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}])^{-1}$ is such that its diagonal elements are always smaller than the diagonal elements of K (element-wise). Hence, the possible effect of larger posterior variance of the latent function values with respect to its prior variance, in the approximation (19), vanishes (see Vanhatalo et al. (2009, Sect. 3.4) and Jylänki et al. (2011), Sect. 5, for details). Kass and Raftery (1995) and Raftery (1996) also point out that the approximation (20) is less precise than the approximation (19), but it remains accurate for great variety of practical purposes.

The computational cost to evaluate (19) is higher than the computational cost to evaluate (20). This comes from the evaluation of the determinant $|K^{-1} + \widehat{\mathrm{W}}|$ which scales to $8\mathcal{O}(n^3)$ computer operations in the approximation (19). In the case of the approximation (20), the evaluation of the determinant $|K^{-1}+\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}(\mathrm{W})|$ scales down to $2\mathcal{O}(n^3)$ computer operations.

4.3 Approximate posterior contraction and outliers

The focus in outlier robust modelling has traditionally been on the behaviour of posterior distribution of f_1 when the observations come increasingly far from the prior mean of f_1. Depending whether the posterior approaches the prior or not, in this situation the probabilistic model assumed for the data can be either outlier prone or not (O'Hagan 1979, 2004; West 1984). However, what has been left for lesser attention is that in some applications it might be of interest to classify individual data points as "normal" or outlier observations. In order to classify observations as outliers we follow Vanhatalo et al. (2009) and check whether the approximate posterior precision of $f_1(\mathbf{x}_i)\,|\,\mathbf{y},\boldsymbol{\theta}$ and $f_2(\mathbf{x}_i)\,|\,\mathbf{y},\boldsymbol{\theta}$, increase (or decrease) as a function of the associated data-point y_i. If there is an increase in the approximate posterior precision, the data-point is considered "normal" observation. If this is not the case, we label y_i as an outlier due to the loss of precision in the approximate posterior distribution.

The above classification is intuitively appealing since, in practice, outlier observations typically originated from gross or systematic errors[4] that either decrease our prior confidence or do not affect it all. In the data analysis using the approach presented in this paper, we will define an outlier and detect it as follows.

Definition 1 *(Outlier & normal observation)* An observation y_i is called normal observation if the following conditions hold,

$$\begin{aligned}&(i)\ P_{i,i} > P_{i,i} - \widehat{\mathrm{W}}_{i,i}\\ &(ii)\ P_{i+n,i+n} > P_{i+n,i+n} - \widehat{\mathrm{W}}_{i+n,i+n}\end{aligned} \qquad (21)$$

where $P_{i,i}$ is the (i,i)-entry of the prior precision matrix $P = K^{-1}$ and $\widehat{\mathrm{W}}_{i,i}$ is the (i,i)-entry of $\widehat{\mathrm{W}}$. Otherwise y_i is called as outlier.

In other words, an observation is normal if the curvature defined by a single negative log-likelihood term at MAP estimate $\hat{\mathbf{f}}$ remains positive.

Theorem 1 (Outlier detection) *Condition* (i) *holds if and only if* $y_i \in \big(\hat{f}_1(\mathbf{x}_i) \pm \exp(\hat{f}_2(\mathbf{x}_i))\nu^{\frac{1}{2}}\big)$. *Condition (ii) holds if and only if* $y_i \neq \hat{f}_1(\mathbf{x}_i)$.

Proof Conditions (i) and (ii) correspond to $\widehat{\mathrm{W}}_{i,i} > 0$ and $\widehat{\mathrm{W}}_{i+n,i+n} > 0$. Let $\hat{z}_i = (y_i - \hat{f}_1(\mathbf{x}_i))/\exp(\hat{f}_2(\mathbf{x}_i))$. Recall the general formulation of W in Appendix B, but instead consider $\widehat{\mathrm{W}}$. Then we have,

$$\begin{aligned}&(i)\ 0 < \widehat{\mathrm{W}}_{i,i}\\ &\quad 0 < \tfrac{1}{[\exp(\hat{f}_2(\mathbf{x}_i))]^2}\big(1+\tfrac{1}{\nu}\big)\Big[\tfrac{2}{(1+\hat{z}_i^2/\nu)^2} - \tfrac{1}{1+\hat{z}_i^2/\nu}\Big]\\ &\quad \hat{z}_i^2 < \nu\\ &\quad y_i \in \big(\hat{f}_1(\mathbf{x}_i) \pm \exp(\hat{f}_2(\mathbf{x}_i))\nu^{\frac{1}{2}}\big).\\ &(ii)\ 0 < \widehat{\mathrm{W}}_{i+n,i+n}\end{aligned}$$

[3] One can imagine that each nonzero elements of W are functions of random variables with respect to each Y_i.

[4] Lack of precision in the measurement instruments, human errors, etc.

$$0 < 2(\tfrac{1}{\nu}+1)\frac{\hat{z}_i^2}{(1+\hat{z}_i^2/\nu)^2}$$
$$0 < \hat{z}_i^2$$
$$y_i \neq \hat{f}_1(\mathbf{x}_i).$$

□

Now, from (i) and (ii) we conclude that, the condition whether a data-point is considered an outlier remains similar as noted by Vanhatalo et al. (2009) and Jylänki et al. (2011). However, the heteroscedastic student-t model now takes into account the variation of the scale parameter as a function of $\hat{f}_2(\mathbf{x}_i)$, as seen in the condition (i) and Theorem 1. For the condition (ii), we surprisingly found that the values $\widehat{\mathrm{W}}_{i+n,i+n}$ are always positive. Therefore, the approximate posterior precision of $f_2(\mathbf{x}_i) \mid \mathbf{y}, \boldsymbol{\theta}$ will be always greater than its prior precision and it does not play any role in the outlier detection methodology presented above.

4.4 Prediction of future outcomes with the Laplace approximation

Let $Y_*|\boldsymbol{\theta}, \mathbf{y}$ be the value of a future outcome under the presence of covariates $\mathbf{x}_*$, given the data and the set of parameters $\boldsymbol{\theta}$. If we use the approximation (19) for (16), the approximate posterior predictive distribution of the vector of latent function values at the new point $\mathbf{x}_*$ is given by (Rasmussen and Williams 2006)

$$\begin{bmatrix} f_1(\mathbf{x}_*) \\ f_2(\mathbf{x}_*) \end{bmatrix} \Big| \boldsymbol{\theta}, \mathbf{y} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1(\mathbf{x}_*) \\ \mu_2(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} \sigma_1^2(\mathbf{x}_*) & \sigma_{12}(\mathbf{x}_*) \\ \sigma_{21}(\mathbf{x}_*) & \sigma_2^2(\mathbf{x}_*) \end{bmatrix} \right) \tag{22}$$

with

$$\begin{bmatrix} \mu_1(\mathbf{x}_*) \\ \mu_2(\mathbf{x}_*) \end{bmatrix} = \mathbf{k}(\mathbf{x}_*) \begin{bmatrix} \nabla_{\mathbf{f}_1} \log L(\mathbf{y}\,|\hat{\mathbf{f}}, \nu) \\ \nabla_{\mathbf{f}_2} \log L(\mathbf{y}\,|\hat{\mathbf{f}}, \nu) \end{bmatrix} \tag{23}$$

and

$$\begin{bmatrix} \sigma_1^2(\mathbf{x}_*) & \sigma_{12}(\mathbf{x}_*) \\ \sigma_{21}(\mathbf{x}_*) & \sigma_2^2(\mathbf{x}_*) \end{bmatrix} = k(\mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*)(K^{-1}+\widehat{\mathrm{W}})^{-1}\mathbf{k}(\mathbf{x}_*)^\top \tag{24}$$

where

$$k(\mathbf{x}_*) = \begin{bmatrix} k_1(\mathbf{x}_*) & 0 \\ 0 & k_2(\mathbf{x}_*) \end{bmatrix} \tag{25}$$

and

$$\mathbf{k}(\mathbf{x}_*) = \begin{bmatrix} \mathbf{k}_1(\mathbf{x}_*) & \mathbf{0}_{1,n} \\ \mathbf{0}_{1,n} & \mathbf{k}_2(\mathbf{x}_*) \end{bmatrix}. \tag{26}$$

$k_1(\mathbf{x}_*) = \mathrm{Cov}(f_1(\mathbf{x}_*), f_1(\mathbf{x}_*)|\gamma_1)$ is the variance of the latent function $f_1(\mathbf{x}_*)$ and $k_2(\mathbf{x}_*) = \mathrm{Cov}(f_2(\mathbf{x}_*), f_2(\mathbf{x}_*)\,|\gamma_2)$ is the variance of the latent function $f_2(\mathbf{x}_*)$. $\mathbf{k}_1(\mathbf{x}_*)$ and $\mathbf{k}_2(\mathbf{x}_*)$ are 1 by n row-vectors which contain the covariances $\mathrm{Cov}(f_1(\mathbf{x}_*), f_1(\mathbf{x}_i)|\gamma_1)$ and $\mathrm{Cov}(f_2(\mathbf{x}_*), f_2(\mathbf{x}_i)\,|\gamma_2)$ for $i = 1, \ldots, n$. More specifically,

$$\mathbf{k}_1(\mathbf{x}_*) = [\mathrm{Cov}(f_1(\mathbf{x}_*), f_1(\mathbf{x}_1)|\gamma_1)\ \ldots \ \mathrm{Cov}(f_1(\mathbf{x}_*), f_1(\mathbf{x}_n)|\gamma_1)] \tag{27}$$

and

$$\mathbf{k}_2(\mathbf{x}_*) = [\mathrm{Cov}(f_2(\mathbf{x}_*), f_2(\mathbf{x}_1)|\gamma_2)\ \ldots \ \mathrm{Cov}(f_2(\mathbf{x}_*), f_2(\mathbf{x}_n)|\gamma_2)]. \tag{28}$$

If we use approximation (20) instead of (19) to approximate the posterior density (16), the approximate posterior predictive distribution (22) has diagonal covariance matrix (24) ($\sigma_{12}(\mathbf{x}_*) = \sigma_{21}(\mathbf{x}_*) = 0$), since $\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}]$ is diagonal. Its mean vector will be equal to (23), given that the mode $\hat{\mathbf{f}}$ remains unchanged for the same $\boldsymbol{\theta}$.

Now, the unconditional expectation (for $\nu > 1$) and unconditional variance (for $\nu > 2$) of the future outcome at $\mathbf{x}_*$ are obtained as

$$\begin{aligned} \mathbb{E}(Y_*|\boldsymbol{\theta}, \mathbf{y}) &= \mathbb{E}[\mathbb{E}(Y_*|f_1(\mathbf{x}_*), f_2(\mathbf{x}_*), \boldsymbol{\theta}, \mathbf{y})] \\ &= \mu_1(\mathbf{x}_*) \end{aligned} \tag{29}$$

and

$$\begin{aligned} \mathbb{V}(Y_*|\boldsymbol{\theta}, \mathbf{y}) &= \mathbb{V}[\mathbb{E}(Y_*|f_1(\mathbf{x}_*), f_2(\mathbf{x}_*), \boldsymbol{\theta}, \mathbf{y})] \\ &\quad + \mathbb{E}[\mathbb{V}(Y_*|f_1(\mathbf{x}_*), f_2(\mathbf{x}_*), \boldsymbol{\theta}, \mathbf{y})] \\ &= \sigma_1^2(\mathbf{x}_*) + \frac{\nu}{\nu-2} e^{2\mu_2(\mathbf{x}_*)+2\sigma_2^2(\mathbf{x}_*)}. \end{aligned} \tag{30}$$

5 On the computational implementation

The main difficulty to make the approximation (19) and (20) useful in practice is in the determination of $\hat{\mathbf{f}}$ for a given $\boldsymbol{\theta}$ (henceforth we refer to it only as $\hat{\mathbf{f}}$). As pointed out by Vanhatalo et al. (2009) and Jylänki et al. (2011), the student-t model is not log-concave and will lead to numerical instability in classical gradient-based algorithms for finding the $\hat{\mathbf{f}}$ if the problem is not approached properly. Besides, the computational algorithm proposed in Rasmussen and Williams (2006) based on Newton's method relies on W being non-negative with log-concave likelihoods. With the student-t model, the log-likelihood is not concave and Newton's method to find the maximum a posteriori $\hat{\mathbf{f}}$ is essentially uncontrolled and not guaranteed to converge (Vanhatalo et al. 2009). In the next subsections, we deal with the problem of finding the maximum a posteriori $\hat{\mathbf{f}}$ and how to choose $\boldsymbol{\theta}$ in the approximations (19) and (20).

