Journal of Machine Learning Research 12 (2011) 3227-3257 Submitted 6/11; Revised 10/11; Published 11/11

Robust Gaussian Process Regression with a Student-*t* Likelihood

Pasi Jylänki PASI.JYLANKI@AALTO.FI
Department of Biomedical Engineering and Computational Science
Aalto University School of Science
P.O. Box 12200
FI-00076 Aalto
Finland

Jarno Vanhatalo JARNO.VANHATALO@HELSINKI.FI
Department of Environmental Sciences
University of Helsinki
P.O. Box 65
FI-00014 Helsinki
Finland

Aki Vehtari AKI.VEHTARI@AALTO.FI
Department of Biomedical Engineering and Computational Science
Aalto University School of Science
P.O. Box 12200
FI-00076 Aalto
Finland

Editor: Neil Lawrence

Abstract

This paper considers the robust and efficient implementation of Gaussian process regression with a Student-*t* observation model, which has a non-log-concave likelihood. The challenge with the Student-*t* model is the analytically intractable inference which is why several approximative methods have been proposed. Expectation propagation (EP) has been found to be a very accurate method in many empirical studies but the convergence of EP is known to be problematic with models containing non-log-concave site functions. In this paper we illustrate the situations where standard EP fails to converge and review different modifications and alternative algorithms for improving the convergence. We demonstrate that convergence problems may occur during the type-II maximum a posteriori (MAP) estimation of the hyperparameters and show that standard EP may not converge in the MAP values with some difficult data sets. We present a robust implementation which relies primarily on parallel EP updates and uses a moment-matching-based double-loop algorithm with adaptively selected step size in difficult cases. The predictive performance of EP is compared with Laplace, variational Bayes, and Markov chain Monte Carlo approximations.

Keywords: Gaussian process, robust regression, Student-*t* distribution, approximate inference, expectation propagation

1. Introduction

In many regression problems observations may include outliers which deviate strongly from the other members of the sample. Such outliers may occur, for example, because of failures in the measurement process or absence of certain relevant explanatory variables in the model. In such cases, a robust observation model is required.

Robust inference has been studied extensively. De Finetti (1961) described how Bayesian inference on the mean of a random sample, assuming a suitable observation model, naturally leads to giving less weight to outlying observations. However, in contrast to simple rejection of outliers, the posterior depends on all data but in the limit, as the separation between the outliers and the rest of the data increases, the effect of outliers becomes negligible. More theoretical results on this kind of outlier rejection were presented by Dawid (1973) who gave sufficient conditions on the observation model $p(y|\theta)$ and the prior distribution $p(\theta)$ of an unknown location parameter θ, which ensure that the posterior expectation of a given function $m(\theta)$ tends to the prior as $y \rightarrow \infty$. He also stated that the Student-t distribution combined with a normal prior has this property.

A more formal definition of robustness was given by O'Hagan (1979) in terms of an outlier-prone observation model. The observation model is defined to be outlier-prone of order n, if $p(\theta|y_1,...,y_{n+1}) \rightarrow p(\theta|y_1,...,y_n)$ as $y_{n+1} \rightarrow \infty$. That is, the effect of a single conflicting observation to the posterior becomes asymptotically negligible as the observation approaches infinity. O'Hagan (1979) showed that the Student-t distribution is outlier prone of order 1, and that it can reject up to m outliers if there are at least $2m$ observations altogether. This contrasts heavily with the commonly used Gaussian observation model in which each observation influences the posterior no matter how far it is from the others.

In nonlinear Gaussian process (GP) regression context the outlier rejection is more complicated and one may consider the posterior distribution of the unknown function values $f_i = f(x_i)$ locally near some input locations x_i. Depending on the smoothness properties defined through the prior on f_i, m observations can be rejected locally if there are at least $2m$ data points nearby. However, already two conflicting data points can render the posterior distribution multimodal making the posterior inference challenging (these issues will be illustrated in the upcoming sections).

In this work, we adopt the Student-t observation model for GP regression because of its good robustness properties which can be altered continuously from a very heavy tailed distribution to the Gaussian model with the degrees of freedom parameter. This allows the extent of robustness to be determined from the data through hyperparameter inference. The Student-t observation model was studied in linear regression by West (1984) and Geweke (1993), and Neal (1997) introduced it for GP regression. Other robust observation models which have been used in GP regression include, for example, mixtures of Gaussians (Kuss, 2006; Stegle et al., 2008), the Laplace distribution (Kuss, 2006), and input dependent observation models (Goldberg et al., 1998; Naish-Guzman and Holden, 2008).

The challenge with the Student-t model is the analytically intractable inference. A common approach has been to use the scale-mixture representation of the Student-t distribution (Geweke, 1993), which enables Gibbs sampling (Geweke, 1993; Neal, 1997), and a factorizing variational approximation (fVB) for the posterior inference (Tipping and Lawrence, 2005; Kuss, 2006). Recently Vanhatalo et al. (2009) compared fVB with the Laplace approximation (see, e.g., Rasmussen and Williams, 2006) and showed that Laplace's method provided slightly better predictive performance with less computational burden. They also showed that fVB tends to underestimate the posterior uncertainties of the predictions because it assumes the scales and the unknown function values a posteriori independent. Another variational approach called variational bounds (VB) is available in the GPML software package (Rasmussen and Nickisch, 2010). The method is based on forming an un-normalized Gaussian lower bound for each non-Gaussian likelihood term independently (see Nickisch and Rasmussen, 2008, for details and comparisons in GP classification). Yet another related variational approach is described by Opper and Archambeau (2009) who studied the Cauchy observation model (Student-*t* with degrees of freedom 1). This method is similar to the KL-divergence minimization approach (KL) described by Nickisch and Rasmussen (2008) and the VB approach can be regarded as a special case of KL. The extensive comparisons by Nickisch and Rasmussen (2008) in GP classification suggest that VB provides better predictive performance than the Laplace approximation but worse marginal likelihood estimates than KL or expectation propagation (EP) (Minka, 2001a). According to the comparisons of Nickisch and Rasmussen (2008), EP is the method of choice since it is much faster than KL, at least in GP classification. The problem with EP, however, is that the Student-*t* likelihood is not log-concave which may lead to convergence problems (Seeger, 2008).

In this paper, we focus on establishing a robust EP implementation for the Student-*t* observation model. We illustrate the convergence problems of standard EP with simple one-dimensional regression examples and discuss how damping, fractional EP updates (or power EP) (Minka, 2004; Seeger, 2005), and double-loop algorithms (Heskes and Zoeter, 2002) can be used to improve the convergence. We present a robust implementation which relies primarily on parallel EP updates (see, e.g., van Gerven et al., 2009) and uses a moment-matching-based double-loop algorithm with adaptively selected step size to find stationary solutions in difficult cases. We show that the implementation enables a robust type-II maximum a posteriori (MAP) estimation of the hyperparameters based on the approximative marginal likelihood. The proposed implementation is general so that it could be applied also to other models having non-log-concave likelihoods. The predictive performance of EP is compared to the Laplace approximation, fVB, VB, and Markov chain Monte Carlo (MCMC) using one simulated and three real-world data sets.

2. Gaussian Process Regression with the Student-*t* Observation Model

We will consider a regression problem, with scalar observations $y_i = f(\mathbf{x}_i) + \varepsilon_i, i = 1,...,n$ at input locations $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, and where the observation errors $\varepsilon_1,...,\varepsilon_n$ are zero-mean exchangeable random variables. The object of inference is the latent function $f(\mathbf{x}) : \Re^d \to \Re$, which is given a Gaussian process prior

$$f(\mathbf{x})|\theta \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x},\mathbf{x}'|\theta)\right), \tag{1}$$

where $m(\mathbf{x})$ and $k(\mathbf{x},\mathbf{x}'|\theta)$ are the mean and covariance functions of the process controlled by hyperparameters θ. For notational simplicity we will assume a zero mean GP. By definition, a Gaussian process prior implies that any finite subset of latent variables, $\mathbf{f} = \{f(\mathbf{x}_i)\}_{i=1}^n$, has a multivariate Gaussian distribution. In particular, at the observed input locations $\mathbf{X}$ the latent variables are distributed as $p(\mathbf{f}|\mathbf{X},\theta) = \mathcal{N}(\mathbf{f}|\mathbf{0},\mathbf{K})$, where $\mathbf{K}$ is the covariance matrix with entries $\mathbf{K}_{ij} = k(\mathbf{x}_i,\mathbf{x}_j|\theta)$. The covariance function encodes the prior assumptions on the latent function, such as the smoothness and scale of the variation, and can be chosen freely as long as the covariance matrices which it produces are symmetric and positive semi-definite. An example of a stationary covariance function is the squared exponential

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

where $\theta = \{\sigma_{\mathrm{se}}^2, l_1, \ldots, l_d\}$, σ_{se}^2 is a magnitude parameter which scales the overall variation of the unknown function, and l_k is a length-scale parameter which governs how fast the correlation decreases as the distance increases in the input dimension k.

The traditional assumption is that given $\mathbf{f}$ the error terms ε_i are i.i.d. Gaussian: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. In this case, the marginal likelihood $p(\mathbf{y}|\mathbf{X}, \theta, \sigma^2)$ and the conditional posterior of the latent variables $p(\mathbf{f}|\mathcal{D}, \theta, \sigma^2)$, where $\mathcal{D} = \{\mathbf{y}, \mathbf{X}\}$, have an analytical solution. This is computationally convenient since approximate methods are needed only for the inference on the hyperparameters θ and σ^2. The robust Student-t observation model

$$p(y_i|f_i, \sigma^2, \nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma} \left(1 + \frac{(y_i - f_i)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2},$$

where $f_i = f(\mathbf{x}_i)$, ν is the degrees of freedom and σ the scale parameter (Gelman et al., 2004), is computationally challenging. The marginal likelihood and the conditional posterior $p(\mathbf{f}|\mathcal{D}, \theta, \sigma^2, \nu)$ are not anymore analytically tractable but require some method for approximate inference.

3. Approximate Inference

In this section, we review the approximate inference methods considered in this paper. First we give a short description of MCMC and the Laplace approximation, as well as two variational methods, fVB and VB. Then we give a more detailed description of the EP algorithm and review ways to improve the convergence in more difficult problems.

3.1 Markov Chain Monte Carlo

The MCMC approach is based on drawing samples from $p(\mathbf{f}, \theta, \sigma^2, \nu|\mathcal{D})$ and using these samples to represent the posterior distribution and to numerically approximate integrals over the latent variables and the hyperparameters. Instead of implementing a Markov chain sampler directly for the Student-t model, a more common approach is to use the Gibbs sampling based on the following scale mixture representation of the Student-t distribution

$$\begin{aligned} y_i|f_i, V_i &\sim \mathcal{N}(f_i, V_i), \\ V_i|\nu, \sigma^2 &\sim \text{Inv-}\chi^2(\nu, \sigma^2), \end{aligned} \tag{3}$$

where each observation has its own Inv-χ^2-distributed noise variance V_i (Neal, 1997; Gelman et al., 2004). Sampling of the hyperparameters θ can be done with any general sampling algorithm, such as the Slice sampling or the hybrid Monte Carlo (HMC) (see, e.g., Gelman et al., 2004). The Gibbs sampler on the scale mixture (3) converges often slowly and may get stuck for long times in small values of σ^2 because of the dependence between V_i and σ^2. This can be avoided by re-parameterization $V_i = \alpha^2 U_i$, where $U_i \sim \text{Inv-}\chi^2(\nu, \tau^2)$, $p(\tau^2) \propto 1/\tau^2$, and $p(\alpha^2) \propto 1/\alpha^2$ (Gelman et al., 2004). This improves mixing of the chains and reduces the autocorrelations but introduces an implicit prior for the scale parameter $\sigma^2 = \alpha^2\tau^2$ of the Student-t model. An alternative parameterization proposed by Liu and Rubin (1995), where $V_i = \sigma^2/\gamma_i$ and $\gamma_i \sim \text{Gamma}(\nu/2, \nu/2)$, also decouples σ^2 and V_i but does not introduce the additional scale parameter τ. It could also lead to better mixing without the implicit scale prior but in the experiments we used the decomposition of Gelman et al. (2004) because the results were not sensitive to the choice of prior on σ^2.

3.2 Laplace Approximation (LA)

The Laplace approximation for the conditional posterior of the latent function is constructed from the second order Taylor expansion of $\log p(\mathbf{f}|\mathcal{D},\theta,\sigma^2,\nu)$ around the mode $\hat{\mathbf{f}}$, which gives a Gaussian approximation to the conditional posterior

$$p(\mathbf{f}|\mathcal{D},\theta,\sigma^2,\nu) \approx q(\mathbf{f}|\mathcal{D},\theta,\sigma^2,\nu) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}},\mathbf{\Sigma}_{\text{LA}}),$$

where $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} p(\mathbf{f}|\mathcal{D},\theta,\sigma^2,\nu)$ (Rasmussen and Williams, 2006). $\mathbf{\Sigma}_{\text{LA}}^{-1}$ is the Hessian of the negative log conditional posterior at the mode, that is,

$$\mathbf{\Sigma}_{\text{LA}}^{-1} = -\nabla\nabla \log p(\mathbf{f}|\mathcal{D},\theta,\sigma^2,\nu)|_{\mathbf{f}=\hat{\mathbf{f}}} = \mathbf{K}^{-1} + \mathbf{W}, \tag{4}$$

where $\mathbf{W}$ is a diagonal matrix with entries $\mathbf{W}_{ii} = \nabla_{f_i}\nabla_{f_i} \log p(y|f_i,\sigma^2,\nu)|_{f_i=\hat{f}_i}$.

