EXPECTATION PROPAGATION FOR NONSTATIONARY HETEROSCEDASTIC GAUSSIAN PROCESS REGRESSION

Ville Tolvanen[1], *Pasi Jylänki*[2] *and Aki Vehtari*[1] *

[1]Department of Biomedical Engineering and Computational Science, Aalto University
[2]Donders Institute for Brain, Cognition, and Behavior, Radboud University Nijmegen

ABSTRACT

This paper presents a novel approach for approximate integration over the uncertainty of noise and signal variances in Gaussian process (GP) regression. Our efficient and straightforward approach can also be applied to integration over input dependent noise variance (heteroscedasticity) and input dependent signal variance (nonstationarity) by setting independent GP priors for the noise and signal variances. We use expectation propagation (EP) for inference and compare results to Markov chain Monte Carlo in two simulated data sets and three empirical examples. The results show that EP produces comparable results with less computational burden.

1. INTRODUCTION

Gaussian processes (GP) are commonly used as flexible nonparametric Bayesian priors for functions [15]. A typical assumption is that the parameters of the GP model stay constant over the input space. However, this is not reasonable when it is clear from the data that the phenomenon changes over the input space.

As an improvement to these cases, Goldberg [5] proposed heteroscedastic noise inference for Gaussian processes using a second GP to infer the log noise variance and doing the inference by Markov chain Monte Carlo (MCMC). More recent work on heteroscedastic noise models include solving the problem by transformation of the mean and variance parameters to the natural parameters of a Gaussian distribution [9], considering a two-component noise model [11], and an expectation maximization like algorithm [7]. Adams [1] used expectation propagation (EP) [10] to model the input-dependent signal variance (signal magnitude) in GPs by factoring the output signal to a product of a strictly positive modulating signal and a non-restricted signal, with independent GP priors for both of them.

Non-stationarity can also be incorporated via input dependent length-scale as proposed by Gibbs [4] and further developed by Paciorek [14] using MCMC for the approximative inference. In general, the length-scale and the signal variance are underidentifiable and the proportion of them is more important to the predictions[3]. Therefore, we assume that an input-dependent signal variance and an input-dependent length-scale would produce similar predictions and here we focus on the input-dependent signal variance.

We present a straightforward and fast approach to integration over the uncertainty of the noise and signal variance in GP regression using EP. This approach can also be applied to input-dependent noise and signal variance by giving them independent GP priors. We extend the heteroscedastic noise model by Goldberg [5] to EP inference, and extend the nonstationary model by Adams [1] to analytical predictions. The scope of this paper is not to compare GPs with other models for non-stationarity, or to compare EP with other approximate inference methods [7, 8]. Therefore, we focus on EP as an experimentally proven and efficient approximate method and use MCMC as the ground truth to form a proof-of-concept for this novel GP modeling framework. We consider the joint posterior of the modulating signal and the non-restricted signal and show that modeling the posterior correlations leads to significant improvements in the convergence of the EP algorithm compared to a factorized approximation. We also obtain stable analytical gradients of the log marginal likelihood.

We still need to infer other covariance function parameters such as the characteristic length-scale by maximizing the marginal likelihood or posterior density, or using quadrature or MCMC integration. The performance of the EP implementation is compared to full MCMC [12] which produces the exact solution in the limit of an infinite sample size.

In Section 2 we briefly go through Gaussian process regression. Section 3 is dedicated to the models and methods including the EP algorithm for posterior approximation, marginal likelihood evaluation and predictions. The experiments in Section 4 present the performance of our EP approach in two simulated data sets and three empirical problems. We conclude with discussion in Section 5.

2. GAUSSIAN PROCESS REGRESSION

In standard GP regression the output y is modeled as a function f plus some additive noise ϵ such that $y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$. If $\epsilon \sim \mathrm{N}(0, \sigma^2)$, y can be expressed as

$$y(\mathbf{x}) \sim \mathrm{N}(f(\mathbf{x}), \sigma^2). \qquad (1)$$

* We acknowledge the computational resources provided by Aalto Science-IT project.