5.1 Natural gradient for finding the mode

The problem of finding $\hat{\mathbf{f}}$ is approached by using a variant of standard gradient-based optimization methods called natural gradient adaptation (Amari 1998). The method uses the curved geometry of the parametric space defined by the Riemannian metric (Amari and Nagaoka 2007) which has been shown to improve efficiency and convergence of the computational algorithms (Amari 1998; Honkela et al. 2010). As shown by Amari (1998) and Ollivier et al. (2017), the steepest ascent direction of a smooth function, say $h : \mathcal{M} \subseteq \mathbb{R}^d \to \mathbb{R}$ in a Riemannian manifold $(\mathcal{M}, g)$ where g is the Riemannian metric, is given by the natural gradient defined as

$$\nabla^G h(p) = G^{-1}(p)\nabla h(p) \tag{31}$$

where ∇ is the gradient operator and $G(\cdot)$ is the matrix of metric coefficients (positive-definite matrix $\forall p \in \mathcal{M}$). The evident challenge at this point is how to specify $G(\cdot)$, which still requires specific knowledge of the problem in question. However, it turns out that, in any regular statistical model (Schervish 2011), a Riemannian manifold can be obtained when the parametric space of the probabilistic model is endowed with the Fisher information matrix (Rao 1945; Atkinson and Mitchell 1981; Girolami and Calderhead 2011; Calderhead 2012). That is, the covariance between the elements of the score vector of the probabilistic model (Schervish 2011). Similar ideas have been successfully applied in many optimization techniques and MCMC methods. See for example works by Jennrich and Sampson (1976), Amari (1998), Honkela et al. (2010), Girolami and Calderhead (2011), Calderhead (2012), Ollivier et al. (2017) and Hasenclever et al. (2017).

Now, the iterative procedure to find $\hat{\mathbf{f}}$ via the natural gradient is given by (Amari 1998; Polak 2006)

$$\mathbf{f}^{\text{new}} = \mathbf{f} + G(\mathbf{f})^{-1}(\nabla \log L(\mathbf{y} \mid \mathbf{f}, \nu) - K^{-1}\mathbf{f}) \tag{32}$$

where G is the matrix of metric coefficients. At this point, note that Eq. (32) is very similar to the Newton-Raphson updating scheme (see Rasmussen and Williams 2006, equation (3.18))

$$\mathbf{f}^{\text{new}} = \mathbf{f} + M(\mathbf{f})^{-1}(\nabla \log L(\mathbf{y} \mid \mathbf{f}, \nu) - K^{-1}\mathbf{f}) \tag{33}$$

where $M(\mathbf{f}) = (K^{-1} + \mathrm{W})$. More specifically, in the case of (32), $G(\mathbf{f})$ is, by construction, always positive-definite (Amari and Nagaoka 2007; Rao 1945; Schervish 2011), while $M(\mathbf{f})$ in (33) may not be, since W is not positive-definite in the domain of the negative logarithm of the likelihood function of the student-t model. Now, $G(\cdot)$ has not been specified yet and as we adopt a Bayesian approach, we would like to consider the geometry of the posterior distribution, which includes the information in the likelihood and in the prior distribution. A possible Riemmanian metric with prior knowledge was used by Girolami and Calderhead (2011) and Calderhead (2012)) (p. 87, Sect. 4.1.4, equation 4.2) and for our settings, their matrix $G(\mathbf{f})$ is given by

$$\begin{aligned} G(\mathbf{f}) &= \mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}\big[-\nabla\nabla_{\mathbf{f}} \log \pi(Y, \mathbf{f} \mid \boldsymbol{\theta})\big] \\ &= \mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}\big[-\nabla\nabla_{\mathbf{f}} \log L(Y \mid \mathbf{f}, \nu)\big] \\ &\quad + \mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}\big[-\nabla\nabla_{\mathbf{f}} \log \mathcal{N}(\mathbf{f}_1 \mid \mathbf{0}, K_1)\mathcal{N}(\mathbf{f}_2 \mid \mathbf{0}, K_2)\big] \\ &= \mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[\mathrm{W}] + \mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[K^{-1}]. \end{aligned} \tag{34}$$

Note again that, $\mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[\mathrm{W}]$ is the expected value of W, that is, the Fisher information matrix which has been already obtained in Sect. 4. The second term $\mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[K^{-1}] = K^{-1}$ is the inverse of the block diagonal covariance matrix of the Gaussian process prior. Hence, Eq. (34) simplifies to $G(\mathbf{f}) = \mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[\mathrm{W}] + K^{-1}$. Plug $G(\mathbf{f})$ into Eq. (32) and rearrange to get

$$\begin{aligned} \mathbf{f}^{\text{new}} &= (K^{-1} + \mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[\mathrm{W}])^{-1} \\ &\quad \times \big(\mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[\mathrm{W}]\,\mathbf{f} + \nabla \log L(\mathbf{y} \mid \mathbf{f}, \nu)\big) \end{aligned} \tag{35}$$

which has the same structural properties as the Newton-update in (2006, equation 3.18) for the binary Gaussian process classification case. Moreover, since $\mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[\mathrm{W}]$ is diagonal, the stable formulation of the computation algorithm provided in Rasmussen and Williams (2006) to find $\hat{\mathbf{f}}$ is straightforwardly applied by replacing W with its expected value, that is $\mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[\mathrm{W}]$ (see Rasmussen and Williams 2006, Sect. 3.4.3, p. 45). Note that, at each iteration proposed in (35), the computational cost to calculate the inverse of $(K^{-1} + \mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[\mathrm{W}])$ is $2\mathcal{O}(n^3)$ instead of $8\mathcal{O}(n^3)$ with the Newton-update (33).

In case of the Gaussian process regression with the homocedastic student-t model ($f_2(\mathbf{x})$ is constant), the GPML (Rasmussen and Nickisch 2010) and GPstuff (Vanhatalo et al. 2013) software packages use the stabilized Newton algorithm to find $\hat{\mathbf{f}}$. In this approach the Newton direction $\mathbf{d} = \big(K^{-1} + \max(\mathbf{0}, \mathrm{diag}(\mathrm{W}))\big)^{-1} \nabla \log \pi(\mathbf{f} \mid \mathbf{y}, \boldsymbol{\theta})$ is used (see Jylänki et al. 2011, p. 3231, Sect. 3.2). We see that the natural gradient adaptation uses the Fisher information matrix $\mathbb{E}_{Y \mid \mathbf{f}, \boldsymbol{\theta}}[\mathrm{W}]$ in place of $\max(\mathbf{0}, \mathrm{diag}(\mathrm{W}))$.

5.2 Approximate marginal likelihood and parameter adaptation

Note that, in Eq. (16), the set of parameters $\boldsymbol{\theta}$ is fixed but unknown. Rasmussen and Williams (2006) proposes a value for $\boldsymbol{\theta}$ such that $\log \pi(\mathbf{y} \mid \boldsymbol{\theta})$ (18) is maximized. Gibbs (1997) and Vanhatalo et al. (2009) considers that, even though $\boldsymbol{\theta}$ is fixed, it is treated as an unknown quantity and so prior distributions are chosen for all its components. Our choice follows the latter and we use the maximum a posterior estimate (MAP) of $\boldsymbol{\theta} \mid \mathbf{y}$ to choose $\boldsymbol{\theta}$, that is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}\in\Theta}{\arg\max} \log \pi(\boldsymbol{\theta} \mid \mathbf{y}) \\ &= \underset{\boldsymbol{\theta}\in\Theta}{\arg\max} \log \pi(\mathbf{y} \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}) \end{aligned} \tag{36}$$

where Θ is a parametric space and $\pi(\boldsymbol{\theta})$ is the prior distribution for $\boldsymbol{\theta}$. A closed-form expression for (18) is not known when the likelihood takes its form from the student-t model. For this reason we use Laplace's method to also approximate the marginal likelihood (18) (Rasmussen and Williams 2006; Rue and Martino 2009; Vanhatalo et al. 2009). The logarithm of the marginal likelihood (18) is then approximated as

$$\begin{aligned} q_{\text{LP}}(\mathbf{y} \mid \boldsymbol{\theta}) = \log L(\mathbf{y} \,|\hat{\mathbf{f}}, \nu) - \tfrac{1}{2}\hat{\mathbf{f}}^{\top} K^{-1}\hat{\mathbf{f}} \\ - \tfrac{1}{2}\log |I_N + \widehat{\mathrm{W}}K|. \end{aligned} \tag{37}$$

However, since W is not guaranteed to be positive-definite, direct evaluation of the above approximate log marginal likelihood can be numerically unstable due to the last term in (37) (see Vanhatalo et al. 2009, Sect. 4.2; Jylänki et al. 2011, Sect. 5.4 for more details).

Similary, as a byproduct of the approximation (20), the approximate log marginal likelihood in the case of the Laplace–Fisher approximation is given by

$$\begin{aligned} q_{\text{LF}}(\mathbf{y} \mid \boldsymbol{\theta}) = \log L(\mathbf{y} \,|\hat{\mathbf{f}}, \nu) - \tfrac{1}{2}\hat{\mathbf{f}}^{\top} K^{-1}\hat{\mathbf{f}} \\ - \tfrac{1}{2}\log |I_N + (\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}])^{\frac{1}{2}} K (\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}])^{\frac{1}{2}}| \end{aligned} \tag{38}$$

where the last term in (38) is now stable to compute since $\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}]$ is positive-definite. The formulation of the approximate log marginal likelihood (38) is the same as the one presented in Rasmussen and Williams (2006) (see equation 3.32, p. 48), which makes its use more attractive due to its stable computational implementational. Besides, in Eqs. (37) and (38), $\hat{\mathbf{f}}$ depends on $\boldsymbol{\theta}$, and the matrices $\widehat{\mathrm{W}}$ and $\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}]$, depends on $\boldsymbol{\theta}$ through $\hat{\mathbf{f}}$. Rasmussen and Williams (2006) present closed-form derivatives of (37) w.r.t $\boldsymbol{\theta}$, which can as well be applied in the case of (38). Hence, their stable computational implementation is fully applicable to the case where we set $\boldsymbol{\theta}$ by maximizing the approximate marginal posterior (36) using (38) (see Rasmussen and Williams 2006, Sect. 5.5.1, p. 125).

In Appendix B, we present the gradient $\nabla \log L(\mathbf{y} \mid \mathbf{f},\, \nu)$, W and $\mathbb{E}_{\,\boldsymbol{Y}\,|\,\mathbf{f},\boldsymbol{\theta}}[\mathrm{W}]$. The derivatives of $\nabla \log L(\mathbf{y} \mid \mathbf{f},\, \nu)$ and W, w.r.t $\mathbf{f}_1$, $\mathbf{f}_2$ and ν are presented in the supplementary material. The derivatives of $\mathbb{E}_{\boldsymbol{Y}\,|\,\mathbf{f},\boldsymbol{\theta}}[\mathrm{W}]$ w.r.t $\mathbf{f}_1$, $\mathbf{f}_2$ and ν are not given in this paper since they are simple to calculate.