The inference in the hyperparameters is done by approximating the conditional marginal likelihood $p(\mathbf{y}|\mathbf{X},\theta,\sigma^2,\nu)$ with Laplace's method and searching for the approximate maximum a posterior estimate for the hyperparameters

$$\{\hat{\theta},\hat{\sigma}^2,\hat{\nu}\} = \underset{\theta,\sigma^2,\nu}{\arg\max}\left[\log q(\theta,\sigma^2,\nu|\mathcal{D})\right] = \underset{\theta,\sigma^2,\nu}{\arg\max}\left[\log q(\mathbf{y}|\mathbf{X},\theta,\sigma^2,\nu) + \log p(\theta,\sigma^2,\nu)\right],$$

where $p(\theta,\sigma^2,\nu)$ is the prior of the hyperparameters. The gradients of the approximate log marginal likelihood can be solved analytically, which enables the MAP estimation of the hyperparameters with gradient based optimization methods. Following Williams and Barber (1998) the approximation scheme is called the Laplace method, but essentially the same approach is named Gaussian approximation by Rue et al. (2009) in their Integrated nested Laplace approximation (INLA) software package for Gaussian Markov random field models (Vanhatalo et al., 2009), (see also Tierney and Kadane, 1986).

The implementation of the Laplace algorithm for this particular model requires care since the Student-*t* likelihood is not log-concave and thus $p(\mathbf{f}|\mathcal{D},\theta,\sigma^2,\nu)$ may be multimodal and some of the $\mathbf{W}_{ii}$ negative. It follows that the standard implementation presented by Rasmussen and Williams (2006) requires some modifications in determining the mode $\hat{\mathbf{f}}$ and the covariance $\mathbf{\Sigma}_{\text{LA}}$ which are discussed in detail by Vanhatalo et al. (2009). Later on Hannes Nickisch proposed a slightly different implementation (personal communication) where the stabilized Newton algorithm is used for finding $\hat{\mathbf{f}}$ instead of the EM algorithm and LU decomposition for determining $\mathbf{\Sigma}_{\text{LA}}$ instead of rank-1 Cholesky updates (see also Section 4.1). This alternative approach is used at the moment in the GPML software package (Rasmussen and Nickisch, 2010).

3.3 Factorizing Variational Approximation (fVB)

The scale-mixture decomposition (3) enables a computationally convenient variational approximation if the latent values $\mathbf{f}$ and the residual variance terms $\mathbf{V} = [V_1,...,V_n]$ are assumed a posteriori independent:

$$q(\mathbf{f},\mathbf{V}) = q(\mathbf{f})\prod_{i=1}^{n} q(V_i). \tag{5}$$

This kind of factorizing variational approximation was introduced by Tipping and Lawrence (2003) to form a robust observation model for linear models within the relevance vector machine framework. For robust Gaussian process regression with the Student-*t* model it was applied by Kuss

(2006) and essentially the same variational approach has also been used for approximate inference on linear models with the automatic relevance determination prior (see, e.g., Tipping and Lawrence, 2005). Assuming the factorizing posterior (5) and minimizing the KL-divergence from $q(\mathbf{f},\mathbf{V})$ to the true posterior $p(\mathbf{f},\mathbf{V}|\mathcal{D},\theta,\sigma^2,\nu)$ results in a Gaussian approximation for the latent values, and inverse-χ^2 (or equivalently inverse gamma) approximations for the residual variances V_i. The parameters of $q(\mathbf{f})$ and $q(V_i)$ can be estimated by maximizing a variational lower bound for the marginal likelihood $p(\mathbf{y}|X,\theta,\sigma^2,\nu)$ with an expectation maximization (EM) algorithm. In the E-step of the algorithm the lower bound is maximized with respect to $q(\mathbf{f})$ and $q(V_i)$ given the current point estimate of the hyperparameters and in the M-step a new estimate of the hyperparameters is determined with fixed $q(\mathbf{f})$ and $q(V_i)$.

The drawback with a factorizing approximation determined by minimizing the reverse KL-divergence is that it tends to underestimate the posterior uncertainties (see, e.g., Bishop, 2006). Vanhatalo et al. (2009) compared fVB with the previously described Laplace and MCMC approximations, and found that fVB provided worse predictive variance estimates compared to the Laplace approximation. In addition, the estimation of ν based on maximizing the variational lower bound was found less robust with fVB.

3.4 Variational Bounds (VB)

This variational bounding method was introduced for binary GP classification by Gibbs and MacKay (2000) and comparisons to other approximative methods for GP classification were done by Nickisch and Rasmussen (2008). The method is based on forming a Gaussian lower bound for each likelihood term independently:

$$p(y_i|f_i) \geq \exp(-f_i^2/(2\gamma_i) + b_i f_i - h(\gamma_i)/2),$$

which can be used to construct a lower bound on the marginal likelihood: $p(\mathbf{y}|\mathbf{X},\theta,\nu,\sigma) \geq Z_{\mathrm{VB}}$. With fixed hyperparameters, γ_i and b_i can be determined by maximizing Z_{VB} to obtain a Gaussian approximation for $p(\mathbf{f}|\mathcal{D},\theta,\nu,\sigma^2)$ and an approximation for the marginal likelihood. With the Student-*t* observation model only the scale parameters γ_i need to be optimized because the location parameter is determined by the corresponding observations: $b_i = y_i/\gamma_i$. Similarly to the Laplace approximation and EP, MAP-estimation of the hyperparameters can be done by optimizing Z_{VB} with gradient-based methods. In our experiments we used the implementation available in the GPML-package (Rasmussen and Nickisch, 2010) augmented with the same hyperprior definitions as with the other approximative methods.

3.5 Expectation Propagation

The EP algorithm is a general method for approximating integrals over functions that factor into simple terms (Minka, 2001a). It approximates the conditional posterior with

$$q(\mathbf{f}|\mathcal{D},\theta,\sigma^2,\nu) = \frac{1}{Z_{\mathrm{EP}}} p(\mathbf{f}|\theta) \prod_{i=1}^{n} \tilde{t}_i(f_i|\tilde{Z}_i,\tilde{\mu}_i,\tilde{\sigma}_i^2) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}), \tag{6}$$

where $Z_{\mathrm{EP}} \approx p(\mathbf{y}|\mathbf{X},\theta,\sigma^2,\nu)$, and the parameters of the approximate conditional posterior distribution are given by $\boldsymbol{\Sigma} = (\mathbf{K}^{-1} + \tilde{\boldsymbol{\Sigma}}^{-1})^{-1}$, $\boldsymbol{\mu} = \boldsymbol{\Sigma}\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\Sigma}} = \mathrm{diag}[\tilde{\sigma}_1^2,...,\tilde{\sigma}_n^2]$, and $\tilde{\boldsymbol{\mu}} = [\tilde{\mu}_1,...,\tilde{\mu}_n]^{\mathrm{T}}$. In Equation (6) the likelihood terms $p(y_i|f_i,\sigma^2,\nu)$ are approximated by un-normalized Gaussian site functions $\tilde{t}_i(f_i|\tilde{Z}_i,\tilde{\mu}_i,\tilde{\sigma}_i^2) = \tilde{Z}_i\mathcal{N}(f_i|\tilde{\mu}_i,\tilde{\sigma}_i^2)$.

The EP algorithm updates the site parameters $\tilde{Z}_i$, $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ and the posterior approximation (6) sequentially. At each iteration (i), first the i'th site is removed from the i'th marginal posterior to obtain a cavity distribution

$$q_{-i}(f_i) \propto q(f_i|\mathcal{D},\theta,\sigma^2,\nu)\tilde{t}_i(f_i)^{-1}.$$

Then the i'th site is replaced with the exact likelihood term to form a tilted distribution $\hat{p}_i(f_i) = \hat{Z}_i^{-1}q_{-i}(f_i)p(y_i|f_i)$ which is a more refined non-Gaussian approximation to the true i'th marginal distribution. Next the algorithm attempts to match the approximative posterior marginal $q(f_i) = q(f_i|\mathcal{D},\theta,\sigma^2,\nu)$ with $\hat{p}_i(f_i)$ by finding first a Gaussian $\hat{q}_i(f_i)$ satisfying

$$\hat{q}_i(f_i) = \mathcal{N}(f_i|\hat{\mu}_i,\hat{\sigma}_i^2) = \arg\min_{q_i} \mathrm{KL}\left(\hat{p}_i(f_i)||q_i(f_i)\right),$$

which is equivalent to matching $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ with the mean and variance of $\hat{p}_i(f_i)$. Then the parameters of the local approximation $\tilde{t}_i$ are updated so that the moments of $q(f_i)$ match with $\hat{q}_i(f_i)$:

$$q(f_i|\mathcal{D},\theta,\sigma^2,\nu) \propto q_{-i}(f_i)\tilde{t}_i(f_i) \equiv \hat{Z}_i\mathcal{N}(f_i|\hat{\mu}_i,\hat{\sigma}_i^2). \tag{7}$$

Finally, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the approximate posterior (6) are updated according to the changes in site $\tilde{t}_i$. These steps are repeated for all the sites at some order until convergence. Since only the means and variances are needed in the Gaussian moment matching only $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ need to be updated during the iterations. The normalization terms $\tilde{Z}_i$ are required for the marginal likelihood approximation $Z_{\mathrm{EP}} \approx p(\mathbf{y}|\mathbf{X},\theta,\sigma^2,\nu)$ which is computed after convergence of the algorithm, and they can be determined by integrating over f_i in Equation (7) which gives $\tilde{Z}_i = \hat{Z}_i(\int q_{-i}(f_i)\mathcal{N}(f_i|\tilde{\mu}_i,\tilde{\sigma}_i^2)df_i)^{-1}$.

In the traditional EP algorithm (from now on referred to as sequential EP), the posterior approximation (6) is updated sequentially after each moment matching (7). Recently an alternative parallel update scheme has been used especially in models with a very large number of unknowns (see, e.g., van Gerven et al., 2009). In parallel EP the site updates are calculated with fixed posterior marginals $\boldsymbol{\mu}$ and $\mathrm{diag}(\boldsymbol{\Sigma})$ for all $\tilde{t}_i$, $i = 1,...,n$, in parallel, and the posterior approximation is refreshed only after all the sites have been updated. Although the theoretical cost for one sweep over the sites is the same ($O(n^3)$) for both sequential and parallel EP, in practice one re-computation of $\boldsymbol{\Sigma}$ using the Cholesky decomposition is much more efficient than n sequential rank-one updates. In our experiments, the number of sweeps required for convergence was roughly the same for both schemes in easier cases where standard EP converges.

The marginal likelihood approximation is given by

$$\log Z_{\mathrm{EP}} = -\frac{1}{2}\log|\mathbf{K}+\tilde{\boldsymbol{\Sigma}}| - \frac{1}{2}\tilde{\boldsymbol{\mu}}^{\mathrm{T}}\left(\mathbf{K}+\tilde{\boldsymbol{\Sigma}}\right)^{-1}\tilde{\boldsymbol{\mu}} + \sum_{i=1}^{n}\log\hat{Z}_i(\sigma^2,\nu) + C_{\mathrm{EP}}, \tag{8}$$

where $C_{\mathrm{EP}} = -\frac{n}{2}\log(2\pi) - \sum_i \log\int q_{-i}(f_i)\mathcal{N}(f_i|\tilde{\mu}_i,\tilde{\sigma}_i^2)df_i$ collects terms that are not explicit functions of θ, σ^2 or ν. If the algorithm has converged, that is, $\hat{p}_i(f_i)$ is consistent (has the same means and variances) with $q(f_i)$ for all sites, C_{EP}, $\tilde{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\mu}}$ can be considered constants when differentiating (8) with respect to the hyperparameters (Seeger, 2005; Opper and Winther, 2005). This enables efficient MAP estimation with gradient based optimization methods.

There is no guarantee of convergence for either sequential or parallel EP. When the likelihood terms are log-concave and the approximation is initialized to the prior, the algorithm converges fine in many cases (see, e.g., Nickisch and Rasmussen, 2008). However, in case of a non-log-concave likelihood such as the Student-*t* likelihood, convergence problems may arise and these will be discussed in Section 5. The convergence can be improved either by damping the EP updates (Minka and Lafferty, 2002) or by using a robust but slower double-loop algorithm (Heskes and Zoeter, 2002). In damping, the site parameters in their natural exponential forms, $\tilde{\tau}_i = \tilde{\sigma}_i^{-2}$ and $\tilde{\nu}_i = \tilde{\sigma}_i^{-2}\tilde{\mu}_i$, are updated to a convex combination of the old and proposed new values, which results in the following update rules:

$$\Delta\tilde{\tau}_i = \delta(\hat{\sigma}_i^{-2} - \sigma_i^{-2}) \quad \text{and} \quad \Delta\tilde{\nu}_i = \delta(\hat{\sigma}_i^{-2}\hat{\mu}_i - \sigma_i^{-2}\mu_i), \tag{9}$$

where μ_i and σ_i^2 are the mean and variance of $q(f_i|\mathcal{D},\theta,\sigma^2,\nu)$, and $\delta \in (0,1]$ is a step size parameter controlling the amount of damping. Damping can be viewed as using a smaller step size within a gradient-based search for saddle points of the same objective function as is used in the double-loop algorithm (Heskes and Zoeter, 2002).