The function f is given a Gaussian process prior,

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{2}$$

defined by its mean and covariance functions. In this work we use a zero-mean Gaussian processes for notational convenience. As for the covariance function, we use the common squared exponential

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\Big(-\sum_{i=1}^{d} (x_i - x_i')^2 \ell_i^{-2}/2 \Big), \tag{3}$$

where $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$, σ_f^2 is the magnitude or signal variance of the covariance function and ℓ_i is the characteristic length-scale corresponding to the ith input dimension.

Given a data matrix $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, we can write our GP prior for the latent function $f(\mathbf{x}) = \mathbf{f}$ as

$$\mathbf{f} \sim \mathrm{N}(0, K(X, X)) = \mathrm{N}(0, \mathbf{K_f}), \tag{4}$$

where the elements $[\mathbf{K_f}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$ are computed with (3). The available data X and $\mathbf{y}$ are denoted with D.

We focus on models where either the noise variance in (1), or both the noise and signal variances in (1) and (3) depend on the input. These cases are handled analogously to (2), by making the noise and signal variances functions of the input, and the observation a combination of the three resulting signals:

$$\begin{aligned} \log(\sigma^2(\mathbf{x})) &\sim \mathcal{GP}(m_{\mathrm{n}}(\mathbf{x}), k_{\mathrm{n}}(\mathbf{x}, \mathbf{x}')), \\ \log(\sigma_f^2(\mathbf{x})) &\sim \mathcal{GP}(m_{\mathrm{m}}(\mathbf{x}), k_{\mathrm{m}}(\mathbf{x}, \mathbf{x}')). \end{aligned} \tag{5}$$

From now on $\boldsymbol{\theta} = \log(\sigma^2(\mathbf{x}))$ and $\boldsymbol{\phi} = \log(\sigma_f^2(\mathbf{x}))$. We set the GP priors for the logarithms of the variances to handle the positive restriction. We use the squared exponential also for $k_{\mathrm{n}}(\mathbf{x}, \mathbf{x}')$ and $k_{\mathrm{m}}(\mathbf{x}, \mathbf{x}')$, although other covariance functions could be used, too.

3. APPROXIMATE INFERENCE

3.1. Expectation Propagation

Expectation propagation is a general algorithm for forming an approximating distribution (from the exponential family) by matching the marginal moments of the approximating distribution to the marginal moments of the true distribution [10]. The notation in this section follows mainly the notation of Rasmussen & Williams [15].

EP forms a Gaussian approximation to the posterior distribution by approximating the independent non-Gaussian likelihood terms with Gaussian site approximations $\tilde{t}_i$. This enables analytical computation of the posterior distribution because both the likelihood approximation and the prior are Gaussian:

$$p(y_i|f_i) \simeq \tilde{Z}_i \tilde{t}_i(f_i) = \tilde{Z}_i \mathrm{N}(f_i|\tilde{\mu}_i, \tilde{\Sigma}_i), \tag{6}$$

where $\tilde{Z}_i$, $\tilde{\mu}_i$ and $\tilde{\Sigma}_i$ are the parameters of the site approximations, or *site parameters*. We use EP to approximate $p(\mathbf{f}|D)$, such that

$$\begin{aligned} p(\mathbf{f}|D) &= \frac{1}{Z} p(\mathbf{f}|X) \prod_i p(y_i|f_i) \\ &\approx \frac{1}{Z_{\mathrm{EP}}} p(\mathbf{f}|X) \prod_i \tilde{t}_i(f_i) = q(\mathbf{f}|D), \end{aligned} \tag{7}$$

where Z is the normalization constant or *marginal likelihood*, Z_{EP} is the EP approximation to the marginal likelihood, $p(\mathbf{f}|X)$ is the prior of the latent variables $\mathbf{f}$, and $q(\mathbf{f}|D)$ is the Gaussian approximation to the exact posterior distribution $p(\mathbf{f}|D)$.