Note also that the evaluation of (37) is slower than the evaluation of (38). The reason is the same as presented in the end of Sect. 4.2. While the calculation of $|I_N + \widehat{\mathrm{W}}K|$ costs $8\mathcal{O}(n^3)$ computer operations in (37), $|I_N + (\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}])^{\frac{1}{2}} K (\mathbb{E}_{\boldsymbol{Y}\,|\hat{\mathbf{f}},\boldsymbol{\theta}}[\mathrm{W}])^{\frac{1}{2}}|$ costs $2\mathcal{O}(n^3)$ computer operations in (38).

6 Experiments

This section illustrates the Laplace approximation (19) and the Laplace–Fisher approximation (20) for the GP regression with the heteroscedastic student-t model. We present two simulated examples to pinpoint practical differences whether conducting data analysis with the traditional Laplace approximation or with the alternative Laplace–Fisher approximation. The predictive performance of both Laplace approximations are compared with several datasets presented in the literature. These comparisons include the gold-standard MCMC method. In the MCMC approximation, the samples from the posterior distribution (16) are obtained with the elliptical slice sampler Murray et al. (2010). Moreover, the predictive comparisons also include the GP regression with the homoscedastic student-t model (Vanhatalo et al. 2009) and the GP regression with the heteroscedastic Gaussian model ($\nu \to \infty$). For the example with simulated data in Sect. 6.2, we also use the methodology proposed by Murray and Adams (2010) to obtain samples from the posterior distribution of the Gaussian process hyperparameters and the degrees-of-freedom parameter ν.

The choice of prior distributions for the Gaussian process hyperparameters and the degrees-of-freedom parameter is discussed in the next subsection, where we also specify the covariance functions for the latent processes f_1 and f_2.

6.1 Priors for the GP hyperparameters and degrees-of-freedom parameter

When the parameter $\nu \to 0$, the student-t model will present higher robustness, in which case the likelihood function may be unbounded and so difficult to evaluate (see Fernandez and Steel 1999; Wang and Yang 2016). Moreover, Gaussian process priors for the function values of the regression model introduce great flexibility into the model's fit capability. For which reason the model can perform poorly and present overfitted regression functions if the prior distributions are not carefully chosen for the covariance function hyperparameters (Simpson et al. 2017).

With the goal of alleviating such scenarios, our choice in the prior distributions for the Gaussian process hyperparameters and for the degrees-of-freedom ν, follows the penalised model-component principles (PC), introduced by Simpson et al. (2017). Under the hierarchical nature of the modelling approach, the main idea of PC-priors rests in the fact that the prior should avoid overly complex models (see desideratas and principles in Simpson et al. (2017).

In this sense, we prefer prior distributions for GP hyperparameters that penalize too flexible regression functions and too small values of the degrees-of-freedom ν. Instead of imposing strict limits to the degrees of freedom, e.g. $\nu > 1$ as was done by Vanhatalo et al. (2009) and Jylänki et al. (2011), we let $\nu \in (0, \infty)$ and choose a prior that penalizes values of $\nu < 2$ (the variance (30) for the data does not exist in this case). Note that, it is the variance of a future outcome (30) that tells us about the uncertainty around the expected value (29) (point estimate). In all subsequent experiments, we will consider that $\nu \sim$ Gumbel-II$(1, \lambda)$ and whose prior distribution is given by

$$\pi(\nu) = \lambda \nu^{-2} e^{-\lambda/\nu}, \tag{39}$$

where $\lambda = -2 \log \mathbb{P}(\nu < 2)$ with $\mathbb{P}(\nu < 2) = 0.1$.

For the covariance function of the latent processes f_1 and f_2, we assume the squared exponential function given by

$$\text{Cov}(f_j(\mathbf{x}), f_j(\mathbf{x}')|\sigma_j^2, \boldsymbol{\ell}_j) = \sigma_j^2 \exp\left(-\tfrac{1}{2}\mathrm{d}^\top D(\boldsymbol{\ell}_j)\mathrm{d}\right) \tag{40}$$

where $\mathrm{d} = \mathbf{x} - \mathbf{x}'$ and $D(\boldsymbol{\ell}_j) = \text{diag}(\ell_{j,1}^2, \ldots, \ell_{j,p}^2)^{-1}$ for $j = 1, 2$. The notation $\text{diag}(\ell_{j,1}^2, \ldots, \ell_{j,p}^2)$ stands for a diagonal matrix whose elements are given by the squared length-scale hyperparameters, $\ell_{j,1}^2, \ldots, \ell_{j,p}^2$. We assume a covariate space with dimension p, accordingly to each experiment in the next subsections. The vector of hyperparameters is then given by $[\sigma_1^2\ \boldsymbol{\ell}_1\ \sigma_2^2\ \boldsymbol{\ell}_2]$ where $\boldsymbol{\ell}_1 = [\ell_{1,1}, \cdots, \ell_{1,p}]^\top$ and $\boldsymbol{\ell}_2 = [\ell_{2,1}, \cdots, \ell_{2,p}]^\top$. The choice of the priors for the hyperparameters combines the weakly informative principle from Gelman (2006) and the PC-priors (Simpson et al. 2017). In this case, the density function for the hyperparameters should give more weight to rigid regression functions (straight lines, planes, etc). That is, the prior should favour small variability of the sample functions in the GP prior and more strongly correlated function values in order to avoid overfitting (see Gelman 2006; Simpson et al. 2017, more for details). In this sense, we assume that $\sigma_1^2, \sigma_2^2 \overset{i.i.d}{\sim} \mathcal{S}_+(0, \sigma_f^2, 4)$ for relatively small values of σ_f^2 and $\boldsymbol{\ell}_1$, $\boldsymbol{\ell}_2 \overset{i.i.d}{\sim}$ inv-$\mathcal{S}_+(0, 1, 4)$. The notation $\mathcal{S}_+$ stands for student-t distribution truncated on $\mathbb{R}_+$ and inv-$\mathcal{S}_+$ stands for inverse student-t distribution truncated on $\mathbb{R}_+$. These prior distributions are respectively given by,

$$\pi(\sigma_j^2) = \frac{2\Gamma\left(\frac{5}{2}\right)}{\Gamma(2)\sigma_f\sqrt{4\pi}}\left[1 + \frac{(\sigma_j^2)^2}{4\sigma_f^2}\right]^{-2.5} \tag{41}$$

and

$$\pi(\ell_{j,r}) = \frac{2\Gamma\left(\frac{5}{2}\right)}{\Gamma(2)\sqrt{4\pi}}\left[1 + \frac{1}{4\ell_{j,r}^2}\right]^{-2.5}\frac{1}{\ell_{j,r}^2} \tag{42}$$

for $j = 1, 2$ and $r = 1, \ldots, p$. The specific choice for σ_f^2 will be given for each dataset in the subsequent sections. With this choice, the prior densities (41) favours small variability of the Gaussian process prior for the function values by assigning higher prior density values to regions where the parameters σ_j^2 is smaller. With respect to the length-scale hyperparameters, the priors (42) induce greater values of length-scales which increase the dependency between the function values. These choices tend to alleviate substantially problems related to hyperparameter identifiability. See for example the work by Vanhatalo et al. (2010), Simpson et al. (2017) and Fuglstad et al. (2018).

6.2 Simulated data with simple regressions

In this first experiment, we simulated a dataset tailored to work well with both approximate marginal likelihoods (37) and (38). We then compared the Laplace approximations (19) and (20) where we set $\boldsymbol{\theta}$ by using (36) either with (37) or (38) respectively. We consider that the probabilistic model for the data is given by (15) where $f_1(\cdot)$ and $f_2(\cdot)$ are unidimensional real-valued functions given by

$$\begin{aligned} f_1(x) &= 0.3 + 0.4x + 0.5\cos(2.7x) + \tfrac{1.1}{1+x^2} \\ f_2(x) &= 0.5\cos(0.5\pi x) + 0.52\cos(\pi x) - 1.2. \end{aligned} \tag{43}$$

Hence, the data generative mechanism is $Y|f_1(x), f_2(x), \nu \sim \mathcal{S}(f_1(x), \exp(f_2(x)), \nu)$ and the number of covariates is $p = 1$. To simulate the dataset, we choose $\nu = 2.5$ and different sample sizes $n \in \{10, 150\}$ with equally spaced points in the interval $A = (-4.5, 4.5)$. The set of parameters is $\boldsymbol{\theta} = [\nu\ \sigma_1^2\ \ell_{1,1}\ \sigma_2^2\ \ell_{2,1}]$ and we choose $\sigma_f^2 = 10$. Note that the vector $\boldsymbol{\theta}$ is *a priori* unknown to us, hence its value will be set by maximizing (in the log scale) the approximate marginal posterior (36) when using the approximate marginal likelihood (37) or (38). When we use (37), the value of $\boldsymbol{\theta}$ which maximizes (36) is denoted as $\hat{\boldsymbol{\theta}}_{\text{LP}}$. When using (38) in the Eq. (36), we denote the previous as $\hat{\boldsymbol{\theta}}_{\text{LF}}$.

We compare approximations (19) and (20) by means of the estimated regression functions $f_1(\cdot), f_2(\cdot)|\,\mathbf{y}, \hat{\boldsymbol{\theta}}_{\text{LP}}$ and $f_1(\cdot), f_2(\cdot)|\,\mathbf{y}, \hat{\boldsymbol{\theta}}_{\text{LF}}$, and local approximate posterior predictive distributions $f_1(x_*), f_2(x_*)|\,\mathbf{y}, \hat{\boldsymbol{\theta}}_{\text{LP}}$ and $f_1(x_*), f_2(x_*)|\,\mathbf{y}, \hat{\boldsymbol{\theta}}_{\text{LF}}$ at $x_* = 0$. The natural gradient adaptation

Table 2 Maximum a posteriori estimates with different approximate marginal likelihoods and sample sizes

	Sample	ν	σ_1^2	$\ell_{1,1}$	σ_2^2	$\ell_{2,1}$
MAP	$n = 10$	6.84	3.59	4.37	0.69	0.83
$\hat{\boldsymbol{\theta}}_{\text{LP}}$	$n = 150$	3.59	3.87	0.93	2.07	0.79
MAP	$n = 10$	3.77	3.77	4.32	0.56	1.08
$\hat{\boldsymbol{\theta}}_{\text{LF}}$	$n = 150$	3.57	3.86	0.93	2.01	0.72
MCMC	$n = 10$	8.07	2.96	2.17	1.47	1.15
$\mathbb{E}(\boldsymbol{\theta} \mid \mathbf{y})$	$n = 150$	2.74	2.44	0.90	1.39	1.02

The estimate $\hat{\boldsymbol{\theta}}_{\text{LP}}$ corresponds to the value of $\boldsymbol{\theta}$ which maximizes (36) when using marginal likelihood estimate (37). The estimate $\hat{\boldsymbol{\theta}}_{\text{LF}}$ corresponds to the value of $\boldsymbol{\theta}$ when marginal likelihood estimate (38) is used in (36). The last row shows the posterior mean of $\boldsymbol{\theta} \mid \mathbf{y}$ estimated via MCMC approximation

(Eq. (35)) is used to find $\hat{\mathbf{f}}$ for both approximations (19) and (20). In both cases, the approximate marginal likelihoods (37) and (38) were stable to evaluate. Hence, $\hat{\boldsymbol{\theta}}_{\text{LP}}$ and $\hat{\boldsymbol{\theta}}_{\text{LF}}$ were obtained without any problems.