3.6 Expectation Propagation, the Double-Loop Algorithm

When either sequential or parallel EP does not converge one may still find approximations satisfying the moment matching conditions (7) by a double loop algorithm. For example, Heskes and Zoeter (2002) present simulation results with linear dynamical systems where the double loop algorithm is able to find useful approximations when damped EP fails to converge. For the model under consideration, the fixed points of the EP algorithm correspond to the stationary points of the following objective function (Minka, 2001b; Opper and Winther, 2005)

$$\begin{aligned}\min_{\boldsymbol{\lambda}_s}\max_{\boldsymbol{\lambda}_-} -\sum_{i=1}^{n}\log\int p(y_i|f_i)\exp\left(\nu_{-i}f_i - \tau_{-i}\frac{f_i^2}{2}\right)df_i - \log\int p(\mathbf{f})\prod_{i=1}^{n}\exp\left(\tilde{\nu}_i f_i - \tilde{\tau}_i\frac{f_i^2}{2}\right)d\mathbf{f} \\ + \sum_{i=1}^{n}\log\int\exp\left(\nu_{s_i}f_i - \tau_{s_i}\frac{f_i^2}{2}\right)df_i\end{aligned} \tag{10}$$

where $\boldsymbol{\lambda}_- = \{\nu_{-i},\tau_{-i}\}$, $\tilde{\boldsymbol{\lambda}} = \{\tilde{\nu}_i,\tilde{\tau}_i\}$, and $\boldsymbol{\lambda}_s = \{\nu_{s_i},\tau_{s_i}\}$ are the natural parameters of the cavity distributions $q_{-i}(f_i)$, the site approximations $\tilde{t}_i(f_i)$, and approximate marginal distributions $q_{s_i}(f_i) = \mathcal{N}(\tau_{s_i}^{-1}\nu_{s_i},\tau_{s_i}^{-1})$ respectively. The min-max problem needs to be solved subject to the constraints $\tilde{\nu}_i = \nu_{s_i} - \nu_{-i}$ and $\tilde{\tau}_i = \tau_{s_i} - \tau_{-i}$, which resemble the moment matching conditions in (7). The objective function in (10) is equal to $-\log Z_{\mathrm{EP}}$ defined in (6) and is also equivalent to the expectation consistent (EC) free energy approximation presented by Opper and Winther (2005). A unifying view of the EC and EP approximations as well as the connection to the Bethe free energies is presented by Heskes et al. (2005).

Equation (10) suggests a double-loop algorithm where the inner loop consist of maximization with respect to $\boldsymbol{\lambda}_-$ with fixed $\boldsymbol{\lambda}_s$ and the outer loop of minimization with respect to $\boldsymbol{\lambda}_s$. The inner maximization affects only the first two terms and ensures that the marginal moments of the current posterior approximation $q(\mathbf{f})$ are equal to the moments of the tilted distributions $\hat{p}_i(f_i)$ for fixed $\boldsymbol{\lambda}_s$. The outer minimization ensures that the moments $q_{s_i}(f_i)$ are equal to marginal moments of $q(\mathbf{f})$. At the convergence, $q(f_i)$, $\hat{p}_i(f_i)$, and $q_{s_i}(f_i)$ share the same moments up to the second order. If $p(y_i|f_i)$ are bounded, the objective is bounded from below and consequently there exists stationary points satisfying these expectation consistency constraints (Minka, 2001b; Opper and Winther, 2005). In the case of multiple stationary points the solution with the smallest free energy can be chosen.

Since the first two terms in (10) are concave functions of $\boldsymbol{\lambda}_{-}$ and $\tilde{\boldsymbol{\lambda}}$ the inner maximization problem is concave with respect to $\boldsymbol{\lambda}_{-}$ (or equivalently $\tilde{\boldsymbol{\lambda}}$) after substitution of the constraints $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_{s_i} - \boldsymbol{\lambda}_{-}$ (Opper and Winther, 2005). The Hessian of the first term with respect to $\boldsymbol{\lambda}_{-}$ is well defined (and negative semi-definite) only if the tilted distributions $\hat{p}_i(f_i) \propto p(y_i|f_i)q_{-i}(f_i)$ are proper probability distributions with finite moments up to the fourth order. Therefore, to ensure that the product of $q_{-i}(f_i)$ and the Student-*t* site $p(y_i|f_i)$ has finite moments and that the inner-loop moment matching remains meaningful, the cavity precisions τ_{-i} have to be kept positive. Furthermore, since the cavity distributions can be regarded as estimates for the leave-one-out (LOO) distributions of the latent values, $\tau_{-i} = 0$ would correspond to a situation where $q(f_i|\mathbf{y}_{-i}, \mathbf{X})$ has infinite variance, which does not make sense given the Gaussian prior assumption (1). On the other hand, $\tilde{\tau}_i$ may become negative for example when the corresponding observation y_i is an outlier (see Section 5).

3.7 Fractional EP Updates

Fractional EP (or power EP, Minka, 2004) is an extension of EP which can be used to reduce the computational complexity of the algorithm by simplifying the tilted moment evaluations and to improve the robustness of the algorithm when the approximation family is not flexible enough (Minka, 2005) or when the propagation of information is difficult due to vague prior information (Seeger, 2008). In fractional EP the cavity distributions are defined as $q_{-i}(f_i) \propto q(f_i|\mathcal{D}, \theta, \nu, \sigma^2)/\tilde{t}_i(f_i)^{\eta}$ and the tilted distribution as $\hat{p}_i(f_i) \propto q_{-i}(f_i)p(y_i|f_i)^{\eta}$ for a fraction parameter $\eta \in (0, 1]$. The site parameters are updated so that the moments of $q_{-i}(f_i)\tilde{t}_i(f_i)^{\eta} \propto q(f_i)$ match with $q_{-i}(f_i)p(y_i|f_i)^{\eta}$. Otherwise the procedure is similar and standard EP can be recovered by setting $\eta = 1$. In fractional EP the natural parameters of the cavity distribution are given by

$$\tau_{-i} = \sigma_i^{-2} - \eta\tilde{\tau}_i \quad \text{and} \quad \nu_{-i} = \sigma_i^{-2}\mu_i - \eta\tilde{\nu}_i, \tag{11}$$

and the site updates (with damping factor δ) by

$$\Delta\tilde{\tau}_i = \delta\eta^{-1}(\hat{\sigma}_i^{-2} - \sigma_i^{-2}) \quad \text{and} \quad \Delta\tilde{\nu}_i = \delta\eta^{-1}(\hat{\sigma}_i^{-2}\hat{\mu}_i - \sigma_i^{-2}\mu_i). \tag{12}$$

The fractional update step $\min_q \mathrm{KL}(\hat{p}_i(f_i)||q(f_i))$ can be viewed as minimization of the α-divergence with $\alpha = \eta$ (Minka, 2005). Compared to the KL-divergence, minimizing the α-divergence with $0 < \alpha < 1$ does not force $q(f_i)$ to cover as much of the probability mass of $\hat{p}_i(f_i)$ whenever $\hat{p}_i(f_i) > 0$. As a consequence, fractional EP tends to underestimate the variance and normalization constant of $q_{-i}(f_i)p(y_i|f_i)^{\eta}$, and also the approximate marginal likelihood Z_{EP}. On the other hand, we also found that minimizing the KL-divergence in standard EP may overestimate the marginal likelihood with some data sets. In case of multiple modes, the approximation tries to represent the overall uncertainty in $\hat{p}_i(f_i)$ the more exactly the closer α is to 1. In the limit $\alpha \to 0$ the reverse KL-divergence is obtained which is used in some form, for example, in the fVB and KL approximations (Nickisch and Rasmussen, 2008). Also the double-loop objective function (10) can be modified according to the different divergence measure of fractional EP (Cseke and Heskes, 2011; Seeger and Nickisch, 2011).

Fractional EP has some benefits over standard EP with the non-log-concave Student-*t* sites. First, when evaluating the moments of $q_{-i}(f_i)p(y_i|f_i)^{\eta}$, setting $\eta < 1$ flattens the likelihood term which alleviates the possible converge problems related to multimodality. This is related to the approximating family being too inflexible and the benefits of different divergence measures in these cases are considered by Minka (2005). Second, the fractional updates help to avoid the cavity precisions becoming too small, or even negative. Equation (11) shows that by choosing $\eta < 1$, a fraction $(1-\eta)$ of the precision $\tilde{\tau}_i$ of the i:th site is left in the cavity. This decreases the cavity variances which in turn makes the tilted moment integrations and the subsequent EP updates (12) more robust. Problems related to cavity precision becoming too small can be present also with log-concave sites when the prior information is vague. For example, Seeger (2008) reports that with an underdetermined linear model combined with a log-concave Laplace prior the cavity precisions remain positive but they may become very small which induces numerical inaccuracies in the analytical moment evaluations. These inaccuracies may accumulate and even cause convergence problems. Seeger (2008) reports that fractional updates improve numerical robustness and convergence in such cases.

4. Robust Implementation of the Parallel EP Algorithm

The sequential EP updates are shown to be stable for models in which the exact site terms (in our case the likelihood functions $p(y_i|f_i)$) are log-concave (Seeger, 2008). In this case, all site variances, if initialized to non-negative values, remain non-negative during the updates. It follows that the variances of the cavity distributions $q_{-i}(f_i)$ are positive and thus also the subsequent moment evaluations of $q_{-i}(f_i)p(y_i|f_i)$ are numerically robust. The non-log-concave Student-*t* likelihood is problematic because both the conditional posterior $p(\mathbf{f}|\mathcal{D},\theta,\nu,\sigma)$ as well as the tilted distributions $\hat{p}_i(f_i)$ may become multimodal. Therefore extra care is needed in the implementation and these issues are discussed in this section.

The double-loop algorithm is a rigorous approach that is guaranteed to converge to a stationary point of the objective function (10) when the site terms $p(y_i|f_i)$ are bounded from below. The downside is that the double-loop algorithm can be much slower than for example parallel EP because it spends much computational effort during the inner loop iterations, especially in the early stages when $q_{s_i}(f_i)$ are poor approximations for the true marginals. An obvious improvement would be to start with damped parallel updates and to continue with the double-loop method if necessary. Since in our experiments parallel EP has proven quite efficient with many easier data sets, we adopt this approach and propose few modifications to improve the convergence in difficult cases. A parallel EP initialization and a double-loop backup is also used by Seeger and Nickisch (2011) in their fast EP algorithm.

Parallel EP can also be interpreted as a variant of the double-loop algorithm where only one inner-loop optimization step is done by moment matching (7) and each such update is followed by an outer-loop refinement of the marginal approximations $q_{s_i}(f_i)$. The inner-loop step consists of evaluating the tilted moments $\{\hat{\mu}_i, \hat{\sigma}_i^2 | i = 1,...,n\}$ with $q_{s_i}(f_i) = q(f_i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$, updating the sites (9), and updating the posterior (6). The outer-loop step consists of setting $q_{s_i}(f_i)$ equal to the new marginal distributions $q(f_i)$. Connections between the message passing updates and the double-loop methods together with considerations of different search directions for the inner-loop optimization can be found in the extended version of Heskes and Zoeter (2002). The robustness of parallel EP can be improved by the following modifications.

1. After each moment matching step check that the objective (10) increases. If the objective does not increase, decrease the damping coefficient δ until increase is obtained. The downside is that this requires one additional evaluation of the tilted moments for every site per iteration,

but if these one-dimensional integrals are implemented efficiently this is a reasonable price for stability.

2. Before updating the sites (9) check that the new cavity variances $\tau_{-i} = \tau_{s_i} - (\tilde{\tau}_i + \Delta\tilde{\tau}_i)$ are positive. If they are negative, choose a smaller damping factor δ so that $\tau_{-i} > 0$. This computationally cheap precaution ensures that the increase of the objective (10) can be verified according to modification 1.

3. With modifications 1 and 2 the site parameters can still oscillate (see Section 5 for an illustration) but according to our experiments the convergence is obtained with all hyperparameters values eventually. The oscillations can be reduced by updating $q_{s_i}(f_i)$ only after the moments of $\hat{p}_i(f_i)$ and $q(f_i)$ are consistent for all $i = 1, ..., n$ with some small tolerance, for example 10^{-4}. At each update, check also that the new cavity precisions are positive, and if not, continue the inner-loop iterations with the previous $q_{s_i}(f_i)$ until better moment consistency is achieved or switch to fractional updates. Actually, this modification corresponds to the maximization in (10) and it results in a double-loop algorithm where the inner-loop optimization is done by moment matching (7). If no parallel initialization is done, often during the first 5-10 iterations when the step size δ is limited according to modification 2, the consistency between $\hat{p}_i(f_i)$ and $q(f_i)$ cannot be achieved. This is an indication of $q(\mathbf{f})$ being a too inflexible approximation for the tilted distributions with the current $q_{s_i}(f_i)$. An outer-loop update $q_{s_i}(f_i) = q(f_i)$ usually helps in these cases.

4. If sufficient increase of the objective is not achieved after an inner-loop update (modification 1), use the gradient information to obtain a better step size δ. The gradients of (10) with respect to the site parameters $\tilde{\nu}_i$ and $\tilde{\tau}_i$ can be calculated without additional evaluations of the objective function for fixed $\boldsymbol{\lambda}_s$. With these gradients, it is possible to determine $g(\delta)$, the gradient of the inner-loop objective function with respect to δ in the current search direction. For parallel EP the search direction is defined by (9) with fixed site updates $\Delta\tilde{\tau}_i = \hat{\sigma}_i^{-2} - \sigma_i^{-2}$ and $\Delta\tilde{\nu}_i = \hat{\sigma}_i^{-2}\hat{\mu}_i - \sigma_i^{-2}\mu_i$ for $i = 1, ..., n$. In case of a too large step, $g(\delta)$ becomes negative. Then, for example, spline interpolation with derivative constraints at the end points can be used to approximate the objective as a function of δ. From this approximation a better estimate for the step size δ can be determined efficiently. In case of a too short step, $g(\delta)$ becomes positive and a better step size can be obtained by extrapolating with constraints based on approximate second order derivatives. This modification corresponds to an approximative line search in the concave inner-loop maximization.