3.2. Noise Variance

To integrate over the uncertainty of the noise variance in GP regression, we approximate the Gaussian likelihood as a product of two independent Gaussian site approximations for the mean f_i and for the logarithm of the noise variance θ:

$$p(y_i|f_i, \sigma^2) = \mathrm{N}(y_i|f_i, e^{\theta}) \approx \tilde{Z}_i \tilde{t}_i(f_i) \tilde{t}_i(\theta). \tag{8}$$

The posterior approximation of the latent variables $\mathbf{f}$ and θ can now be written in a factorized form, if we set independent prior distributions for $\mathbf{f}$ and θ

$$p(\mathbf{f}, \theta|D) \approx q(\mathbf{f}|D)\, q(\theta|D). \tag{9}$$

3.3. Signal Variance

To use a similar approach for the signal variance, we move the signal variance from the GP prior to the likelihood function. Otherwise we would need to integrate over an n-by-n matrix determinant, which is computationally expensive. We reparameterize $\mathbf{f}$ as $\mathbf{f} = \sigma_f \tilde{\mathbf{f}}$, where σ_f is the square root of the signal variance. Now, if $\mathrm{Cov}[\mathbf{f}] = \sigma_f^2 \mathbf{K}_{\tilde{\mathbf{f}}}$, then $\mathrm{Cov}[\tilde{\mathbf{f}}] = \mathbf{K}_{\tilde{\mathbf{f}}}$, where $\mathbf{K}_{\tilde{\mathbf{f}}}$ is a covariance matrix computed with identity signal variance in (3). Because $\tilde{\mathbf{f}}$ and $e^{\phi/2}$ are multiplied together when modeling the mean of $p(y_i|\tilde{f}_i, \theta, \phi)$, $\tilde{\mathbf{f}}$ and ϕ have strong posterior dependency. Instead of doing a factorized approximation as for the noise variance, we approximate the likelihood with two site approximations: one for the noise variance and a joint two-dimensional Gaussian for $\boldsymbol{v}_i = (\tilde{f}_i, \phi)$:

$$p(y_i|\tilde{f}_i, \theta, \phi) = \mathrm{N}(y_i|e^{\phi/2}\tilde{f}_i, e^{\theta}) \approx \tilde{Z}_i \tilde{t}_i(\tilde{f}_i, \phi)\, \tilde{t}_i(\theta). \tag{10}$$

Assuming independent priors for the latent variables $\tilde{\mathbf{f}}$, ϕ and θ, the posterior approximation is also analogous to the noise variance case

$$p(\tilde{\mathbf{f}}, \theta, \phi|D) \approx q(\tilde{\mathbf{f}}, \phi|D)\, q(\theta|D). \tag{11}$$

We also tested the fully factorized approximation $\tilde{t}_i(\tilde{f}_i, \phi) = \tilde{t}_i(f_i)\tilde{t}_i(\phi)$, but it gave worse predictions, and the EP algorithm needed more iterations to converge.

3.4. Input-Dependent Noise and Signal Variance

We can easily extend the presented likelihood approximations to include also input-dependency on signal and noise variances (or either of them), by setting independent GP priors for both the logarithm of the noise variance and logarithm of the signal variance:

$$p(\boldsymbol{\theta}|X) = \mathrm{N}(0, \mathbf{K}_{\boldsymbol{\theta}}), \qquad p(\boldsymbol{v}|X) = \mathrm{N}(0, \mathbf{K}_{\boldsymbol{v}}). \tag{12}$$

If we integrate over the input-dependent signal variance, we have

$$\mathbf{K}_{\boldsymbol{v}} = \begin{bmatrix} \mathbf{K}_{\tilde{\mathbf{f}}} & 0 \\ 0 & \mathbf{K}_{\boldsymbol{\phi}} \end{bmatrix}, \tag{13}$$

otherwise we have $\mathbf{K}_{\boldsymbol{v}} = \mathbf{K}_{\mathbf{f}}$. By using the GP priors, we assume that the signal and noise variances are also some unknown functions that depend on the input $\mathbf{x}$.

The site approximations are of the same form independent of the input-dependency of the parameters

$$\tilde{t}_i(\theta_i) = \mathrm{N}(\tilde{\mu}_{\theta,i}, \tilde{\Sigma}_{\theta,i}), \qquad \tilde{t}_i(\boldsymbol{v}_i) = \mathrm{N}(\tilde{\boldsymbol{\mu}}_{v,i}, \tilde{\boldsymbol{\Sigma}}_{v,i}). \tag{14}$$

If we integrate over the (input-dependent) signal variance, we have

$$\tilde{\boldsymbol{\mu}}_{v,i} = \begin{bmatrix} \tilde{\mu}_{\tilde{f},i} \\ \tilde{\mu}_{\phi,i} \end{bmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}}_{v,i} = \begin{bmatrix} \tilde{\Sigma}_{\tilde{f},i} & \tilde{\Sigma}_{\tilde{f}\phi,i} \\ \tilde{\Sigma}_{\tilde{f}\phi,i} & \tilde{\Sigma}_{\phi,i} \end{bmatrix}, \tag{15}$$

otherwise $\tilde{\boldsymbol{\mu}}_{v,i} = \tilde{\mu}_{f,i}$ and $\tilde{\boldsymbol{\Sigma}}_{v,i} = \tilde{\Sigma}_{f,i}$. Here we have used Σ for both the scalar variance of the univariate Gaussian and the covariance matrix of the bivariate Gaussian, but it should be clear from the context which one it represents.