Table 2 displays the maximum a posterior estimate for $\boldsymbol{\theta}$ using the approximate marginal likelihoods (37) and (38). The posterior mean of $\boldsymbol{\theta} \mid \mathbf{y}$ obtained with MCMC methods is also presented. Figures 2 and 3 show the model performance for the Laplace approximations (19) and (20) for $\boldsymbol{\theta}$ fixed as $\hat{\boldsymbol{\theta}}_{\text{LP}}$ and $\hat{\boldsymbol{\theta}}_{\text{LF}}$ respectively. In Fig. 2, the Laplace approximation (19) gives slightly different performance when compared to (20) in the case where $n = 10$. In the case where $n = 150$, the approximations (19) and (20) completely match. Figure 3 shows the result of the same experiment, however with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{LF}}$ for both approximations. We note that, for $n = 10$, the approximations (19) and (20) show very similar performance. When $n = 150$, the approximations match again. In general, the Laplace approximations (19) and (20) are slightly different for small sample sizes, but very similar when the number of data points increase, no matter whether $\boldsymbol{\theta}$ is chosen as $\hat{\boldsymbol{\theta}}_{\text{LP}}$ or $\hat{\boldsymbol{\theta}}_{\text{LF}}$.

In Fig. 4, we compare the approximate posterior predictive distributions (22) with both Laplace approximations and with the MCMC approximation. We consider $x_* = 0$ with the sample size $n = 10$. In the first row of Fig. 4, all approximations of (16) consider $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{LP}}$. The Laplace–Fisher approximation estimates similar variances in both cases. In the second row of Fig. 4, we redo the same, but instead we set $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{LF}}$. In this case, the difference between the approximate posterior predictive distributions whether considering the traditional Laplace approximation (19) or the Laplace–Fisher

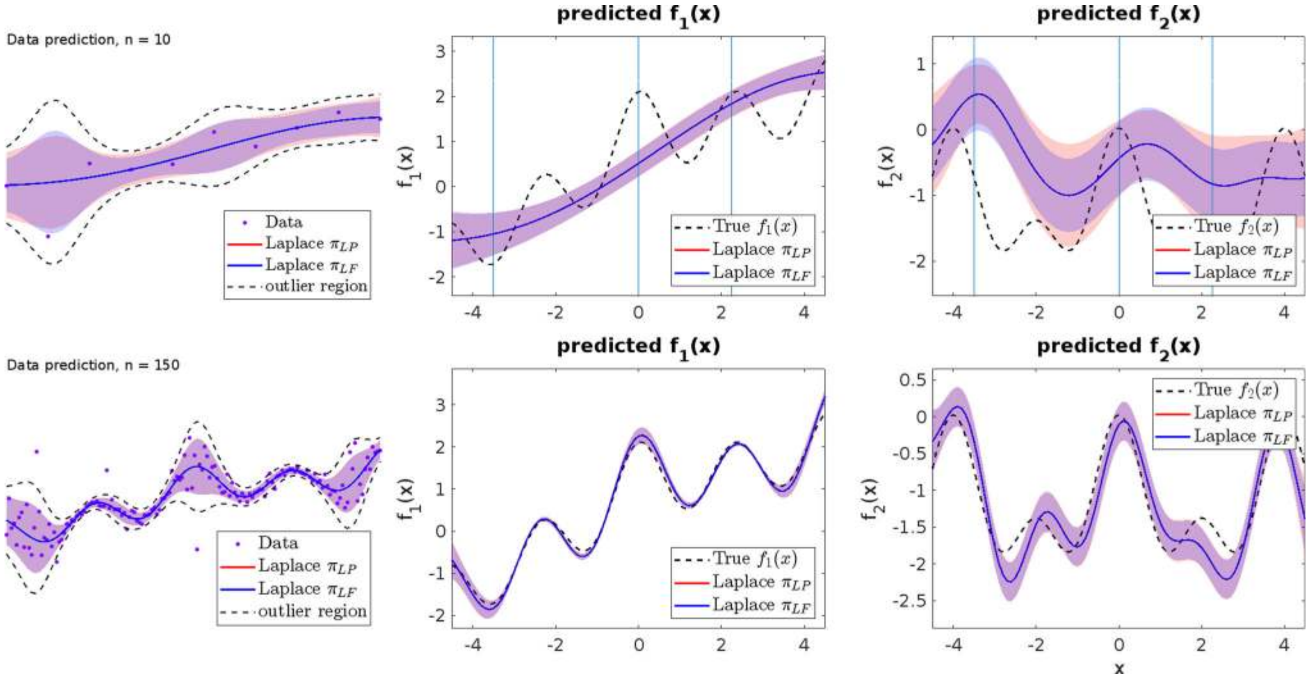

Fig. 2 Comparisons between the Laplace approximations (19) and (20) where $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{LP}}$. In the first row, the number of data points in the x-axis is $n = 10$ and in the second row it is $n = 150$. The red color shows the approximate posterior predictive distributions for the regression functions $f_1(x)$ and $f_2(x)$ with the Laplace approximation (19). The blue color shows the approximate posterior predictive distribution for the regression functions $f_1(x)$ and $f_2(x)$ with the Laplace–Fisher approximation (20). Note that, since $\boldsymbol{\theta}$ is the same in both Laplace approximations, the MAP estimate $\hat{\mathbf{f}}$ is also the same for both approximations. In the second row, with a larger dataset, both approximations completely match. The dashed line on the left-hand side plots (above and below) represent the outlier region defined via Theorem 1. (Color figure online)

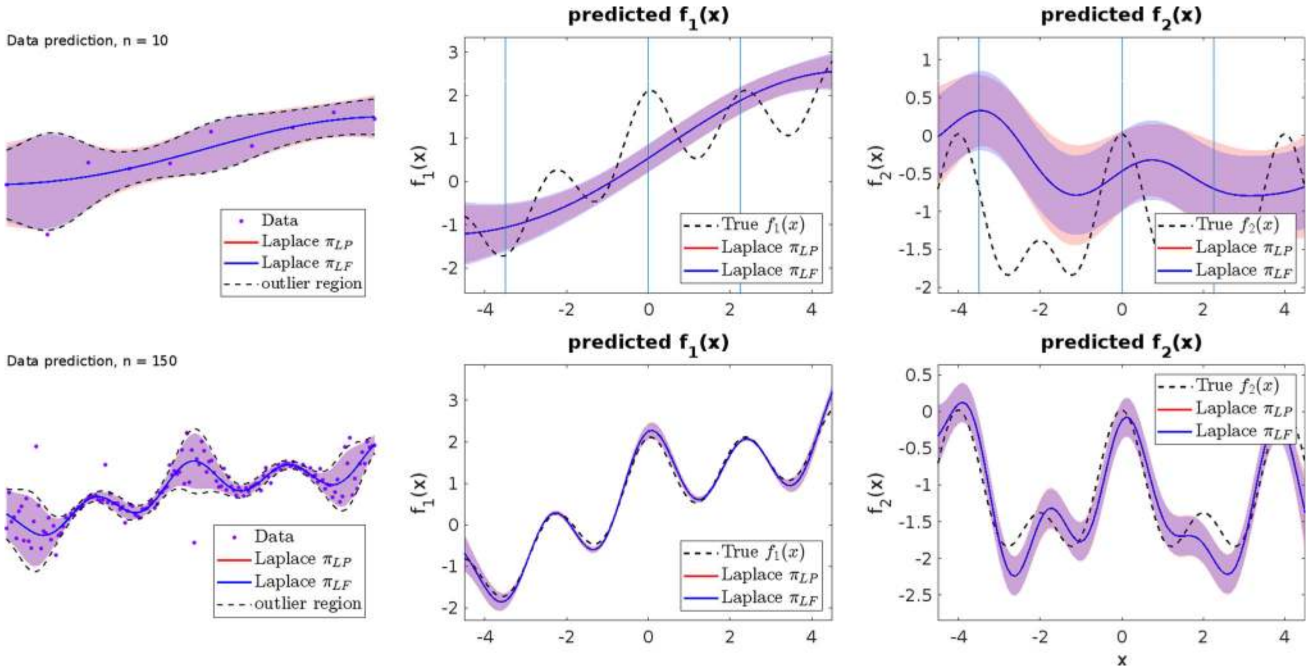

Fig. 3 Comparison between the Laplace approximations (19) and (20) where $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{LF}}$. In the first row, the sample size is $n = 10$ and in the second row the sample size is $n = 150$. The red color shows the approximate posterior predictive distribution for the regression functions $f_1(x)$ and $f_2(x)$ with the Laplace approximation (19). The blue color shows the approximate posterior predictive distribution for the regression functions $f_1(x)$ and $f_2(x)$ with the Laplace approximation (20). Note that, since $\boldsymbol{\theta}$ is the same in both Laplace approximations, the MAP estimate $\hat{\mathbf{f}}$ is the same for both approximations. In the first row the approximations are very similar and in the second row the approximations completely match each other again. On the left-hand side, the plots with dashed lines represents the outlier region defined via Theorem 1. (Color figure online)

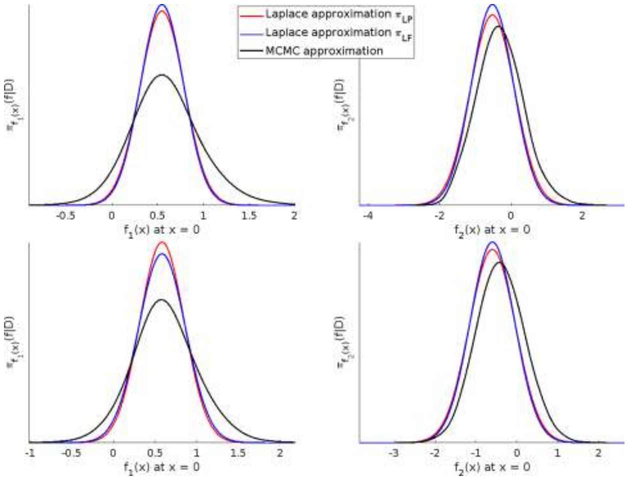

Fig. 4 Local comparisons between the approximate posterior predictive marginal distributions of the Laplace approximations (19), (20) and MCMC approximation. The upper row displays the approximate posterior predictive marginal distributions for $f_1(x)|\,\mathbf{y}, \boldsymbol{\theta}$ and $f_2(x)|\,\mathbf{y}, \boldsymbol{\theta}$ at $x = 0$ where $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{LP}}$. The lower row displays the approximate posterior predictive marginal distribution for $f_1(x)|\,\mathbf{y}, \boldsymbol{\theta}$ and $f_2(x)|\,\mathbf{y}, \boldsymbol{\theta}$ at $x = 0$ where $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{LF}}$. In all cases the dataset in the same and the sample size is $n = 10$

approximation (20) is also almost unnoticed. The MCMC approximation for the true marginal predictive distribution also shows very similar performance.

6.3 A simulation study in parameter/hyperparameter inference

In order to evaluate the goodness of the approximate marginal likelihoods (37) and (38) in the estimation of $\boldsymbol{\theta}$, we conduct the following simulation experiment. First, we choose covariance function (40) with $p = 1$ for both of the processes, f_1 and f_2, and fix true values for the parameter ν and Gaussian process hyperparameters as $\boldsymbol{\theta}_{\text{true}} = [\nu\ \sigma_1^2\ \ell_{1,1}\ \sigma_2^2\ \ell_{2,1}] = [2.5\ 1\ 1.5\ 1\ 1.5]^\top$. Then, we replicate $M = 1000$ times the following experiment for sample sizes $n \in \{10, 30, 50, 150\}$. For a sample size n, we uniformly select n points in the interval A and generated random realizations of the processes f_1 and f_2 at those particular points in A. Then, for each of those realizations, we use Eq. (14) with $\varepsilon_i|\nu \sim \mathcal{S}(0, 1, \nu)$, to generate sample data $Y_i = y_i$. In the prior (41) for the hyperparameters σ_j^2, $j = 1, 2$, the paramater $\sigma_f^2 = 5$.