In the comparisons of Section 6 we start with 10 damped ($\delta = 0.8$) parallel iterations because with a sensible hyperparameter initialization this is enough to achieve convergence in most hyperparameter optimization steps with the empirical data sets. If no convergence is achieved this parallel initialization also speeds up the convergence of the subsequent double-loop iterations (see Section 5.3). If after any of the initial parallel updates the posterior covariance $\boldsymbol{\Sigma}$ becomes ill-conditioned, that is, many of the $\tilde{\tau}_i$ are too negative, or any of the cavity variances become negative we reject the new site configuration and proceed with more robust updates using the previously described modifications. To reduce the computational costs we limited the maximum number of inner loop iterations (modification 3) to two with two possible additional step size adjustment iterations (modification 4). This may not be enough to suppress all oscillations of the site parameters but in practice more frequent outer loop refinements of $q_{s_i}(f_i)$ were found to require fewer computationally expensive objective evaluations for convergence.

In some rare cases, for example, when the noise level σ is very small, the outer-loop update of $q_{s_i}(f_i)$ may result in negative values for some of the cavity variances even though the inner-loop optimality is satisfied. In practise this means that $[\mathbf{\Sigma}_{ii}]^{-1}$ is smaller than $\tilde{\tau}_i$ for some i. This may be a numerical problem or an indication of a too inflexible approximating family but switching to fractional updates helps. However, in our experiments, this happened only when the noise level was set to too small values and with a sensible hyperparameter initialization such problems did not emerge.

4.1 Other Implementation Details

The EP updates require evaluation of moments $m_k = \int f_i^k g_i(f_i) df_i$ for $k = 0, 1, 2$, where we have defined $g_i(f_i) = q_{-i}(f_i) p(y_i|f_i)^\eta$. With the Student-$t$ likelihood and an arbitrary $\eta \in (0, 1]$ numerical integration is required. Instead of the standard Gauss quadrature we used the adaptive Gauss-Kronrod quadrature described by Shampine (2008) because it can save function evaluations by reusing the existing nodes during the adaptive interval subdivisions. For further computational savings all the required moments were calculated simultaneously using the same function evaluations. The integrand $g_i(f_i)$ may have one or two modes between the cavity mean μ_{-i} and the observation y_i. In the two-modal case the first mode is near μ_{-i} and the other near $\mu_\infty = \sigma_\infty^2(\sigma_{-i}^{-2}\mu_{-i} + \eta_i\sigma^{-2}y_i)$, where μ_∞ and $\sigma_\infty^2 = (\sigma_{-i}^{-2} + \eta_i\sigma^{-2})^{-1}$ correspond to the mean and variance of the limiting Gaussian tilted distribution as $\nu \to \infty$. The integration limits were set to $\min(\mu_{-i} - 6\sigma_{-i}, \mu_\infty - 10\sigma_\infty)$ and $\max(\mu_{-i} + 6\sigma_{-i}, \mu_\infty + 10\sigma_\infty)$ to cover all the relevant mass around the both possible modes.

Both the hyperparameter estimation and monitoring the convergence of EP requires that the marginal likelihood $q(\mathbf{y}|\mathbf{X}, \theta, \sigma^2, \nu)$ can be evaluated in a numerically robust manner. Assuming a fraction parameter η the marginal likelihood is given by

$$\log Z_{\text{EP}} = \frac{1}{\eta} \sum_{i=1}^{n} \left(\log \hat{Z}_i + \frac{1}{2} \log \tau_{s_i} \tau_{-i}^{-1} + \frac{1}{2} \tau_{-i}^{-1} \nu_{-i}^2 - \frac{1}{2} \tau_{s_i}^{-1} \nu_{s_i}^2 \right) - \frac{1}{2} \log |\mathbf{I} + \mathbf{K}\tilde{\mathbf{\Sigma}}^{-1}| - \frac{1}{2} \tilde{\boldsymbol{\nu}}^{\mathrm{T}} \boldsymbol{\mu},$$

where $\nu_{s_i} = \nu_{-i} + \eta\tilde{\nu}_i$ and $\tau_{s_i} = \tau_{-i} + \eta\tilde{\tau}_i$. The first sum term can be evaluated safely if the cavity precisions τ_{-i} and the tilted variances $\hat{\sigma}_i^2$ remain positive during the EP updates because at convergence $\tau_{s_i} = \hat{\sigma}_i^{-2}$.

Evaluation of $|\mathbf{I} + \mathbf{K}\tilde{\mathbf{\Sigma}}^{-1}|$ and $\mathbf{\Sigma} = (\mathbf{K}^{-1} + \tilde{\mathbf{\Sigma}}^{-1})^{-1}$ needs some care because many of the diagonal entries of $\tilde{\mathbf{\Sigma}}^{-1} = \text{diag}[\tilde{\tau}_1, ..., \tilde{\tau}_n]$ may become negative due to outliers and thus the standard approach presented by Rasmussen and Williams (2006) is not suitable. One option is to use the rank one Cholesky updates as described by Vanhatalo et al. (2009) or the LU decomposition as is done in the GPML implementation of the Laplace approximation (Rasmussen and Nickisch, 2010). In our parallel EP implementation we process the positive and negative sites separately. We define $\mathbf{W}_1 = \text{diag}(\tilde{\tau}_i^{1/2})$ for $\tilde{\tau}_i \geq 0$ and $\mathbf{W}_2 = \text{diag}(|\tilde{\tau}_i|^{1/2})$ for $\tilde{\tau}_i < 0$, and divide $\mathbf{K}$ into corresponding blocks $\mathbf{K}_{11}$, $\mathbf{K}_{22}$, and $\mathbf{K}_{12} = \mathbf{K}_{21}^{\mathrm{T}}$. We compute the Cholesky decompositions of two symmetric matrices

$$\mathbf{L}_1\mathbf{L}_1^{\mathrm{T}} = \mathbf{I} + \mathbf{W}_1\mathbf{K}_{11}\mathbf{W}_1 \quad \text{and} \quad \mathbf{L}_2\mathbf{L}_2^{\mathrm{T}} = \mathbf{I} - \mathbf{W}_2(\mathbf{K}_{22} - \mathbf{U}_2\mathbf{U}_2^{\mathrm{T}})\mathbf{W}_2,$$

where $\mathbf{U}_2 = \mathbf{K}_{21}\mathbf{W}_1\mathbf{L}_1^{-\mathrm{T}}$. The required determinant is given by $|\mathbf{I} + \mathbf{K}\tilde{\mathbf{\Sigma}}^{-1}| = |\mathbf{L}_1|^2|\mathbf{L}_2|^2$. The dimension of $\mathbf{L}_1$ is typically much larger than that of $\mathbf{L}_2$ and it is always positive definite. $\mathbf{L}_2$ may not be positive definite if the site precisions have too small negative values, and therefore if the second Cholesky decomposition fails after a parallel EP update we reject the proposed site parameters and reduce the step size. The posterior covariance can be evaluated as $\mathbf{\Sigma} = \mathbf{K} - \mathbf{U}\mathbf{U}^{\mathrm{T}} + \mathbf{V}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U} = [\mathbf{K}_{11}, \mathbf{K}_{12}]^{\mathrm{T}} \mathbf{W}_1 \mathbf{L}_1^{-\mathrm{T}}$ and $\mathbf{V} = [\mathbf{K}_{21}, \mathbf{K}_{22}]^{\mathrm{T}} \mathbf{W}_2 \mathbf{L}_2^{-\mathrm{T}} - \mathbf{U}\mathbf{U}_2^{\mathrm{T}} \mathbf{W}_2 \mathbf{L}_2^{-\mathrm{T}}$. The regular observations reduce the posterior uncertainty through $\mathbf{U}$ and the outliers increase uncertainty through $\mathbf{V}$.

5. Properties of EP with a Student-t Likelihood

In GP regression the outlier rejection property of the Student-*t* model depends heavily on the data and the hyperparameters. If the hyperparameters and the resulting unimodal approximation (6) are suitable for the data there are usually only a few outliers and there is enough information to handle them given the smoothness assumptions of the GP prior and the regular observations. This is usually the case during the MAP estimation if the hyperparameters are initialized sensibly. On the other hand, unsuitable hyperparameters may produce a very large number of outliers and also considerable uncertainty on whether certain data points are outliers or not. For example, a small ν combined with a too small σ and a too large lengthscale (i.e., a too inflexible model) can result into a very large number of outliers because the model is unable to explain large quantity of the observations. Unsuitable hyperparameters may not necessarily induce convergence problems for EP if there exists only one plausible posterior hypothesis capable of handling the outliers. However, if the conditional posterior distribution has multiple modes, convergence problems may occur unless sufficient amount of damping is used. In some difficult cases either fractional updates or double-loop iterations may be needed to achieve convergence. In this section we discuss the convergence properties of EP with the Student-*t* likelihood, demonstrate the effects of the different EP modifications described in the sections 3 and 4, and also compare the quality of the EP approximation to the other methods described in Section 3 with the help of simple regression examples.

An outlying observation y_i increases the posterior uncertainty on the unknown function at the input space regions a priori correlated with $\mathbf{x}_i$. The amount of increase depends on how far the posterior mean estimate of the unknown function value, $\mathrm{E}(f_i|\mathcal{D})$, is from the observation y_i. Some insight into this behavior is obtained by considering the negative Hessian of $\log p(y_i|f_i,\nu,\sigma^2)$, that is, $W_i = -\nabla^2_{f_i} \log p(y_i|f_i)$, as a function of f_i (compare to the Laplace approximation in Section 3.2). W_i is positive when $y_i - \sigma\sqrt{\nu} < f_i < y_i + \sigma\sqrt{\nu}$, attains its negative minimum when $f_i = y_i \pm \sigma\sqrt{3\nu}$ and approaches zero as $|f_i| \to \infty$. Thus, with the Laplace approximation, y_i satisfying $\hat{f}_i - \sigma\sqrt{\nu} < y_i < \hat{f}_i + \sigma\sqrt{\nu}$ can be interpreted as regular observations because they decrease the posterior covariance $\mathbf{\Sigma}_{\mathrm{LA}}^{-1}$ in Equation (4). The rest of the observations increase the posterior uncertainty and can therefore be interpreted as outliers. Observations that are far from the mode $\hat{f}_i$ are clear outliers in the sense that they have very little effect on the posterior uncertainty. Observations that are close to $\hat{f}_i \pm \sigma\sqrt{3\nu}$ are not clearly outlying because they increase the posterior uncertainty the most. The most problematic situations arise when the hyperparameters are such that many $\hat{f}_i$ are close to $y_i \pm \sigma\sqrt{3\nu}$. However, despite the negative $\mathbf{W}_{ii}$, the covariance matrix $\mathbf{\Sigma}_{LA}$ is positive definite if $\hat{\mathbf{f}}$ is a local maximum of the conditional posterior.

EP behaves similarly as well. If there is a disagreement between the cavity distribution $q_{-i}(f_i) = \mathcal{N}(\mu_{-i}, \sigma^2_{-i})$ and the likelihood $p(y_i|f_i)$ but the observation is not a clear outlier, the uncertainty in the tilted distribution increases towards the observation and the tilted distribution can even become two-modal. The moment matching (7) results in an increase of the marginal posterior variance, $\hat{\sigma}_i^2 > \sigma_i^2$, which causes $\tilde{\tau}_i$ to decrease (9) and possibly to become negative. Sequential EP usually runs smoothly when all the outliers are clear and $p(\mathbf{f}|\mathcal{D},\theta,\nu,\sigma^2)$ has a unique mode. The site precisions corresponding to the outlying observations may become negative but their absolute values remain small compared to the site precisions of the regular observations. However, if some of the negative sites become very small they may notably decrease the approximate marginal precisions $\tau_i = \sigma_i^{-2}$ of the a priori dependent sites because of the prior correlations defined by $\mathbf{K}$. It follows that the uncertainty in the cavity distributions may increase considerably, that is, the cavity precisions, $\tau_{-i} = \tau_i - \tilde{\tau}_i$, may become very small or negative. This may cause both stability and convergence problems which will be illustrated in the following sections with the help of simple regression examples.

5.1 Simple Regression Examples

Figure 1 shows two one-dimensional regression problems in which standard EP may run into problems. In example 1 (the left subfigures), there are two outliers y_1 and y_2 providing conflicting information in a region with no regular observations ($1 < x < 3$). In this example the posterior mass of the length-scale is concentrated to sufficiently large value so that the GP prior is stiff and keeps the marginal posterior $p(\mathbf{f}|\mathcal{D})$ (shown in the lower left panel) and the conditional posterior $p(\mathbf{f}|\mathcal{D},\hat{\theta},\hat{\nu},\hat{\sigma}^2)$ at the MAP estimate unimodal. Both sequential and parallel EP converge with the MAP estimate for the hyperparameters.