The posterior distributions are given by

$$\begin{aligned} q(\boldsymbol{v}|D) = \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) &\propto p(\boldsymbol{v}|X) \prod_i \tilde{t}_i(\boldsymbol{v}_i) \\ &\propto \mathrm{N}(\boldsymbol{v}|0, \mathbf{K}_{\boldsymbol{v}}) \mathrm{N}(\boldsymbol{v}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{v}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}), \qquad (16) \\ q(\boldsymbol{\theta}|D) = \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) &\propto p(\boldsymbol{\theta}|X) \prod_i \tilde{t}_i(\theta_i) \\ &\propto \mathrm{N}(\boldsymbol{v}|0, \mathbf{K}_{\boldsymbol{\theta}}) \mathrm{N}(\boldsymbol{\theta}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}), \qquad (17) \end{aligned}$$

where $\boldsymbol{\mu}_\theta = \boldsymbol{\Sigma}_\theta \tilde{\boldsymbol{\Sigma}}_\theta^{-1} \tilde{\boldsymbol{\mu}}_\theta$, $\boldsymbol{\Sigma}_\theta = (\mathbf{K}_\theta^{-1} + \tilde{\boldsymbol{\Sigma}}_\theta^{-1})^{-1}$, $\boldsymbol{\mu}_{\boldsymbol{v}} = \boldsymbol{\Sigma}_{\boldsymbol{v}} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}^{-1} \tilde{\boldsymbol{\mu}}_{\boldsymbol{v}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{v}} = (\mathbf{K}_{\boldsymbol{v}}^{-1} + \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}^{-1})^{-1}$. The joint site covariance $\tilde{\boldsymbol{\Sigma}}_\theta$ is diagonal while $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$ has a block form if we integrate over the input-dependent signal variance:

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{\tilde{f}} & \tilde{\boldsymbol{\Sigma}}_{\tilde{f}\phi} \\ \tilde{\boldsymbol{\Sigma}}_{\phi\tilde{f}} & \tilde{\boldsymbol{\Sigma}}_{\phi} \end{bmatrix}, \tag{18}$$

where each block is diagonal. Cross-diagonal terms, $\tilde{\boldsymbol{\Sigma}}_{\tilde{f}\phi} = \tilde{\boldsymbol{\Sigma}}_{\phi\tilde{f}}$, collect the marginal covariances $\tilde{\Sigma}_{\tilde{f}\phi,i}$ and the main-diagonal terms, $\tilde{\boldsymbol{\Sigma}}_{\tilde{\mathbf{f}}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\phi}}$, collect the marginal variances $\tilde{\Sigma}_{\tilde{f},i}$ and $\tilde{\Sigma}_{\phi,i}$. If we do not integrate over the signal variance, we have $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} = \tilde{\boldsymbol{\Sigma}}_f$.

3.5. EP Algorithm

The full EP algorithm is presented in Algorithm 1. The main points in the algorithm are the same as in the standard EP approach for Gaussian processes [15, pp. 52–60]. There are some implementation details that should be noted:

1. The overall stability of the EP updates can be improved by using the natural parameterization, $\tilde{\nu} = \tilde{\Sigma}^{-1}\tilde{\mu}$ and $\tilde{\tau} = \tilde{\Sigma}^{-1}$.This way we can avoid inverting the site covariance matrices at every iteration.