Table 3 The perfomance of the MAP estimate of the hyperparameters $\boldsymbol{\theta}$ based on the different approximate marginal likelihoods (37) and (38) in the simulation study

$M = 1000$	Laplace		Laplace–Fisher	
n	$\tilde{B}_{\text{LP}}$	$\tilde{E}_{\text{LP}}$	$\tilde{B}_{\text{LF}}$	$\tilde{E}_{\text{LF}}$
15				
ν	3.23	3.89	1.46	2.19
σ_1^2	0.44	1.29	0.41	0.75
$\ell_{1,1}$	0.30	1.30	0.44	1.39
σ_2^2	0.51	1.80	0.49	0.61
$\ell_{1,2}$	0.32	1.52	0.27	1.07
30				
ν	2.05	2.96	1.31	1.95
σ_1^2	0.40	0.91	0.39	0.73
$\ell_{1,1}$	0.29	1.18	0.42	1.29
σ_2^2	0.34	0.82	0.37	0.59
$\ell_{1,2}$	0.21	1.45	0.22	1.05
50				
ν	1.33	2.32	1.21	1.58
σ_1^2	0.39	0.80	0.37	0.72
$\ell_{1,1}$	0.28	1.16	0.34	1.21
σ_2^2	0.33	0.80	0.32	0.57
$\ell_{1,2}$	0.18	1.01	0.16	0.99
150				
ν	0.43	1.06	0.40	1.06
σ_1^2	0.32	0.65	0.32	0.66
$\ell_{1,1}$	0.24	0.97	0.26	0.98
σ_2^2	0.27	0.67	0.25	0.56
$\ell_{1,2}$	0.15	0.93	0.14	0.85

The quantities $\tilde{B}_{\text{LP}}$, $\tilde{B}_{\text{LF}}$, $\tilde{E}_{\text{LP}}$ and $\tilde{E}_{\text{LF}}$ denote the biases and root mean squared errors for each sample size and each element of $\boldsymbol{\theta}$

Let's denote by $\hat{\boldsymbol{\theta}}_{\text{LP}}^{(m)}$ and $\hat{\boldsymbol{\theta}}_{\text{LF}}^{(m)}$, the MAP estimate obtained when using, respectively, marginal likelihood approximations (37) and (38) for the mth replicate data with $m = 1, \ldots, M$. To evaluate the performance of the approximate marginal likelihoods in parameter and hyperparameter inference, two criteria were considered. The bias and the root mean squared error (rmse), which are defined respectively as $\tilde{B}_{\text{LP}} = \frac{1}{M}(\sum_{m=1}^{M} \hat{\boldsymbol{\theta}}_{\text{LP}}^{(m)} - \boldsymbol{\theta}_{\text{true}})$, $\tilde{B}_{\text{LF}} = \frac{1}{M}(\sum_{m=1}^{M} \hat{\boldsymbol{\theta}}_{\text{LF}}^{(m)} - \boldsymbol{\theta}_{\text{true}})$, $\tilde{E}_{\text{LP}} = (\frac{1}{M}\sum_{m=1}^{M}(\hat{\boldsymbol{\theta}}_{\text{LP}}^{(m)} - \boldsymbol{\theta}_{\text{true}})^2)^{\frac{1}{2}}$ and $\tilde{E}_{\text{LF}} = (\frac{1}{M} \sum_{m=1}^{M} (\hat{\boldsymbol{\theta}}_{\text{LF}}^{(m)} - \boldsymbol{\theta}_{\text{true}})^2)^{\frac{1}{2}}$. Table 3 summarizes the results of this experiment. For small sample size, we note marginally smaller bias $\tilde{B}_{\text{LF}}$ compared to $\tilde{B}_{\text{LP}}$. Also, we note smaller root mean squared error $\tilde{E}_{\text{LF}}$ compared to $\tilde{E}_{\text{LP}}$ when considering the degrees-of-freedom ν. As the number n increases, those difference tend to disappear, thus showing that the inferential procedure over $\boldsymbol{\theta}_{\text{true}}$ performed using (36) via (37) or (38) will provide similar results.

6.4 Predictive performance on real datasets

In this section, the performance of the Laplace approximation (19) and (20) for the Gaussian process regression with the heteroscedastic student-t model is examined with real data. Experiments with five datasets were conducted to evaluate the performance of different models in terms of predictive performance (see Appendix A for a short description of the datasets).

We provide comparisons for the predictive performance of the Laplace approximations (19) and (20) with the Gaussian process regression with homoscedastic student-t model (Vanhatalo et al. 2009) and the Gaussian process regression with heteroscedastic Gaussian model. We also compare these models with the MCMC approximation of (16) in the heteroscedastic student-t model. These models are respectively denoted by HT-ST-LP, HT-ST-LF, HM-ST-LP, HT-G-LP and HT-ST-MCMC.

The predictive performance of the models were compared by splitting the datasets into training data (n_{Train}) and test data (n_{Test}). Three measures of predictive quality are proposed to compare all the models. 1) The absolute mean error $\mathcal{R}_1 = 1/n_{\text{Test}} \sum_{i=1}^{n_{\text{Test}}} |y_{i,*} - \mathbb{E}(Y_{i,*}|\,\boldsymbol{\theta}, \mathbf{y})|$. 2) The root mean squared error $\mathcal{R}_2 = (1/n_{\text{Test}} \sum_{i=1}^{n_{\text{Test}}} (y_{i,*} - \mathbb{E}(Y_{i,*}|\,\boldsymbol{\theta}, \mathbf{y}))^2)^{\frac{1}{2}}$. 3) The log predictive density statistic $\mathcal{P} = \sum_{i=1}^{n_{\text{Test}}} \log \pi(y_{i,*}|\,\boldsymbol{\theta}, \mathbf{y})$ (Gelman et al. 2014, the greater the value of $\mathcal{P}$, the better the model is for the data analysis, see][]. This experiment was replicated 20 times with randomly selected training and test data. For each replicate, we obtain the values of $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{P}$, after which we average each predictive statistics and obtain their mean values. This is denoted as $\overline{\mathcal{R}}_1$, $\overline{\mathcal{R}}_2$ and $\overline{\mathcal{P}}$.

For all the models, inference on $\boldsymbol{\theta}$ is done by maximizing the approximate marginal posterior (36) using the approximate marginal likelihood (37) and (38) of corresponding Laplace approximation (19) and (20) and $\hat{\mathbf{f}}$ is searched by the natural gradient method (35). For model HM-ST-LP, we set $\boldsymbol{\theta}$ by maximizing the approximate marginal likelihood as done by Vanhatalo et al. (2009) and $\hat{\mathbf{f}}$ is obtained via the stabilized Newton algorithm (see Jylänki et al. 2011, p. 3231, Sect. 3.2). Model HT-G-LP was implemented as HT-ST-LP with fixed $\nu = 5 \times 10^4$. In this case the student-t model practically corresponds to the Gaussian model.

Table 4 shows the predictive performance for all the models in terms of average predictive statistics. We see that all the models perform similarly in terms of $\overline{\mathcal{R}}_1$ and $\overline{\mathcal{R}}_2$. Model HT-G-LP shows slighlty worse predictive performance with respect to $\overline{\mathcal{R}}_1$, and this is reasonable. The Gaussian model for the data is not an outlier-prone model, if some test data point $y_{i,*}$ is possibly an outlier, then the predictive value $\mathbb{E}(Y_{i,*}|\,\boldsymbol{\theta}, \mathbf{y})$ will try to match that point. This is not the case with the student-t model for data. Note that, both of the statistics $\mathcal{R}_1$ and $\mathcal{R}_2$ use the discrepancy between $y_{i,*}$ and

Table 4 Model comparisons. $\overline{\mathcal{R}}_1$ stands for the average values of absolute mean squared error, $\overline{\mathcal{R}}_2$ is the average root mean squared error and $\overline{\mathcal{P}}$ is the average log-predictive density statistics. The number n_{Train} is the sample size, n_{Test} is the number of test points and p is the number of covariates for each dataset. The second column shows the models examined in the experiments and the last column shows the priors chosen for the variance parameter of the Gaussian processes. The model abbreviations stand for: 1) HM-ST-LP - Laplace approximation for the GP regression with the homoscedastic student-t model, 2) HT-ST-LP - Laplace approximation for the GP regression with the heteroscedastic student-t model, 3) HT-ST-LF - Laplace–Fisher approximation for the GP regression with the heteroscedastic student-t model, 4) HT-G-LP - Laplace approximation for the GP regression with the heteroscedastic Gaussian model and 5) HT-ST-MCMC - MCMC approximation for the GP regression with the heteroscedastic student-t model

Dataset	Models	$\overline{\mathcal{R}}_1$	$\overline{\mathcal{R}}_2$	$\overline{\mathcal{P}}$	Priors
Neal	HM-ST-LP	0.12	0.26	54.20	$\sigma_1^2, \sigma_2^2 \overset{i.i.d}{\sim} \mathcal{S}_+(0, 15, 4)$
$n_{\text{Train}} = 100$	HT-ST-LP	0.11	0.28	54.42	$\boldsymbol{\ell}_1, \boldsymbol{\ell}_2 \overset{i.i.d}{\sim}$ inv-$\mathcal{S}_+(0, 1, 4)$
$n_{\text{Test}} = 100$	HT-ST-LF	0.11	0.28	54.46	$\nu \sim$ Gumbel-II$(1, 4.60)$
$p = 1$	HT-G-LP	0.13	0.28	44.12	
	HT-ST-MCMC	0.12	0.27	57.45	
Motorcycle	HM-ST-LP	17.83	24.28	– 305.17	$\sigma_1^2, \sigma_2^2 \overset{i.i.d}{\sim} \mathcal{S}_+(0, 500, 4)$
$n_{\text{Train}} = 67$	HT-ST-LP	18.43	24.67	– 300.02	$\boldsymbol{\ell}_1, \boldsymbol{\ell}_2 \overset{i.i.d}{\sim}$ inv-$\mathcal{S}_+(0, 1, 4)$
$n_{\text{Test}} = 66$	HT-ST-LF	18.24	24.40	– 302.28	$\nu \sim$ Gumbel-II$(1, 4.60)$
$p = 1$	HT-G-LP	18.43	24.69	– 303.83	
	HT-ST-MCMC	18.18	24.29	– 290.73	
Boston	HM-ST-LP	0.28	0.46	– 56.56	$\sigma_1^2, \sigma_2^2 \overset{i.i.d}{\sim} \mathcal{S}_+(0, 15, 4)$
$n_{\text{Train}} = 253$	HT-ST-LP	0.28	0.46	– 48.40	$\boldsymbol{\ell}_1\, \boldsymbol{\ell}_2 \overset{i.i.d}{\sim}$ inv-$\mathcal{S}_+(0, 1, 4)$
$n_{\text{Test}} = 253$	HT-ST-LF	0.28	0.46	– 46.22	$\nu \sim$ Gumbel-II$(1, 4.60)$
$p = 13$	HT-G-LP	0.27	0.45	– 48.90	
	HT-ST-MCMC	0.27	0.45	−39.65	
Friedman	HM-ST-LP	1.69	2.15	– 217.44	$\sigma_1^2, \sigma_2^2 \overset{i.i.d}{\sim} \mathcal{S}_+(0, 15, 4)$
$n_{\text{Train}} = 100$	HT-ST-LP	1.63	2.09	– 213.34	$\boldsymbol{\ell}_1\, \boldsymbol{\ell}_2 \overset{i.i.d}{\sim}$ inv-$\mathcal{S}_+(0, 1, 4)$
$n_{\text{Test}} = 100$	HT-ST-LF	1.67	2.03	– 213.64	$\nu \sim$ Gumbel-II$(1, 4.60)$
$p = 5$	HT-G-LP	1.68	2.15	– 223.15	
	HT-ST-MCMC	1.60	2.06	−210.62	
Compressive	HM-ST-LP	6.19	9.22	– 317.14	$\sigma_1^2, \sigma_2^2 \overset{i.i.d}{\sim} \mathcal{S}_+(0, 500, 4)$
$n_{\text{Train}} = 515$	HT-ST-LP	7.00	8.64	– 305.06	$\boldsymbol{\ell}_1\, \boldsymbol{\ell}_2 \overset{i.i.d}{\sim}$ inv-$\mathcal{S}_+(0, 1, 4)$
$n_{\text{Test}} = 515$	HT-ST-LF	6.30	8.37	– 312.71	$\nu \sim$ Gumbel-II$(1, 4.60)$
$p = 8$	HT-G-LP	9.28	10.19	– 316.14	
	HT-ST-MCMC	6.21	9.24	– 292.36	

$\mathbb{E}(Y_{i,*} | \boldsymbol{\theta}, \mathbf{y})$. In the case of $\mathcal{R}_2$, this discrepancy is squared, which penalizes the predictive quality of the model too much if the discrepancy for some particular data points are too high (or too small). With respect to the statistic $\mathcal{R}_1$, there is no harsh penalization. Hence the models HT-ST-LP, HT-ST-LF shows slightly better predictive performance when compared to HT-G-LP. Overall, the model HM-ST-LP shows slightly better predictive performance with respect to $\overline{\mathcal{R}}_1$ and $\overline{\mathcal{R}}_2$, this means that this model tends to overfit to a small degree, since it does not allow for heteroscedasticity in the data.