The corresponding predictive distribution is visualized in the upper left panel of Figure 1 showing a considerable increase in the posterior uncertainty when $1 < x < 3$. The lower left panel shows comparison of the predictive distribution of $f(x)$ at $x = 2$ obtained with the different approximations described in Section 3. The hyperparameters are estimated separately for each method. The smooth MCMC estimate of the predictive density of the latent value $f_* = f(x_*)$ at input location x_* is calculated by integrating analytically over $\mathbf{f}$ for each posterior draw of the residual variances $\mathbf{V}$ and averaging the resulting Gaussian distributions $q(f_*|x_*,\mathbf{V},\theta)$. The MCMC estimate (with integration over the hyperparameters) is unimodal but shows small side bumps when the latent function value is close to the observations y_1 and y_2. The standard EP estimate covers well the posterior uncertainty on the latent value but both the Laplace method and fVB underestimate it. At the other input locations where the uncertainty is small, all methods give very similar estimates.

Even though EP remains stable in example 1 with the MAP estimates of the hyperparameters, it is not stable with all hyperparameter values. If ν and σ^2 were sufficiently small, so that the likelihood $p(y_i|f_i)$ was narrow as a function of f_i, and the length-scale was small inducing small correlations between inputs far apart, there would be significant posterior uncertainty about the unknown $f(x)$ when $1 < x < 3$ and the true conditional posterior would be multimodal. Due to the small prior covariances of the observations y_1 and y_2 with the other data points $y_3,\ldots,y_n$, the cavity distributions $q_{-1}(f_1)$ and $q_{-2}(f_2)$ would differ strongly from the approximative marginal posterior distributions $q(f_1)$ and $q(f_2)$. This difference would lead to a very small (or even negative) cavity precisions τ_{-1} and τ_{-2} during the EP iterations which causes stability problems as will be illustrated in section 5.2.

The second one-dimensional regression example, visualized in the upper right panel of Figure 1, is otherwise similar with example 1 except that the nonlinearity of the true function is much stronger when $-5 < x < 0$, and the observations y_1 and y_2 are closer in the input space. The stronger nonlinearity requires a much smaller length-scale for a good data fit and the outliers y_1 and y_2 provide more conflicting information (and stronger multimodality) due to the larger prior covariance. The lower right panel shows comparison of the approximative predictive distributions of $f(x)$ when $x = 2$.

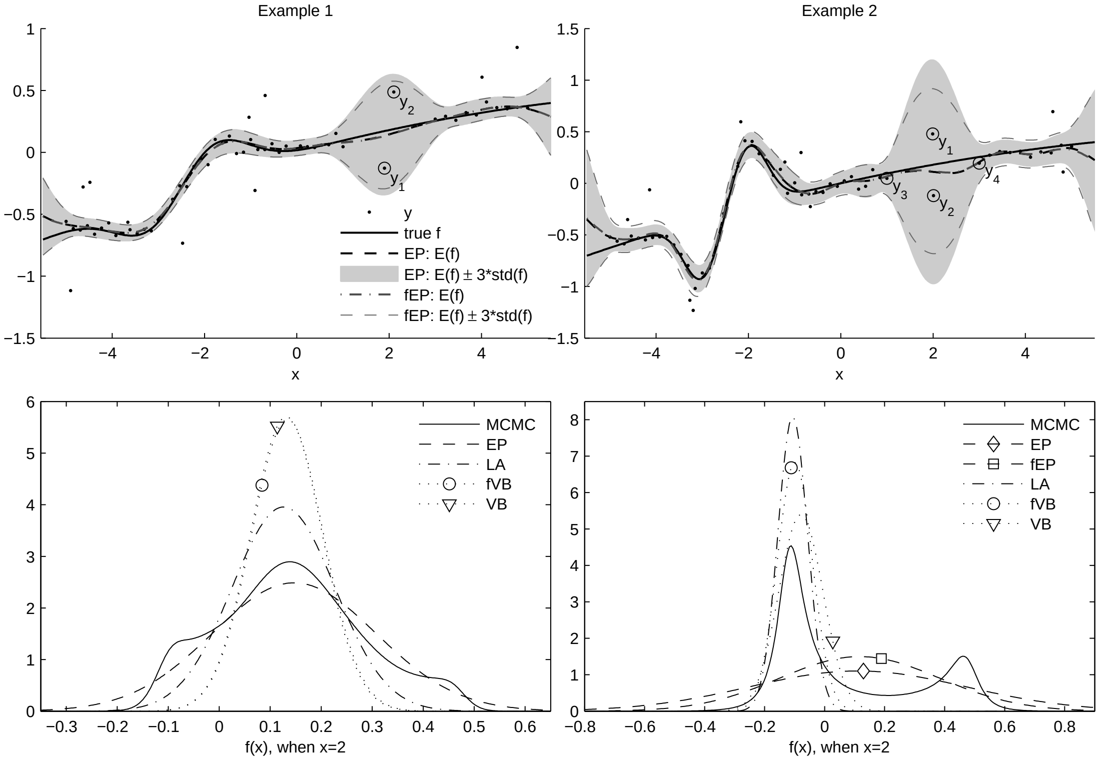
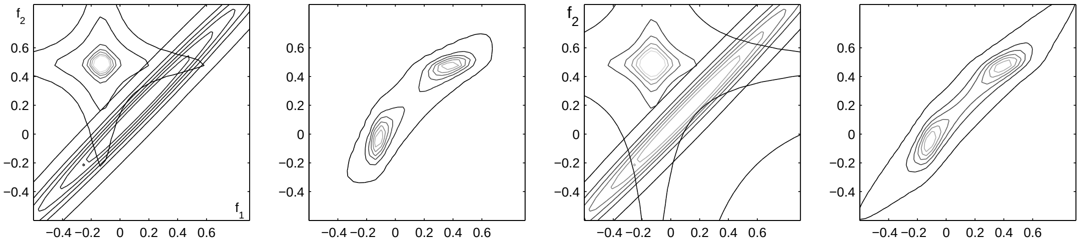


Figure 1: The upper row: Two one-dimensional regression examples, where standard EP may fail to converge with certain hyperparameter values, unless damped sufficiently. The EP approximations obtained by both the regular updates $\eta = 1$ (EP) and the fractional updates $\eta = 0.5$ (fEP) are visualized. The lower row: Comparison of the approximative predictive distributions of the latent value $f(x)$ at $x = 2$. With MCMC all the hyperparameters are sampled and for all the other approximations (except fVB in example 2, see the text for explanation) the hyperparameters are fixed to the corresponding MAP estimates. Notice that the MCMC estimate of the predictive distribution is unimodal in example 1 and multimodal in example 2. With smaller lengthscale values the conditional posterior $p(\mathbf{f}|\mathcal{D}, \theta, \nu, \sigma^2)$ can be multimodal also in example 1.

The MCMC estimate has two separate modes near the observations y_1 and y_2. The Laplace and fVB approximations are sharply localized at the mode near y_1 but the standard EP approximation (EP1) is very wide trying to preserve the uncertainty about the both modes. Contrary to example 1, also the conditional posterior $q(\mathbf{f}|\mathcal{D}, \theta, \nu, \sigma)$ is two-modal if the hyperparameters are set to their MAP-estimates.

5.2 EP Updates with the Student-*t* Sites

Next we discuss the problems with the standard EP updates with the help of example 1. Figure 2 illustrates a two-dimensional tilted distribution of the latent values f_1 and f_2 related to the observations y_1 and y_2 in example 1. A relatively small lengthscale (0.9) is chosen so that there is



(a) $q(f_1,f_2|y_3,...,y_n)$ with $p(y_1|f_1)p(y_2|f_2)$

(b) $q(f_1,f_2|y_3,...,y_n)$ $\times p(y_1|f_1)p(y_2|f_2)$

(c) $q(f_1,f_2|y_3,...,y_n)$ with $p(y_1|f_1)^\eta p(y_2|f_2)^\eta$

(d) $q(f_1,f_2|y_3,...,y_n)$ $\times p(y_1|f_1)^\eta p(y_2|f_2)^\eta$

Figure 2: An illustration of a two-dimensional tilted distribution related to the two problematic data points y_1 and y_2 in example 1. Compared to the MAP value used in Figure 1, shorter lengthscale (0.9) is selected so that the true conditional posterior is multimodal. Panel (a) visualizes the joint likelihood $p(y_1|f_1)p(y_2|f_2)$ together with the generalized 2-dimensional cavity distribution $q(f_1,f_2|y_3,...,y_n)$ obtained by one round of undamped sequential EP updates on sites $\tilde{t}_i(f_i)$, for $i=3,...,n$. Panel (b) visualizes the corresponding two-dimensional tilted distribution $\hat{p}_i(f_1,f_2) \propto q(f_1,f_2|y_3,...,y_n)p(y_1|f_1)p(y_2|f_2)$. Panels (c) and (d) show the same with only a fraction $\eta = 0.5$ of the likelihood terms included in the tilted distribution, which corresponds to fractional EP updates on these sites.

quite strong prior correlation between f_1 and f_2. Suppose that all other sites have already been updated once with undamped sequential EP starting from a zero initialization ($\tilde{\tau}_i = 0$ and $\tilde{\nu}_i = 0$ for $i=1,...,n$). Panel (a) visualizes a generalized 2-dimensional cavity distribution $q(f_1,f_2|y_3,\ldots,y_n)$ together with the joint likelihood $p(y_1,y_2|f_1,f_2) = p(y_2|f_2)p(y_2|f_2)$, and panel (b) shows the contours of the resulting two dimensional tilted distribution which has two separate modes. If the site $\tilde{t}_1(f_1)$ is updated next in the sequential manner with no damping, $\tilde{\tau}_1$ will get a large positive value and the approximation $q(f_1,f_2)$ fits tightly around the mode near the observation y_1. After this, when the site $\tilde{t}_2(f_2)$ is updated, it gets a large negative precision, $\tilde{\tau}_2 < 0$, since the approximation needs to be expanded towards the observation y_2. It follows that, the marginal precision of f_1 is updated to a smaller value than $\tilde{\tau}_1$. Therefore, during the second sweep the cavity precision $\tau_{-1} = \sigma_1^{-2} - \tilde{\tau}_1$ becomes negative, and site 1 can no longer be updated. If the EP updates were done in parallel, both the cavity and the site precisions would be positive after the first posterior update, but $q(f_1,f_2)$ would be tightly centered between the modes. After a couple of parallel loops over all the sites, one of the problematic sites gets a too small negative precision because the approximation needs to be expanded to cover all the marginal uncertainty in the tilted distributions which leads to a negative cavity precision for the other site.

Skipping updates on the sites with negative cavity variances can keep the algorithm numerically stable (see, for example, Minka and Lafferty, 2002). Also increasing damping reduces $\Delta\tilde{\tau}_i$ so that the negative cavity precisions are less likely to emerge. However, these modifications are not enough to ensure convergence. After a few EP iterations, the marginal posterior distribution of a problematic site, for instance $q(f_1)$, is centered between the observations (see, for example, Figure 1). At the same time, the respective cavity distribution, $q_{-1}(f_1)$, is centered near the other problematic observation, y_2. Combining such cavity distribution with the likelihood term, $p(y_1|f_1)$, gives a tilted distribution with significant mass around both observations. If the site precisions, $\tilde{\tau}_1$ and $\tilde{\tau}_2$, are sufficiently large (corresponding to a tight posterior approximation), the variance of the tilted distribution will be larger than that of the marginal posterior and thus the site precision, $\tilde{\tau}_1$ will be decreased. The same happens for the other site. The site precisions are decreased for a few iterations after which the posterior marginals are so wide that the variances of the tilted distributions are smaller than the posterior marginal variances. At this point the site precisions start again to increase gradually. This leads to oscillation between small and large site precisions as illustrated in Figure 3.

With a smaller δ the oscillations are slower and with a sufficiently small δ the amplitude of the oscillations may gradually decrease leading to convergence, as in the panel (b) of Figure 3. However, the convergence is not guaranteed since the conditions of the inner-loop maximization in (10) are not guaranteed to be fulfilled in sequential or parallel EP. For example, a sequential EP update can be considered as a one inner-loop step where only one site is updated, followed by an outer-loop step which updates all the marginal posteriors as $q_{s_i}(f_i) = q(f_i)$. Since the update of one site does not maximize the inner-loop objective, the conditions used to form the upper bound of the convex part in (10) are not met (Opper and Winther, 2005). Therefore, the outer-loop objective is not guaranteed to decrease and the new approximate marginal posteriors may be worse than in the previous iteration.

Example 2 is more difficult in the sense that convergence requires damping at least with $\delta = 0.5$. With sequential EP the convergence depends also on the update order of the sites and $\delta < 0.3$ is needed for convergence with all permutations. Furthermore, if the double-loop approach of Section 4 is considered, the best step size, that minimizes the inner-loop objective in the current search direction, can change (and also increase) considerably between subsequent inner-loop iterations which makes the continuous step-size adjustments very useful.