Model HT-ST-LF has almost the same predictive performance as model HT-ST-LP with respect to $\overline{\mathcal{R}}_1$ and $\overline{\mathcal{R}}_2$. Based on the simulation studies in Sect. 6.3 this was expected. The number of data points in all datasets is relatively high, so the estimate of $\boldsymbol{\theta}$, whether from (37) or (38) are expected to be similar. This implies similar $\hat{\mathbf{f}}$ in both of the approximations (19) and (20). Hence, according to Eqs. (23) and (29), the predictive measures $\overline{\mathcal{R}}_1$ and $\overline{\mathcal{R}}_2$ are close for both models HT-ST-LP and HT-ST-LF. The performance of HM-ST-LP has also shown good predictive performance with respect to $\overline{\mathcal{R}}_1$ and $\overline{\mathcal{R}}_2$ for all datasets, but it does not present good values with respect to statistic $\overline{\mathcal{P}}$. Note, however, that $\mathcal{R}_1$ and $\mathcal{R}_2$ are measures of dispersion based on the estimate $\mathbb{E}(Y_{i,*} | \boldsymbol{\theta}, \mathbf{y})$, which does not take into account the uncertainty encoded in the predictive distribution of $Y_{*,i} | \mathbf{y}, \boldsymbol{\theta}$.

With respect to the $\overline{\mathcal{P}}$ statistics, model HT-ST-LP dominates when compared to the models HT-G-LP and HM-ST-LP. For the model HT-ST-LF, the statistics $\overline{\mathcal{P}}$ is only slightly smaller compared to HT-ST-LP. These outcomes are still quite reasonable. The $\mathcal{P}$ statistics calculates the value of the predictive density for a future outcome at the measured values. If the random variable $Y_{*,i}\,|\,\mathbf{y},\boldsymbol{\theta}$ has small variance, its predictive density does not cover much region of the sample space, therefore, if the mode of the predictive density function is distant from the observed value, the density $\pi(y_{*,i}\,|\,\mathbf{y},\boldsymbol{\theta})$ is small. On the other hand, if $Y_{*,i}\,|\,\mathbf{y},\boldsymbol{\theta}$ has greater variance, its predictive density covers greater regions of the sample space, therefore, even if the mode is distant from the observation, the density function of $Y_{*,i}\,|\,\mathbf{y},\boldsymbol{\theta}$ evaluated at $y_{*,i}$ will be higher. This is exactly what happens with the models HT-ST-LP and HT-ST-LF. The predictive distributions of $Y_{*,i}\,|\,\mathbf{y},\boldsymbol{\theta}$, with models HT-ST-LP and HT-ST-LF have similar expectations since, in both approximate posteriors (19) and (20), the estimates for $\hat{\mathbf{f}}$ are similar. However, since the approximate variance of $\mathbf{f}\,|\,\mathbf{y},\boldsymbol{\theta}$ is generally higher in the approximation (19), $\pi(y_{*,i}\,|\,\mathbf{y},\boldsymbol{\theta})$ will be wider (see Eq. (30)), hence leading to a higher $\overline{\mathcal{P}}$ statistics.

The aforementioned behaviour is also analogous to the case where the Gaussian model for the data is assumed, since the Gaussian density function will always have thinner tails compared to the student-*t* model. Once we have chosen the probabilistic approach to conduct the data analysis, the statistic $\mathcal{P}$ may be considered a better suitable measure of predictive quality since it takes into account the degrees of uncertainty which is encoded in the posterior predictive distributions (Bernardo and Smith 1994; Vehtari and Ojanen 2012).

As expected, the HT-ST-MCMC model presents very similar results with respect to the predictive measures $\overline{\mathcal{R}}_1$ and $\overline{\mathcal{R}}_2$ compared to all other models. This model also presents the best predictive performance with respect to the predictive measure $\overline{\mathcal{P}}$. This is also confirmatory in the sense of the previous explanation about $\mathcal{P}$, since this model approximates the true predictive distributions $\pi(y_{*,i}\,|\,\boldsymbol{\theta},\mathbf{y})$ better than Laplace's method.

Even though model HT-ST-LF only had showed slightly worse predictive performance compared to HT-ST-LP, model HT-ST-LF still provided very similar results in all predictive measures. This result suggests that the Laplace–Fisher approximation (20), based on the Fisher information matrix in place of the Hessian matrix of the negative log-likelihood function, can also be a good candidate to approximate the posterior density (16).

6.5 Computational performance in simulated and real data

The optimization of (16) based on the natural gradient provided clear benefits compared to the traditional approach. In our experiments, the natural gradient adaptation was always able to converge, whereas the Newton's method was very sensitive to initial values of $\mathbf{f}$ and to the values of the vector $\boldsymbol{\theta}$ (a general discussion on this is given by, e.g.,Vanhatalo et al. 2009; Jylänki et al. 2011). This is not unexpected. In the Newton update (33), $(K^{-1}+\mathrm{W})^{-1}$ is not always positive-definite (as it should be in the traditional Newton's method) and if the initial value for $\mathbf{f}$ is far from the mode of (16), the Newton's method will not converge. For this reason and in order to be able to compare to Laplace–Fisher with the traditional Laplace approximation, we used the natural gradient in the real data experiments to find the MAP of estimate of (19) if the Newton-Raphson algorithm did not converge.

In all the experiments with simulated and real data, the initial values for the latent function values are $\mathbf{f}_1 = \mathbf{0}$ and for $\mathbf{f}_2 = \mathbf{3}$ (a vector where each element is equal to 3). This choice means that $\sigma(\mathbf{x}) = \exp(3) \approx 20$, in other words, at initialization the data has "large" variance compared to the prior variance of f_1 everywhere in the covariate space. This also avoids possible multimodality of the posterior density (19) since the initial values for $\sigma(\mathbf{x})$ are relatively high (see the analysis done by Vanhatalo et al. 2009, Sect. 3.4; Jylänki et al. 2011, Sect. 5, second paragraph). This helped the regular Newton method for most of the datasets, but for the motorcycle dataset, where the data varies from -130 to 100 (see Silverman 1985, Fig. 2), the initial value for $\mathbf{f}_2$ is far from optimal. However, the natural gradient approach did not encounter any trouble with this data either. In summary, we did not encounter any problems in optimization of (19) with any dataset using the natural gradient adaptation.

Table 5 compares computational speed in performing parameter and hyperparameter estimation with full real data sets whether using the approximate marginal likelihoods (37) or (38) in (36). Optimization using Eq. (38) has smaller computational times compared to the (37) in all cases. This is expected, since the computational effort to evaluate the approximate marginal likelihood (38) is less than the com-

Table 5 Elapsed CPU times (in minutes) to obtain the MAP estimate w.r.t to the approximate marginal posterior in (36) whether using (37) or (38). In this experiment, all the data sets have been considered in full. The last column comprises the time (in minutes) needed to find $\hat{\boldsymbol{\theta}}$ in (36) using either the marginal likelihood approximation (37) or (38)

Full dataset	Marginal likelihood	CPU times (in minutes)
Neal	q_{LP}	0.52
$n_{\text{Train}} = 200,\ p = 1,$	q_{LF}	0.27
Motorcycle	q_{LP}	0.22
$n_{\text{Train}} = 133,\ p = 1$	q_{LF}	0.19
Boston	q_{LP}	1.62
$n_{\text{Train}} = 506,\ p = 13$	q_{LF}	1.35
Friedman	q_{LP}	1.02
$n_{\text{Train}} = 200,\ p = 5$	q_{LF}	0.42
Compressive	q_{LP}	22.60
$n_{\text{Train}} = 1030,\ p = 8$	q_{LF}	11.84

putation effort to evaluate (37) and, in this sense, tend to converge faster. This is in line with the number of computational operations presented in the end of the Sects. 4.2 and 5.2, and the Sect. 5.1, where the natural gradient is presented and discussed.

Comparison to the MCMC approximation has not been included here since it has surely higher computational times to obtain a representative sample of $\boldsymbol{\theta} \,|\, \mathbf{y}$ and $\mathbf{f} \,|\, \mathbf{y}$. Moreover, the computational load in calculating the predictive statistics $\mathcal{P}$ with MCMC methods can be time-consuming and high. First because there is a inversion of the diagonal block matrix K for each MCMC iteration. This costs $2\mathcal{O}(n^3)$ computer operations. Second, we have to perform numerical integration over $\mathbb{R}^2$ to obtain the value of $\pi(y_{*,i} \,|\, \boldsymbol{\theta}, \mathbf{y})$. Specifically, for a MCMC sample $\mathbf{f}'$ from the true posterior (16), we get the posterior predictive distribution of $f_1(\mathbf{x}_*)$ and $f_2(\mathbf{x}_*)$ conditioned on $\mathbf{f}'$ as,

$$f_1(\mathbf{x}_*),\, f_2(\mathbf{x}_*) \,|\, \mathbf{f}', \boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{k}(\mathbf{x}_*)K^{-1}\mathbf{f}', k(\mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*)K^{-1}\mathbf{k}(\mathbf{x}_*)^\top\right) \tag{44}$$

where $k(\mathbf{x}_*)$ and $\mathbf{k}(\mathbf{x}_*)$ are respectively given by (25) and (26). Then the posterior predictive distribution for the test data conditioned on $\mathbf{f}'$ is given by,

$$\pi(y_{*,i} \,|\, \boldsymbol{\theta}, \mathbf{f}') = \int\int_{\mathbb{R}^2} \pi(y_{*,i} \,|\, f_1(\mathbf{x}_*), f_2(\mathbf{x}_*), \nu) \times \pi(f_1(\mathbf{x}_*), f_2(\mathbf{x}_*) \,|\, \mathbf{f}', \boldsymbol{\theta})\mathrm{d}f_1(\mathbf{x}_*)\mathrm{d}f_2(\mathbf{x}_*) \tag{45}$$

which is calculated via numerical integration since no closed-form is known for the above expression. Finally, the value of $\pi(y_{*,i} \,|\, \mathbf{y}, \boldsymbol{\theta})$ is obtained by taking the mean value of $\pi(y_{*,i} \,|\, \boldsymbol{\theta}, \mathbf{f}')$ w.r.t all MCMC samples (Monte Carlo estimate). This is due to the fact that we can write $\pi(y_{*,i} \,|\, \mathbf{y}, \boldsymbol{\theta})$ as a expected value,

$$\pi(y_{*,i} \,|\, \mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbb{R}^N} \pi(y_{*,i} \,|\, \boldsymbol{\theta}, \mathbf{f})\pi(\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\theta})\mathrm{d}\,\mathbf{f} = \mathbb{E}_{\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\theta}}[\pi(y_{*,i} \,|\, \boldsymbol{\theta}, \mathbf{f})]. \tag{46}$$

In the previous experiments presented in Table 4, the number of MCMC samples were, in total, 6200. We used 200 samples as burn-in and took each second sample to form the MCMC chain of 3000 samples. After that, we use (44), (45) and (46) to obtain the value of predictive statistics $\mathcal{P}$.

7 Concluding remarks and discussion

Recently, many approximative methods have been proposed to approximate the posterior distribution of the Gaussian process model with homoscedastic student-*t* probabilistic model for the data (see Vanhatalo et al. 2009; Jylänki et al. 2011). With a non log-concave likelihood, those methods require special treatment by tuning certain values in the mechanism of the estimation process to incur convergence in the computational algorithm (see Vanhatalo et al. 2009, Sect. 4.2 and Jylänki et al. 2011, Sect. 4).

In this paper, we extended the models presented by Vanhatalo et al. (2009) and Jylänki et al. (2011), by additionally modelling the scale parameter of the student-*t* model with a Gaussian process prior. In general, the Gaussian process regression with the heteroscedastic student-*t* model has been shown to perform very well. With respect to the statistic $\mathcal{P}$, it has shown the best performance when compared to known models such as the Gaussian process regression with the homocesdastic student-*t* model of Vanhatalo et al. (2009) and the Gaussian process regression with the heteroscedastic Gaussian model for the data.

Saul et al. (2016) introduced chained Gaussian processes, which uses variational methods to approximate the posterior distribution of the Gaussian process regression with the heteroscedastic student-*t* model for the data. Additionally, their approach allow the use of large datasets via sparse GP approximations (Snelson and Ghahramani 2005; Titsias 2009; Hensman et al. 2015). Our methodology could easily be extended to include sparse GP approximation as well. However, in this work, we have focused in the aspects of parametrization in statistical models and exploited the orthogonal parametrization of the student-*t* model. Due to this particular property, we have recovered well-known algorithms (Rasmussen and Williams 2006) to perform approximate inference with the Laplace approximation and with the Laplace–Fisher approximation.

Although the Laplace approximation based on the Fisher information matrix has already been proposed in the literature, its application in the context of Gaussian process regression has not been investigated yet. In our case, with the student-*t* model, this approximation delivered very similar results in the experiments with simulated and real datasets. Thus, the methodology presented here provides an alternative approximation method for Gaussian process regression. This also concerns approximation methods with other probabilistic models and parametrization in the same lines of Kuss and Rasmussen (2005) and Nickisch and Rasmussen (2008). Moreover, the choice of the parameters $\boldsymbol{\theta}$ through the approximate marginal likelihood $q_{\mathrm{LF}}(\mathbf{y} \,|\, \boldsymbol{\theta})$ (38), can also be seen as a new way of adapting the unknown covariance function hyperparameters and the probabilistic model parameters. In difficult cases, where the dataset leads to difficult evaluation of $q_{\mathrm{LP}}(\mathbf{y} \,|\, \boldsymbol{\theta})$ (37), one can always use $q_{\mathrm{LF}}(\mathbf{y} \,|\, \boldsymbol{\theta})$ to choose $\boldsymbol{\theta}$ and use the Laplace approximation $\pi_{\mathrm{LF}}(\mathbf{f} \,|\, \mathbf{y}, \boldsymbol{\theta})$ (19) if wanted.

We also point out that, there are two possible avenues of improvement in the optimization of (19) via natural gradient. Firstly, as studied by Hang and Amari (1998), Amari (1998) and Fukumizu and Amari (2000) the natural gradient adaptation is a robust learning rule in the sense that the method might avoid plateaus and local maxima. Hence, the natural gradient may be better suited than Newton's method given that (16) is not guaranteed to be unimodal. Secondly, as empirically evaluated by Honkela et al. (2010), the natural gradient might increase the convergence speed of the optimization method and there might be stability with the simplification of the computational code. The latter holds true with the heteroscedastic student-t model. The structure of the natural gradient update (35) provides stable implementation. But it is hard to state whether the natural gradient will always provide faster convergence. Some theoretical studies of the convergence speed and statistical properties of the natural gradient can be found in Martens (2014, Sect. 12).

By carefully noting the particular orthogonal property of the parameters in the student-t model, the natural gradient for finding the parameters of the Laplace approximation proposed here becomes attractive. With this approach the Laplace approximation is available for non-log-concave likelihoods and likelihoods that depend on more than one Gaussian process with the same stability and easiness of implementation as the Laplace approximation for log-concave likelihoods presented by Rasmussen and Williams (2006) (see their book for pseudocode).

The choice of the matrix of metric coefficient G, which may be difficult to obtain in general optimization settings, can always be induced through the probabilistic model for the data. Thus, due to the probabilistic nature of our approach, the natural gradient is better suited to optimize the posterior density of the Gaussian process than the Newton's method. Moreover, for the most of the probabilistic models presented in the literature, the Fisher information matrix is available in closed-form (see Johnson et al. 1995). Hence, one can always investigate a new parametrization for the probabilistic model such that the Fisher information matrix is diagonal (see Sect. 2). Besides, this is not restricted to the case where two parameters of a probabilistic model are modelled with Gaussian process priors, as shown in this paper. In fact, the approach presented here can also be used in the homoscedastic student-t model of Vanhatalo et al. (2009) as well as in other uniparametric models, such as the Bernoulli and Poisson. These uniparametric models are commonly used within the context of Gaussian process regression and some type of reparametrization could be beneficial to improve posterior approximations and the estimation process. The studies by Achcar and Smith (1990), Kass and Slate (1994), Achcar (1994) and MacKay (1998) indicate and discuss possible ways to do so.

More generally, concepts of reparametrization in statistical modelling within the Gaussian process regression context deserve more attention. There is freedom of choice in the parametrization of the probabilistic model. If the posterior "normality" or inferential procedures can be improved under different parametrizations, then approximation methods may be reassessed. That is, all of the well known approximation methods such as variational-Bayes, expectation-propagation or Laplace's method, approximate the target density with a Gaussian density. If the target density in some new parametrization is closer to a Gaussian, then the choice of the approximation method may not be as crucial as its computational aspects.

These aspects of reparametrization are also important for MCMC methods. If there are difficulties to sample from a posterior density in some specific parametrization of the model, one can also investigate a new set of parameters so that the sampling problem is alleviated. For example, in the state-of-the-art Riemann manifold Hamiltonian Monte Carlo method (RMHMC) (Girolami and Calderhead 2011) the choice of the Riemannian metric (the Fisher information matrix) is essential for achieving good performance of the sampler. However, its computational implementation is hard and costly since G is full matrix in most practical applications. If there is a possibility to find an orthogonal parametrization for the model parameters such that G is diagonal, or at least it is not full matrix, then the computational aspects of the method could be further simplified. In this sense, the attractiveness of the method due to its properties would increase its use in practical applications.

The code implementing the model and the natural gradient approach as well as the Newton method are freely available at https://github.com/mahaa2/LP-approximation-and-NG-for-GPs-with-heteroscedastic-Student-t-mode. A demo code also follows in the aforementioned link.

Acknowledgements The work was funded by the Academy of Finland (Grant 304531) and the research funds of University of Helsinki. The first author thanks Luiz Roberto Hartmann Junior, Susan Chumbimune, Marco Pollo Almeida, Teo Calvo and Donald Smart for their comments and suggestions that helped improve the paper.

Appendix A: Datasets

A short description of the benchmark datasets used to evaluate the predictive performance of the models proposed in this paper. See Sect. 6, Table 4.

Neal. This is a simulated dataset with the presence of strong outliers. The dataset was also used by Neal (1997) (see p. 21, Figure 5) for the Gaussian process regression with the homocesdastic student-*t* model.

Motorcycle. This dataset consists of motorcycle accelerometer readings versus the time of impact in order to study the efficacy of helmets. This case illustrates a unidimensional nonlinear regression problem which was studied by Silverman (1985).

Boston housing. A well-known study case on housing prices, which was used to investigate whether clean air influenced the price of houses within the Boston metropolitan area in 1978. The dataset is composed by 506 measurements (census tracts) where each measurement consists of 13 covariates and 1 dependent variable, which is the median house price for that census tract. The detailed description of each explanatory variable can be consulted in Harrison and Rubinfeld (1978) table IV.

Friedman. A special regression function provided by Friedman (1991) and Jylänki et al. (2011), which involves a nonlinear regression function with 5 covariates. To make the experiment more challenging for the inference algorithm, 5 extra random covariates were generated as described by Jylänki et al. (2011). In this experiment a dataset with 200 observations is generated with 10 randomly selected outliers.

Compressive. A dataset for which the task is to predict concrete compressive strength based on 8 covariates and 1030 measurements. More details are described in Yeh (1998).

Appendix B: Extra formulas

In all the equations presented below we consider that $z_i = \frac{y_i - f_1(\mathbf{x}_i)}{\exp(f_2(\mathbf{x}_i))}$ for $i = 1, \ldots, n$.

B.1 The elements of the matrix W

$$\mathrm{W}_{i,j} = \begin{cases} \frac{1}{[\exp(f_2(\mathbf{x}_i))]^2}\left(1+\frac{1}{\nu}\right)\left[\frac{2}{(1+z_i^2/\nu)^2} - \frac{1}{1+z_i^2/\nu}\right] \\ \text{for } i = j = 1, \ldots, n \\ \frac{2}{\exp(f_2(\mathbf{x}_i))}\left(1+\frac{1}{\nu}\right)\frac{z_i}{(1+z_i^2/\nu)^2} \\ \text{for } i = 1, \ldots, N \text{ and} \\ j = (i+n)\mathbf{1}_{\{1,\ldots,n\}}(i) + (i-n)\mathbf{1}_{\{n+1,\ldots,N\}}(i) \\ 2(\frac{1}{\nu}+1)\frac{z_i^2}{(1+z_i^2/\nu)^2} \\ \text{for } i = j = n+1, \ldots, N \\ 0, \text{ otherwise.} \end{cases} \tag{47}$$

B.2 The elements of the Fisher information matrix $\mathbb{E}_{Y\,|\,\mathbf{f},\boldsymbol{\theta}}[\mathbf{W}]$

$$\mathbb{E}_{Y\,|\,\mathbf{f},\boldsymbol{\theta}}[\mathrm{W}]_{i,j} = \begin{cases} \frac{\nu+1}{\nu+3}\exp(-2f_2(\mathbf{x}_i)) \\ \text{for } i = j = 1, \ldots, n \\ \frac{2\nu}{\nu+3} \\ \text{for } i = j = n+1, \ldots, N \\ 0, \text{ otherwise.} \end{cases} \tag{48}$$

B.3 Derivatives of the log-likelihood

For each $i = 1, \ldots, n$ the elements of the gradient $\nabla_{\mathbf{f}} \log L(\mathbf{y} \mid \mathbf{f}, \nu)$ are given by

$$\begin{aligned} \frac{\partial \log \pi(y_i \mid f_1(\mathbf{x}_i), f_2(\mathbf{x}_i), \nu)}{\partial f_1(\mathbf{x}_i)} &= \left(1+\frac{1}{\nu}\right)\frac{z_i}{\exp(f_2(\mathbf{x}_i))(1+z_i^2/\nu)} \\ \frac{\partial \log \pi(y_i \mid f_1(\mathbf{x}_i), f_2(\mathbf{x}_i), \nu)}{\partial f_2(\mathbf{x}_i)} &= \frac{z_i^2-1}{(1+z_i^2/\nu)}. \end{aligned} \tag{49}$$