Also fractional updates improve the stability of EP. Figures 2(c)–(d) illustrate the same approximate tilted distribution as Figures 2(a)–(b) but now only a fraction $\eta = 0.5$ of the likelihood terms are included. This corresponds to the first round fractional updates on these sites with zero initialization. Because of the flattened likelihood $p(y_1|f_1)^\eta p(y_2|f_2)^\eta$ the 2-dimensional tilted distribution is still two-modal but less sharply peaked compared to standard EP on the left. It follows that also the one-dimensional tilted distributions have smaller variances and the consecutive fractional updates (12) of the sites 1 and 2 do not widen the marginal variances σ_1^2 and σ_2^2 as much. This helps to keep the cavity precisions positive by increasing the approximate marginal posterior precisions and reducing the possible negative increments on the site precisions $\tilde{\tau}_1$ and $\tilde{\tau}_2$. This is possible because the different divergence measure allows for a more localized approximation at $1 < x < 3$. In addition, the property that a fraction $(1-\eta)$ of the site precisions is left in the cavity distributions helps to keep the cavity precisions positive during the algorithm. Figure 1 shows a comparison of standard (EP) and fractional EP (fEP, $\eta = 0.5$) with the MAP estimates of the hyperparameters. In example 1 both methods produce very similar predictive distribution because the posterior is unimodal. In example 2 (the lower right panel) fractional EP gives a much smaller predictive uncertainty estimate when $x = 2$ than standard EP which in turn puts more false posterior mass in the tails when compared to MCMC.

The practical guidelines presented in Section 4 bring additional stability in the above described problematic situations. Modification 1 helps to avoid immediate problems from a too large step size by ensuring that each parallel EP update increases the inner-loop objective defined by (10). Modification 2 reduces the step size δ so that the cavity variances, defined as $\tau_{-i} = \tau_{s_i} - \tilde{\tau}_i$ with fixed $\boldsymbol{\lambda}_s = \{\nu_{s_i}, \tau_{s_i}\}$, will remain positive during the inner-loop updates. Modification 3 reduces the oscillations by ensuring that the inner-loop maximization is done within some tolerance, that is,

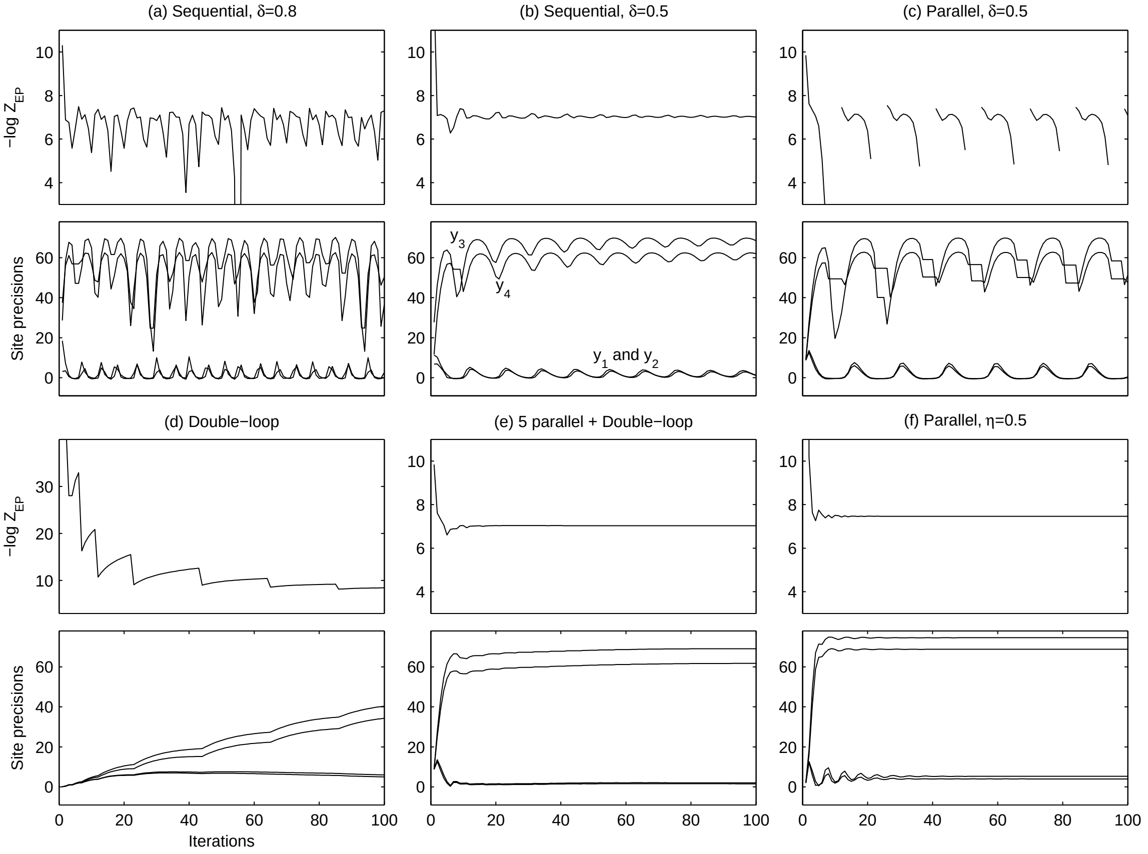


Figure 3: A convergence comparison between sequential and parallel EP as well as the double-loop algorithm in example 2 (the right panel in Figure 1). For each method both the objective $-\log Z_{EP}$ and the site precisions $\tilde{\tau}_i$ related to data points $y_1, ..., y_4$ (see Figure 1) are shown. See Section 5.3 for explanation.

the moments of $\hat{p}_i(f_i)$ and $q(f_i)$ are consistent for fixed $\boldsymbol{\lambda}_s$ before updating $q_{s_i}(f_i)$. For example, a poor choice of δ may require many iterations for achieving inner-loop consistency in the examples 1 or 2, and a too large δ can easily lead to a decrease of the inner-loop objective function or even negative cavity precisions for the sites 1 or 2. Finally, if an unsuccessful update is made due to an unsuitable δ, modification 4 enables automatic determination of a better step size by making use of the concavity of the inner-loop maximization as well as the tilted and marginal moments evaluated at the previous steps with the same $\boldsymbol{\lambda}_s$.

5.3 Convergence Comparisons

Figure 3 illustrates the convergence properties of the different EP algorithms using the data from example 2. The hyperparameters were set to: $\nu = 2$, $\sigma = 0.1$, $\sigma_{\text{se}} = 3$ and $l_k = 0.88$. Panel (a) shows the negative marginal likelihood approximation during the first 100 sweeps with sequential EP and the damping set to $\delta = 0.8$. The panel below shows the site precisions corresponding to the observations $y_1,...,y_4$ marked in the upper right panel of Figure 1. With this damping level the site parameters keep oscillating with no convergence and there are also certain parameter values between iterations 50-60 where the marginal likelihood is not defined because of negative cavity precisions (the updates for such sites are skipped until the next iteration). Whenever $\tilde{\tau}_1$ and $\tilde{\tau}_2$ become very small they also inflict large decrease in the site precisions of the nearby sites 3 and 4. These fluctuations affect other sites the more the larger their prior correlations are (defined by the GP prior) with the sites 1 and 2. Panel (b) shows the same graphs with larger amount of damping $\delta = 0.5$. Now the oscillations gradually decrease as more iterations are done but convergence is still very slow. Panel (c) shows the corresponding data with parallel EP and the same amount of damping. The algorithm does not converge and the oscillations are much larger compared to sequential EP. Also the marginal likelihood is not defined at many iterations because of negative cavity precisions.

Panel (d) in Figure 3 illustrates the convergence of the double-loop algorithm with no parallel initialization. There are no oscillations present because the increase of the objective (10) is verified at every iteration and sufficient inner-loop optimality is obtained before proceeding with the outer-loop minimization. However, compared to sequential or parallel EP, the convergence is very slow and it takes over 100 iterations to get the site parameters to the level that sequential EP attains with only a couple of iterations. Panel (e) shows that much faster convergence can be obtained by initializing with 5 parallel iterations and then switching to the double-loop algorithm. There is still some slow drift visible in the site parameters after 20 iterations but changes in the marginal likelihood estimate are very small. Small changes in the site parameters indicate inconsistencies in the moment matching conditions (7) and consequently also the gradient of the marginal likelihood estimate may be slightly inaccurate if the implicit derivatives of $\log Z_{EP}$ with respect to $\boldsymbol{\lambda}_-$ and $\boldsymbol{\lambda}_s$ are assumed zero in the gradient evaluations (Opper and Winther, 2005). Panel (f) shows that parallel EP converges without damping if fractional updates with $\eta = 0.5$ are applied. Because of the different divergence measure the posterior approximation is more localized (see Figure 1) and also the cavity distributions are closer to the respective marginal distributions. It follows that the site precisions related to y_1 and y_2 are larger and no damping is required to keep the updates stable.

5.4 The Marginal Likelihood Approximation

Figure 4 shows contours of the approximate log marginal likelihood with respect to $\log(l_k)$ and $\log(\sigma^2_{\text{se}})$ in the examples of Figure 1. The contours in the first column are obtained by applying first sequential EP with $\delta = 0.8$ and using the double-loop algorithm if it does not converge. The hyperparameter values for which the sequential algorithm does not converge are marked with black dots and the maximum marginal likelihood estimate of the hyperparameters is marked with ($\times$). The second column shows the corresponding results obtained with fractional EP ($\eta = 0.5$) and the corresponding hyperparameter estimates are marked with ($\circ$). For comparison, log marginal likelihood estimates determined with the annealed importance sampling (AIS) (Neal, 2001) are shown in the third column.

In the both examples there is an area of problematic EP updates with smaller length-scales which corresponds to the previously discussed ambiguity about the unknown function near data points y_1 and y_2 in Figure 1. There is also a second area of problematic updates at larger length-scale values in example 2. With larger length-scales the model is too stiff and it is unable to explain large proportion of the data points in the strongly nonlinear region ($-4 < x < -1$) and consequently there

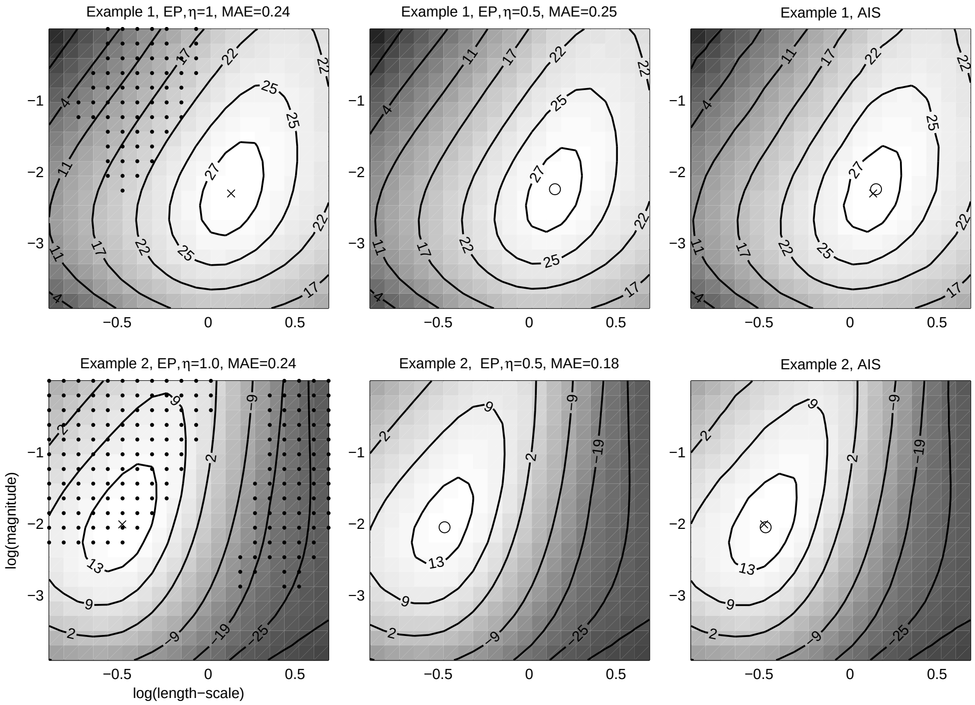


Figure 4: The approximate log marginal likelihood $\log p(\mathbf{y}|\mathbf{X},\theta,\nu,\sigma^2)$ as a function of the log-length-scale $\log(l_k^2)$ and the log-magnitude $\log(\sigma_{\text{se}}^2)$ in the examples shown in Figure 1. The marginal likelihood approximation is visualized with both standard EP ($\eta = 1$) and fractional EP ($\eta = 0.5$). The mode of the hyperparameters is marked with × and ∘ for standard and fractional EP respectively. For comparison the marginal is also approximated by annealed importance sampling (AIS). For both standard and fractional EP the mean absolute errors (MAE) over the region with respect to the AIS estimate are also shown. The noise parameter σ^2 and the degrees of freedom ν are fixed to the MAP-estimates obtained with $\eta = 1$. The hyperparameter values in which sequential EP with $\delta = 0.8$ does not converge are marked with black dots in the two leftmost panels.

exist no unique unimodal solution. It is clear that with the first artificial example the optimization of the hyperparameters with sequential EP can fail if not initialized carefully or not enough damping is used. In the second example the sequential EP approximation corresponding to the MAP values cannot even be evaluated because the mode lies in the area of nonconvergent hyperparameter values. In visual comparison with AIS both standard and fractional EP give very similar and accurate approximations in the first example (the contours are drawn at the same levels for each method). In the second example there are more visible differences: standard EP tends to overestimate the marginal likelihood due to the larger posterior uncertainties (see Figure 1) whereas fractional EP underestimates it slightly. This is congruent with the properties of the different divergence measure used in the moment matching. The difference between the hyperparameter values at the modes between standard and fractional EP is otherwise less than 5% except that in the second example σ and ν are ca. 30% larger with fractional EP.

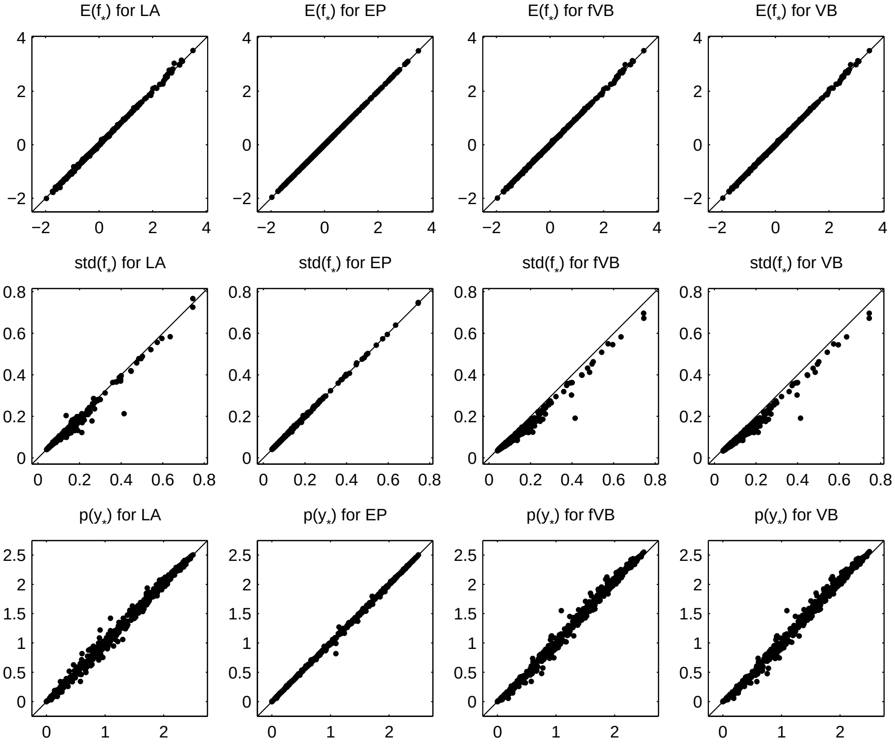


Figure 5: A comparison of the approximative predictive means $\mathrm{E}(f_*|\mathbf{x}_*,\mathcal{D})$, standard deviations $\mathrm{std}(f_*|\mathbf{x}_*,\mathcal{D})$, and predictive densities $q(y_*|\mathbf{x}_*,\mathcal{D})$ provided by the different approximation methods using 10-fold cross-validation on the Boston housing data. The hyperparameters are fixed to the posterior means obtained by a MCMC run on all data. Each dot corresponds to one data point for which the x-coordinate is the MCMC estimate and the y-coordinate the corresponding approximative value obtained with LA, EP, fVB, or VB.

6. Experiments

Four data sets are used to compare the approximative methods: 1) An artificial regression example by Friedman (1991) involving a nonlinear function of 5 inputs. To create a feature selection problem, five irrelevant input variables were added to the data. We generated 10 data sets with 100 training points and 10 randomly selected outliers as described by Kuss (2006). 2) Boston housing data with 506 observations for which the task is to predict the median house prices in the Boston metropolitan area with 13 input variables (see, e.g., Kuss, 2006). 3) Data that involves the prediction of concrete quality based on 27 input variables for 215 experiments (Vehtari and Lampinen, 2002). 4) Data for which the task is to predict the compressive strength of concrete based on 8 input variables for 1030 observations (Yeh, 1998).

6.1 Predictive Comparisons with Fixed Hyperparameters

First we compare the quality of the approximate predictive distributions $q(f_*|\mathbf{x}_*,\mathcal{D},\theta,\nu,\sigma^2)$, where $\mathbf{x}_*$ is the prediction location and $f_* = f(\mathbf{x}_*)$, between all the approximative methods. We ran a full

MCMC on the housing data to determine the posterior mean estimates for the hyperparameters. Then the hyperparameters were fixed to these values and a 10-fold cross-validation was done with all the approximations including MCMC. The predictive means and standard deviations of the latent values as well as the predictive densities of the test observations obtained with Laplace's method (LA), EP, fVB, and VB are plotted against the MCMC estimate in Figure 5. Excluding MCMC, the predictive densities were approximated by numerically integrating over the Gaussian approximation of f_* in $q(y_*|\mathbf{x}_*,\mathcal{D},\theta,\nu,\sigma^2) = \int p(y_*|f_*,\nu,\sigma^2)q(f_*|\mathbf{x}_*,\mathcal{D},\theta,\nu,\sigma^2)df_*$. EP gives the most accurate estimates for all the predictive statistics, and clear differences to MCMC can only be seen in the predictive densities of y_* which indicates that accurate mean and variance estimates of the latent value may not always be enough when deriving other predictive statistics. This contrast somewhat to the corresponding results in GP classification where Gaussian approximation was shown to be very accurate in estimating predictive probabilities (Nickisch and Rasmussen, 2008). Both fVB and VB approximate the mean well but are overconfident in the sense that they underestimate the standard deviations, overestimate the larger predictive densities, and underestimate the smaller predictive densities. LA gives similar mean estimates with the VB approximations, but approximates the standard deviations slightly better especially with larger values. Put together, all methods provide decent estimates with fixed hyperparameters but larger performance differences are possible with other hyperparameter values (depending on the non-Gaussianity of the true conditional posterior) and especially when the hyperparameters are optimized.

6.2 Predictive Comparisons with Estimation of the Hyperparameters

In this section we compare the predictive performance of LA, EP, fVB, VB, and MCMC with estimation of the hyperparameters. The predictive performance was measured with the mean absolute error (MAE) and the mean log predictive density (MLPD). These were evaluated for the Friedman data using a test set of 1000 latent variables for each of the 10 simulated data sets. A 10-fold cross validation was used for the Boston housing and concrete quality data whereas a 2-fold cross-validation was used for the compressive strength data because of the large number of observations. To assess the significance of the differences between the model performances, 95% credible intervals of the MLPD measures were approximated by Bayesian bootstrap as described by Vehtari and Lampinen (2002). Gaussian observation model (GA) is selected as a baseline model for comparisons. With GA, LA, EP, and VB the hyperparameters were estimated by optimizing the marginal posterior densities whereas with MCMC all parameters were sampled. The fVB approach was implemented following Kuss (2006) where the hyperparameters are adapted in the M-step of the EM-algorithm. The variational lower bound associated with the M-step was augmented with the same hyperpriors that were used with the other methods.

Since the MAP inference on the degrees of freedom parameter ν proved challenging due to possible identifiability issues, the LA, EP, fVB, and VB approximations are tested both with ν fixed to 4 (LA1, EP1, fVB1, VB1) and optimized together with the other hyperparameters (LA2, EP2, fVB2, VB2). $\nu = 4$ was chosen as a robust default alternative to the normal distribution which allows for outliers but still has finite variance compared to the extremely wide-tailed alternatives with $\nu \leq 2$. With EP we also tested a simple approach (from now on EP3) to approximate the integration over the posterior uncertainty of ν. We selected 15 values ν_j from the interval $[1.5, 20]$ linearly in the log-log scale and ran the optimization of all the other hyperparameters with ν fixed to these values. The conditional posterior of the latent values was approximated as

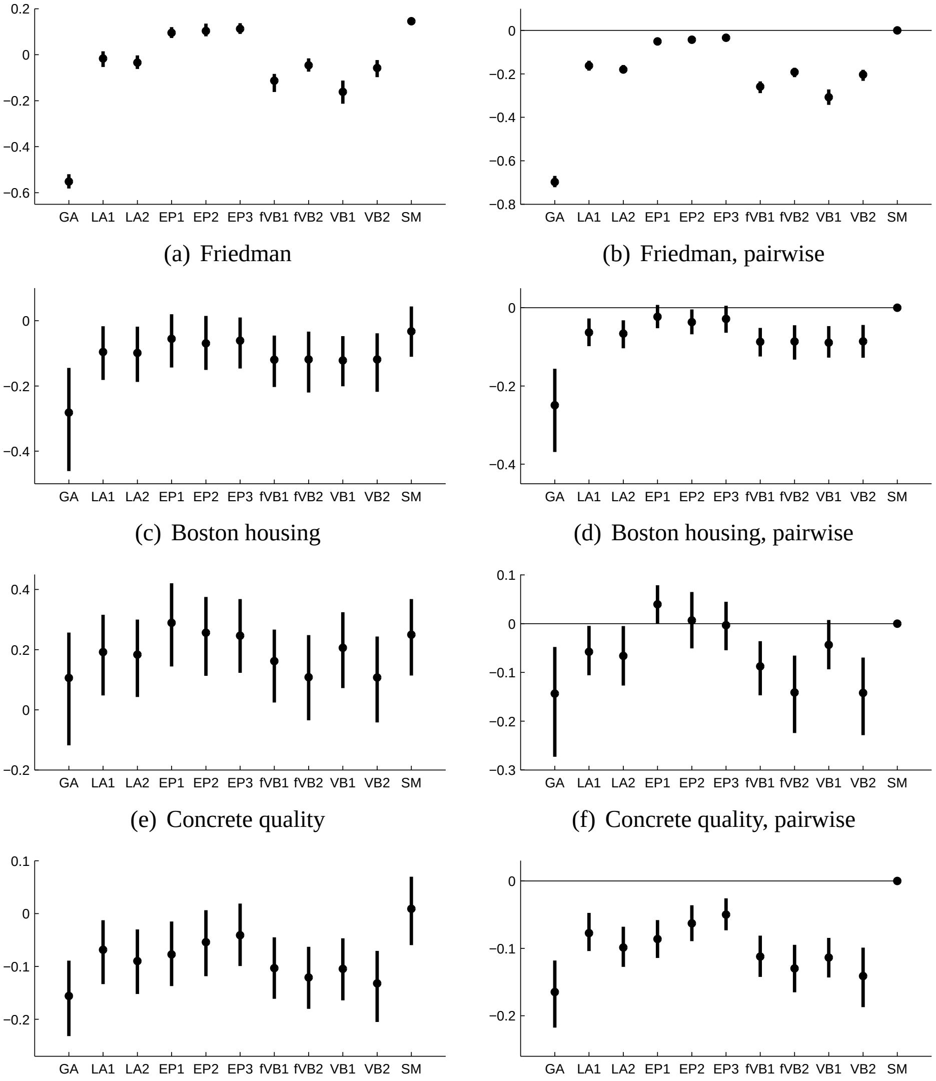
$$p(f_*|\mathbf{x}_*,\mathcal{D}) \approx \sum_j w_j q(f_*|\mathbf{x}_*,\mathcal{D},\theta_j,\sigma_j^2,\nu_j),$$

where $\{\theta_j,\sigma_j^2\} = \arg\max_{\theta,\sigma^2} q(\theta,\sigma^2|\mathcal{D},\nu_j)$ and $w_j = q(\theta_j,\sigma_j^2,\nu_j|\mathcal{D})/\left(\sum_k q(\theta_k,\sigma_k^2,\nu_k|\mathcal{D})\right)$. This can be viewed as a crude approximation of the integration over ν where $p(\theta,\sigma^2|\nu,\mathcal{D})$ is assumed to be very narrowly distributed around the mode. This approximation requires optimization of θ and σ^2 with all the preselected values of ν and therefore θ and σ^2 were initialized to the previous mode to speed up the computations.

The squared exponential covariance (2) was used for all models. Uniform priors were assumed for θ and σ^2 on log-scale and for ν on log-log-scale. The input and target variables were scaled to zero mean and unit variances. ν was initialized to 4, σ to 0.5 and the magnitude σ_{se}^2 to 1. The optimization was done with different random initializations for the length-scales $l_1,...,l_d$ and the result with the highest posterior marginal density $q(\theta,\nu,\sigma^2|\mathcal{D})$ was chosen. The MCMC inference on the latent values was done with both Gibbs sampling based on the scale-mixture model (3) and direct application of the scaled HMC as described by Vanhatalo and Vehtari (2007). The sampling of the hyperparameters was tested with both slice sampling and HMC. The scale-mixture Gibbs sampling (SM) combined with the slice sampling of the hyperparameters resulted in the best mixing of the chains and gave the best predictive performance which is why only those results are reported. The convergence and quality of the MCMC runs was checked by both visual inspections as well as by calculating the potential scale reduction factors, the effective number of independent samples, and the autocorrelation times (Gelman et al., 2004; Geyer, 1992). Based on the convergence diagnostics, burn-in periods were excluded from the beginning of the chains and the remaining draws were thinned to form the final MCMC estimates.

Figures 6(a), (c), (e) and (g) show the MLPD values together with their 95% credible intervals for all the methods in the four data sets. To illustrate the differences between the approximations more clearly figures 6(b), (d), (f) and (h) show the pairwise comparisons of the log posterior predictive densities to SM. The mean values of the pairwise differences together with their 95% credible intervals are visualized. The Student-t model with the SM implementation is significantly better than the Gaussian model with a probability above 95% in all data sets. SM also performs significantly better than all the other approximations on the Friedman and compressive strength data, and on the housing data only EP1 is not significantly worse. The differences are considerably smaller in the concrete quality data on which EP1 actually performs better than SM. One possible explanation for this is a wrong assumption on the noise model (evidence for a covariate dependent noise was found in other experiments). Another possibility is the experimental design used in the data collection; a large proportion of the observations can be classified based on one of the input variables with a very small length scale which is why averaging over this parameter may lead to worse performance.

Additional pairwise comparisons not shown in Figure 6 reveal that either EP1 or EP2 is significantly better than LA, VB, and fVB in all data sets except the compressive strength data for which significant difference is not found when compared to LA1. If the better performing method for estimating ν is selected in either LA, fVB, or VB, LA is better than fVB and VB on the Friedman data and the compressive strength data. No significant differences were found between fVB or VB in pairwise comparisons.