References

Achcar, J.A.: Some aspects of reparametrization in statistical models. Pak. J. Stat. **10**(3), 597–616 (1994)

Achcar, J.A., Smith, A.F.: Aspects of reparametrization in approximate Bayesian inference. Bayesian Likelihood Methods Stat. Econ. **4**(2), 439–452 (1990)

Amari, S.: Natural gradient works efficiently in learning. Neural Comput. (communicated by Steven Nowlan and Erkki Oja) **10**, 251–276 (1998)

Amari, S., Nagaoka, H.: Methods of Information Geometry. Translations of mathematical monographs. American Mathematical Society (2007)

Atkinson, A., Riani, M.: Robust Diagnostic Regression Analysis. Springer Series in Statistics. Springer, New York (2000)

Atkinson, C., Mitchell, A.F.S.: Rao's distance measure. Sankhyä Ser. A **43**, 345–365 (1981)

Bernardo, J.-M.: Reference posterior distributions for Bayesian-inference. J. R. Stat. Soc. Ser. B Methodol. **41**(2), 113–147 (1979)

Bernardo, J.-M., Smith, A.F.M.: Bayesian Theory. Wiley, Chichester (1994)

Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, New York Inc (2006)

Box, G.E., Tiao, G.C.: Bayesian Inference in Statistical Analysis. Addison-Wesley Pub. Co, Reading (1973)

Calderhead, B.: Differential Geometric MCMC Methods and Applications. Ph.D. thesis, University of Glasgow (2012)

Cox, D.R., Reid, N.: Parameter orthogonality and approximate conditional inference. J. R. Stat. Soc. Ser. B Methodol., (pp. 1–39) (1987)

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

Fernandez, C., Steel, M.F.J.: Multivariate Student-*t* regression models: pitfalls and inference. Biometrika **86**(1), 153–167 (1999)

Finetti, B.D.: The Bayesian approach to the rejection of outliers. In: Proceeding of Fourth Berkeley Symposium on Mathematical Statistics and Probability, (pp. 199–210). University of California Press (1961)

Fonseca, T.C.O., Ferreira, M.A.R., Migon, H.S.: Objective Bayesian analysis for the student-*t* regression model. Biometrika **95**(2), 325 (2008)

Friedman, J.H.: Multivariate adaptive regression splines. Ann. Stat. **19**(1), 1–67 (1991)

Fuglstad, G.-A., Simpson, D., Lindgren, F., Rue, H.: Constructing priors that penalize the complexity of Gaussian random fields. J. Am. Stat. Assoc. **1**(1), 1–8 (2018)

Fukumizu, K., Amari, S.: Local minima and plateaus in hierarchical structures of multilayer perceptrons. Neural Netw. **13**(3), 317–327 (2000)

Gelman, A.: Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal. **1**(3), 515–534 (2006)

Gelman, A., Hwang, J., Vehtari, A.: Understanding predictive information criteria for Bayesian models. Stat. Comput. **24**(6), 997–1016 (2014)

Geweke, J.: Bayesian treatment of the independent student-*t* linear model. J. Appl. Econ. **8**(S1), S19–S40 (1993)

Gibbs, M.N.: Bayesian Gaussian Processes for Regression and Classification.. Ph.D. thesis, Department of Physics, University of Cambridge (1997)

Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. Stat. R. Soc. B **73**(2), 123–214 (2011)

Gosset, W.S.: The probable error of a mean. Biometrika **6**(1), 1 (1908)

Gupta, R.D., Kundu, D.: On the comparison of Fisher information of the Weibull and generalized-exponential distributions. J. Stat. Plan. Inference **136**(9), 3130–3144 (2006)

Hang, H.H., Amari, S.: The efficiency and the robustness of the natural gradient descent learning rule. In: Advances in Neural Information Processing Systems (1998)

Harrison, D., Rubinfeld, D.L.: Hedonic housing prices and the demand for clean air. J. Environ. Econ. Manag. **5**(1), 81–102 (1978)

Hasenclever, L., Webb, S., Lienart, T., Vollmer, S., Lakshminarayanan, B., Blundell, C., Teh, Y.W.: Distributed Bayesian learning with stochastic natural gradient expectation propagation and the posterior server. J. Mach. Learn. Res. **18**(106), 1–37 (2017)

Hensman, J., Matthews, A.G.d.G., Ghahramani, Z.: Scalable variational Gaussian process classification. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (2015)

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., Karhunen, J.: Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. J. Mach. Learn. **11**, 3235–3268 (2010)

Huzurbazar, V.S.: Probability distribution and orthogonal parameters. In: Mathematical Proceedings of the Cambridge philosofical society, vol. 46, (pp. 281–284) (1950)

Huzurbazar, V.S.: Sufficient statistics and orthogonal parameters. Sankhyä Indian J. Stat. (1933–1960) **17**(3), 217–220 (1956)

Jeffreys, H.: The Theory of Probability. Oxford Classic Texts in the Physical Sciences. OUP Oxford, 3rd edn (1998)

Jennrich, R.I., Sampson, P.F.: Newton–Raphson and related algorithms for maximum likelihood variance component estimation. Technometrics **18**(1), 11–17 (1976)

Johnson, N., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, Hoboken (1995)

Jylänki, P., Vanhatalo, J., Vehtari, A.: Robust Gaussian process regression with a student-*t* likelihood. J. Mach. Learn. Res. **12**, 3227–3257 (2011)

Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**(430), 773–795 (1995)

Kass, R.E., Slate, E.H.: Some diagnostics of maximum likelihood and posterior nonnormality. Ann. Stat. **22**(2), 668–695 (1994)

Kass, R.E., Vaidyanathan, S.K.: Approximate Bayes factors and orthogonal parameters with application to testing equality of two binomial proportions. J. R. Stat. Soc. Ser. B Methodol. **54**, 129–144 (1992)

Kuss, M., Rasmussen, C.E.: Assessing approximate inference for binary Gaussian process classification. J. Mach. Learn. Res. **6**, 1679–1704 (2005)

Lange, K.L., Little, R.J.A., Taylor, J.M.G.: Robust statistical modeling using the *t*-Distribution. J. Am. Stat. Assoc. **84**, 881–896 (1989)

MacKay, D.J.: Choice of basis for Laplace approximation. Mach. Learn. **33**(1), 77–86 (1998)

MacKay, D.J.C.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge (2002)

Martens, J.: New perspectives on the natural gradient method (2014). arXiv:1412.1193

Migon, H.S., Gamerman, D., Louzada, F.: Statistical inference: An integrated approach, Second Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis (2014)

Minka, T.: A Family of Algorithms for Approximate Bayesian Inference. Ph.D. thesis, Massachusetts Institute of Tecnology (2001a)

Minka, T.: Expectation propagation for approximate Bayesian inference. In: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, UAI '01, (pp. 362–369). Morgan Kaufmann Publishers Inc (2001b)

Murray, I., Adams, R.A., Mackay, D.J.: Elliptical slice sampling. In: Journal of Machine Learning Research: Workshop and Conference Proceedings. International Conference on Artificial Intelligence and Statistics, vol. 9, (pp. 541–548) (2010)

Murray, I., Adams, R.P.: Slice sampling covariance hyperparameters of latent Gaussian models. In: Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., Culotta, A. (Eds.) Advances in Neural Information Processing Systems 23, (pp. 1732–1740). Curran Associates, Inc (2010)

Neal, R.: Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. Technical report, Department of Statistics, Department of Computer Science, University of Toronto (1997)

Nickisch, H., Rasmussen, C.E.: Approximations for binary Gaussian process classification. J. Mach. Learn. **9**, 2035–2078 (2008)

O'Hagan, A.: Curve fitting and optimal design for prediction. J. R. Stat. Soc. B Methodol. **40**(1), 1–42 (1978)

O'Hagan, A.: On outlier rejection phenomena in Bayes inference. J. R. Stat. Soc. B Methodol. **41**(3), 358–367 (1979)

O'Hagan, A.: Kendall's Advanced Theory of Statistics: Bayesian Inference. Oxford University Press, Oxford (2004)

Ollivier, Y., Arnold, L., Auger, A., Hansen, N.: Information-geometric optimization algorithms: a unifying picture via invariance principles. J. Mach. Learn. Res. **18**(18), 1–65 (2017)

Polak, B.T.: Newton's method and its use in optimization. Eur. J. Oper. Res. **181**, 1086–1096 (2006)

Raftery, A.E.: Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika **83**(2), 251 (1996)

Rao, R.C.: Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81–91 (1945)

Rasmussen, C.E., Nickisch, H.: Gaussian processes for machine learning (GPML) toolbox. J. Mach. Learn. Res. **11**, 3011–3015 (2010)

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

Rue, H., Martino, S.: Approximate Bayesian inference for latent Gaussian models by using integrated Laplace approximations. J. R. Stat. Soc. Ser. B Methodol. **71**(2), 319–392 (2009)

Saul, A., Hensman, J., Vehtari, A., Lawrence, N.: Chained Gaussian processes. In: Journal of Machine Learning Research: Workshop and Conference Proceedings. International Conference on Artificial Intelligence and Statistics, vol. 51, pp. 1431–1440 (2016)

Schervish, M.J.: Theory of Statistics. Springer Series in Statistics. Springer, Berlin (2011)

Seber, G.A.F., Lee, A.J.: Linear Regression Analysis. Wiley Series in Probability and Statistics. Wiley, Hoboken (2012)

Seber, G.A.F., Wild, C.: Nonlinear Regression. Wiley Series in Probability and Statistics. Wiley, Hoboken (2003)

Silverman, B.W.: Some aspects of the spline smoothing approach to non-parametric regression curve fitting. J. R. Stat. Soc. Ser. B Methodol. **47**(1), 1–52 (1985)

Simpson, D.P., Rue, H., Martins, T.G., Riebler, A., Sørbye, S.H.: Penalising model component complexity: a principled, practical approach to constructing priors. Stat. Sci. **32**(1), 1–28 (2017)

Snelson, E., Ghahramani, Z.: Sparse Gaussian processes using pseudo-inputs. In: Proceedings of the 18th International Conference on Neural Information Processing Systems, NIPS'05, (pp. 1257–1264). MIT Press (2005)

Tierney, L., Kadane, J.B.: Accurate approximation for posterior moments and marginal densities. J. Am. Stat. Assoc. **81**(393), 82–86 (1986)

Tierney, L., Kass, R.E., Kadane, J.B.: Fully exponential Laplace approximations to expectations and variances of nonpositive functions. J. Am. Stat. Assoc. **84**(407), 710–716 (1989)

Tipping, M.E., Lawrence, N.D.: Variational inference for Student-*t* models: Robust Bayesian interpolation and generalised component analysis. Neurocomputing **69**(1–3), 123–141 (2005)

Titsias, M.K.: Variational learning of inducing variables in sparse Gaussian processes. In: Artificial Intelligence and Statistics 12, (pp. 567–574) (2009)

Vanhatalo, J., Jylänki, P., Vehtari, A.: Gaussian process regression with a Student-*t* likelihood. In: Advances in Neural Information Processing Systems (2009)

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

Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., Vehtari, A.: GPstuff: Bayesian modeling with Gaussian processes. J. Mach. Learn. Res. **14**(1), 1175–1179 (2013)

Vehtari, A., Ojanen, J.: A survey of bayesian predictive methods for model assessment, selection and comparison. Stat. Surv. **6**, 141–228 (2012)

Wang, M., Yang, M.: Posterior property of Student-*t* linear regression model using objective priors. Stat. Probab. Lett. **113**, 23–29 (2016)

West, M.: Outlier models and prior distributions in Bayesian linear regression. J. R. Stat. Soc. Ser. B Methodol. **46**(3), 431–439 (1984)

Yeh, I.-C.: Modeling of strength of high-performance concrete using artificial neural networks. Cement Concrete Res. **28**, 1797–1808 (1998)