(a) Friedman

(b) Friedman, pairwise

(c) Boston housing

(d) Boston housing, pairwise

(e) Concrete quality

(f) Concrete quality, pairwise

(g) Compressive strength

(h) Compressive strength, pairwise

Figure 6: Left column: The mean log posterior predictive density (MLPD) and its 95% central credible interval. The Gaussian observation model (GA) is shown for reference. The Student-*t* model is inferred with LA, EP, fVB, VB, and scale-mixture based Gibbs sampling (SM). Number 1 after a method means that ν is fixed, number 2 that it is optimized, and number 3 stands for the simple approximative numerical integration over ν. Right column: Pairwise comparisons of the log posterior predictive densities with respect to SM. The mean together with its 95% central credible interval are shown. Values greater than zero indicate that a method is better than SM.

The optimization of ν proved challenging and sensitive to the initialization of the hyperparameters. The most difficult was fVB for which ν often drifted slowly towards infinity. This may be due to our implementation that was made following Kuss (2006) or more likely to the EM style

	GA	LA1	LA2	EP1	EP2	EP3	fVB1	fVB2	VB1	VB2	SM
mean	0.07	1.0	0.8	0.8	7.0	13	15	8.9	1.6	1.8	280
max	0.09	1.0	1.2	1.1	16	26	39	22	3.3	3.8	440
fixed	0.1	1.0		5.5			2.4		1.9		–

Table 1: Two upper rows: The relative CPU times required for the hyperparameter inference. The times are scaled to yield 1 for LA1 separately for each of the four data sets. Both the relative mean (mean) as well as the maximum (max) over the data sets are reported. The third row: The average relative CPU times over the four data sets with the hyperparameters fixed to 28 preselected configurations.

optimization of the hyperparameters. With LA, EP, and VB the integration over $\mathbf{f}$ is redone in the inner-loop for all objective evaluations in the hyperparameter optimization, whereas with fVB the optimization is pursued with fixed approximation $q(\mathbf{f}|\mathcal{D},\theta,\nu,\sigma^2)$. The EP-based marginal likelihood estimate was the most robust with regards to the hyperparameter initialization. According to pairwise comparisons LA2 was significantly worse than LA1 only in the compressive strength data. EP2 was significantly better than EP1 in the housing and compressive strength data but significantly worse with the housing data. With fVB and VB optimization of ν gave significantly better performance only with the simulated Friedman data, and significant decrease was observed with VB2 in the housing and compressive strength data. In pairwise comparisons, the crude numerical integration over ν (EP3) was significantly better than EP1 and EP2 with the housing and compressive strength data, but never significantly worse. These results give evidence that the EP approximation is more reliable in the hyperparameter inference because of the more accurate marginal likelihood estimates which is in line with the results in GP classification (Nickisch and Rasmussen, 2008).

In terms of MAE the Student-*t* model was significantly better than GA in all data sets besides the concrete quality data, in which only EP1 gave better results. If the best performing hyperparameter inference scheme is selected for each method, EP is significantly better than the others on all the data sets excluding the compressive strength data in which the differences were not significant. EP was better than SM on the Friedman and concrete quality data but no other significant differences were found in comparisons with SM. LA was significantly better than fVB and VB on the compressive strength data whereas on the simulated Friedman data VB was better than LA and fVB.

Table 1 summarizes the total CPU times required for the posterior inference including the hyperparameter optimization and the predictions. The CPU times are scaled to give one for LA1 and both the mean and maximum over the four data sets are reported. The running times of the fastest Student-*t* approximations are roughly 10-fold compared to the baseline method GA. EP1, where $\nu = 4$, is surprisingly fast compared to LA but it gets much slower with the optimization of ν (EP2). This is explained by the increasing number of double-loop iterations required to achieve convergence with the larger number of difficult posterior distributions as ν gets smaller values. EP3 is clearly more demanding compared to EP1 or EP2 because the optimization has to be repeated with every preselected value of ν. fVB is quite slow compared to LA or VB because of the slowly progressing EM-based hyperparameter adaptation. The running times of LA and VB are quite similar with ν both fixed and optimized. The running times are suggestive since they depend much on the implementations, convergence thresholds and the hyperparameter initializations. Table 1 shows also the average relative running times over the four data sets (excluding MCMC) with the hyperparameters fixed to 28 different configurations (fixed). The configurations were created by first including the MCMC mean for each data set and then generating all combinations of three clearly different values of ν, σ, and σ_{se} around the MCMC mean with randomly selected lengthscales. The average relative running time is higher with EP because many difficult hyperparameter configurations were created.

7. Discussion

Much research has been done on EP and it has been found very accurate and computationally efficient in many practical applications. Although non-log-concave site functions may be problematic for EP it has been used and found effective for many potentially difficult models such as the Gaussian mixture likelihoods (Kuss, 2006; Stegle et al., 2008) as well as ”spike and slab” priors (Hernández-Lobato et al., 2008). Modifications such as the damping and fractional updates as well as alternative double-loop algorithms have been proposed to improve the stability in difficult cases but the practical implementation issues have not been discussed that much. In this work we have given another demonstration of the good predictive performance of EP in a challenging model but also analyzed the convergence problems and the EP improvements from a practical point of view. In addition, we have presented practical guidelines for a robust parallel EP implementation that can be applied for other non-log-concave likelihoods as well.

We have described the properties of the EP algorithm and its modifications with the Student-*t* observation model, but the same key challenges can also be considered with respect to a general observation model with a non-log-concave likelihood. With a Gaussian prior on $\mathbf{f}$ and a log-concave likelihood, each site approximation increases the posterior precision and all the site precisions remain positive throughout the EP iterations as was shown by Seeger (2008). With a non-log-concave likelihood, however, negative site precisions may occur. The negative site precisions are natural and well justified because a non-log-concave likelihood can generate local increases of the posterior uncertainty which cannot otherwise be modeled with the Gaussian approximation. For example, as discussed here and by Vanhatalo et al. (2009), with the Student-*t* model the negative site precisions correspond to the outlying observations. Through the prior covariances of $\mathbf{f}$, the negative site precisions decrease also the approximate marginal posterior precisions of the other site approximations with positive site precisions. This may become a problem during the sequential or the parallel EP iterations if some of the approximate marginal posterior precisions decrease close to the level of the corresponding site precisions. In such cases the respective cavity precisions become very small which can both induce numerical instabilities in the tilted moment integrations (Seeger, 2008) and make the respective sites very sensitive to the subsequent EP updates. If the EP updates are not constrained some of the cavity precisions may also become negative in which case the tilted moments and the following updates are no longer well defined.

Both of the well-known EP modifications help to alleviate the above described problem. Damping takes more conservative update steps so that the negative site precision increments are less likely to decrease the other cavity precisions too much. Fractional EP keeps the cavity precisions larger by leaving a fraction of the site precisions in the cavity but leads to different approximation which may underestimate the posterior uncertainties. The double-loop algorithm is computationally demanding but admissible steps in the concave inner-loop maximization ensure that the cavity and the tilted distributions remain well defined at all times. And most importantly, the inner-loop maximization forms an upper-bound which provably converges to a stationary solution satisfying the moment matching conditions (7).

The general modifications described in Section 4 bring additional stability with reasonable computational cost. Modification 1 is a principled way to avoid immediate problems arising from a too large step size. It ensures that each parallel EP update results in an increase of the inner-loop objective, and it is computationally cheap with likelihoods for which the tilted moments can be determined analytically (e.g., finite Gaussian mixtures). Modification 2 is also computationally cheap and it ensures that the cavity distributions (defined with fixed $\boldsymbol{\lambda}_s$) remain well defined at all times. If the current step size does not result in a sufficient decrease of the energy the extra tilted moment evaluations required in modification 1 can be used in determining a better step length based on the gradient information according to modification 4 with little additional computational cost.

Modification 3 comes with a considerable computational cost if a small tolerance is required for the inner-loop iterations. However, in our experiments with the Student-*t* model, a relaxed double-loop scheme with a maximum of two inner-loop iterations and two step-size adjustments steps (only if required) were sufficient to achieve convergence. In practice this requires at most three additional matrix inversions per iteration compared to the regular parallel EP but unfortunately also the number of outer loop iterations tended to increase with the more difficult data sets and hyperparameter values. In these cases the main challenge was the difficult inner-loop moment matching which can be partly related to a too inflexible approximating family and partly to a suboptimal search direction defined by parallel EP. Considering the better convergence properties of sequential EP (see Section 5.3), for instance a scheme, where the inner-loop optimization of the more difficult sites (whose cavity distributions differ notably from the respective marginals) was done sequentially and the remaining sites were optimized with parallel updates, could lead to better overall performance.

The nonlinear GP regression combined with the Student-*t* model makes the inference problem challenging because the potential multimodality of the conditional posterior depends on the hyperparameter values. As we have demonstrated by examples, standard EP may not converge with the MAP estimates of the hyperparameters. Therefore, in practical applications, one cannot simply discard all problematic hyperparameter values. Instead some estimate of the marginal likelihood is required also in the more difficult cases. In our examples these situations were related to two modes in the conditional posterior (caused by two outliers) quite far away from each other which requires a very large local increase of the marginal variances from the unimodal posterior approximation. (It should also be noted that moderately damped sequential EP worked fine with many other multimodal posterior distributions.) The globally unimodal assumption is not the best in such cases although the true underlying function is unimodal, but we think that it is important to get some useful posterior approximation. Whether one prefers the possible false certainty provided by the Laplace or VB approximations, or the possible false uncertainty of EP, is a matter of taste but we prefer the latter one.

It is also important that the inference procedure gives some clue of the potential inadequacy of the approximating family so that more elaborate models can be considered. In addition to the examination of the posterior approximation, the need for double-loop iterations with the MAP hyperparameter estimates may be one indication of an unsuitable model. One can also compare the cavity distributions, which can be regarded as the LOO estimates of the latent values, with the respective marginal approximations. If for certain sites most of the LOO information comes from the corresponding site approximations there is reason to suspect that the approximation is not suitable. Our EP implementation enables a robust way of forming such approximations and in case of problems it also enables automatic switching to fractional updates.

The presented EP approach for approximative inference with GP models is implemented in the freely available GPstuff software package (`http://www.lce.hut.fi/research/mm/gpstuff/`). The software also allows experimenting with other non-log-concave likelihoods by implementing the necessary tilted moment integrations in a separate likelihood function.

Acknowledgments

We would like to thank Janne Ojanen and the anonymous reviewers for their valuable comments to improve the manuscript. The study was supported financially by the Academy of Finland (grants 218248, 129230, 129670) and by the Finnish Foundation for Technology Promotion.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science +Business Media LLC, 2006.

Botond Cseke and Tom Heskes. Properties of Bethe free energies and message passing in Gaussian models. *Journal of Artificial Intelligence Research*, 41:1–24, 2011.

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

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

Jerome H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, 1991.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition, 2004.

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

Charles J. Geyer. Practical markov chain monte carlo. *Statistical Science*, 7(12):473–483, 1992.

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

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

José Miguel Hernández-Lobato, Tjeerd Dijkstra, and Tom Heskes. Regulator discovery from gene expression time series of malaria parasites: a hierarchical approach. In J. C. Platt, D. Koller,

Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 649–656. MIT Press, Cambridge, MA, 2008.

Tom Heskes and Onno Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In A. Darwiche and N. Friedman, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, pages 216–233. Morgan Kaufmann, San Francisco, CA, 2002.

Tom Heskes, Manfred Opper, Wim Wiegerinck, Ole Winther, and Onno Zoeter. Approximate inference techniques with expectation constraints. *Journal of Statistical Mechanics: Theory and Experiment*, 2005:P11015, 2005.

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

Chuanhai Liu and Donald B. Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5:19–39, 1995.

Thomas P. Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001a.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI-2001)*, pages 362–369. Morgan Kaufmann, San Francisco, CA, 2001b.

Thomas P. Minka. Power EP. Technical report, Microsoft Research, Cambridge, 2004.

Thomas P. Minka. Divergence measures and message passing. Technical report, Microsoft Research, Cambridge, 2005.

Thomas P. Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-2002)*, pages 352–359. Morgan Kaufmann, San Francisco, CA, 2002.

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

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

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Hannes Nickisch and Carl E. Rasmussen. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 2008.

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

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

Manfred Opper and Ole Winther. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6:2177–2204, 2005.

Carl E. Rasmussen and Hannes Nickisch. Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research*, 11:3011–3015, 2010.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal statistical Society (Series B)*, 71(2):1–35, 2009.

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

Matthias Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

Matthias Seeger and Hannes Nickisch. Fast convergent algorithms for expectation propagation approximate Bayesian inference. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 652–660. JMLR W&CP, 2011.

Lawrence F. Shampine. Vectorized adaptive quadrature in MATLAB. *Journal of Computational and Applied Mathematics*, 211:131–140, 2008.

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

Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

Michael E. Tipping and Neil D. Lawrence. A variational approach to robust Bayesian interpolation. In *In Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing*, pages 229–238. IEEE, 2003.

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

Marcel van Gerven, Botond Cseke, Robert Oostenveld, and Tom Heskes. Bayesian source localization with the multivariate Laplace prior. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1901–1909, 2009.

Jarno Vanhatalo and Aki Vehtari. Sparse log Gaussian processes via MCMC for spatial epidemiology. *JMLR Workshop and Conference Proceedings*, 1:73–89, 2007.

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

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

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

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

I-Cheng Yeh. Modeling of strength of high performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808, 1998.