2. Even though the algorithm should be stable and robust, there are some cases where the site updates exhibit oscillations, for example, due to weird hyperparameter values.Thus, the updates should be damped after computing the new site approximations in step 4,

$$\begin{aligned} \Delta\tau_i &= \delta(\tau_i^{\text{new}} - \tau_i^{\text{old}}), \quad \Delta\nu_i = \delta(\nu_i^{\text{new}} - \nu_i^{\text{old}}) \\ \tau_i^{\text{new}} &= \tau_i^{\text{old}} + \Delta\tau_i, \quad \nu_i^{\text{new}} = \nu_i^{\text{old}} + \Delta\nu_i, \end{aligned}$$

with some suitable damping factor δ, for example $\delta = 0.8$.

3. In step 3 of the algorithm we minimize KL divergence with respect to Gaussian distributions. This means that we match the first and second moments of the one-dimensional distributions and in addition to these the cross-moment if we have a bivariate Gaussian $\tilde{t}_i(\boldsymbol{v}_i)$. The integrals over f_i or $\tilde{f}_i$ can be computed analytically in every case in steps 2 and 3. If we don't integrate over signal variance, this can be done trivially as both the cavity and likelihood are Gaussian with respect to f_i. If we integrate over signal variance, we can utilize the standard factorization of the multivariate Gaussian $q_{-i}(\tilde{f}_i, \phi_i) = q_{-i}(\tilde{f}_i|\phi_i)q_{-i}(\phi_i)$. The integrals over θ and ϕ must be computed numerically, but this can be done effectively, for example, with Simpson's method.

4. We use parallel EP updates for the site parameters. This means that we compute the site updates for every site approximation before we update the posterior distribution and compute the marginal likelihood. This usually results in a few more EP iterations than sequential EP, but the overall speed of the algorithm is faster.

3.5.1. Marginal Likelihood

Marginal likelihood can be used for model selection under GP framework as it has good calibration and the maximum of the marginal likelihood usually corresponds to good predictions [15, 13, 16]. Marginal likelihood in Gaussian processes is defined as

$$Z = p(\mathbf{y}|X) = \int p(\mathbf{f}|X)p(\mathbf{y}|\mathbf{f})\,\mathrm{d}\mathbf{f}. \tag{19}$$

For our noise and signal variance GPs, an EP approximation to the marginal likelihood is

$$Z_{\text{EP}} = \int p(\boldsymbol{v}|X)p(\boldsymbol{\theta}|X) \prod_i \tilde{Z}_i \tilde{t}_i(\boldsymbol{v}_i)\tilde{t}_i(\theta_i)\,\mathrm{d}\boldsymbol{v}\,\mathrm{d}\boldsymbol{\theta}, \tag{20}$$

Algorithm 1 Parallel EP algorithm

```
Initialize $\tilde{\mu}_{i,\theta} = \tilde{\mu}_{i,v} = \tilde{\Sigma}_{i,\theta}^{-1} = \tilde{\Sigma}_{i,v}^{-1} = 0$ for $i = 1, 2, \ldots, n$.
Set $q(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta}|X)$ and $q(\boldsymbol{v}|D) = p(\boldsymbol{v}|X)$.
repeat
  for $i = 1$ to $n$ do
    if input-dependent signal variance then
      $\boldsymbol{v}_i = (\tilde{f}_i, \phi_i)$
    else
      $\boldsymbol{v}_i = f_i$
    end if
    1. Compute the cavity distributions:
         $q_{-i}(\boldsymbol{v}_i) \propto q_i(\boldsymbol{v}_i)/\tilde{t}_i(\boldsymbol{v}_i)$
         $q_{-i}(\theta) \propto q_i(\theta)/\tilde{t}_i(\theta)$
    with
         $\Sigma_{-i,\cdot}^{-1} = \Sigma_{i,\cdot}^{-1} - \tilde{\Sigma}_{i,\cdot}^{-1}$
         $\mu_{-i,\cdot} = \Sigma_{-i,\cdot}(\Sigma_{i,\cdot}^{-1}\mu_{i,\cdot} - \tilde{\Sigma}_{i,\cdot}^{-1}\tilde{\mu}_{i,\cdot}),$
    when
         $q_i(\cdot) \sim \mathrm{N}(\mu_{i,\cdot}, \Sigma_{i,\cdot})$
         $\tilde{t}_i(\cdot) \sim \mathrm{N}(\tilde{\mu}_{i,\cdot}, \tilde{\Sigma}_{i,\cdot})$
    2. Compute the normalization $\hat{Z}_i$:
         $\hat{Z}_i = \iint p(y_i|\boldsymbol{v}_i, \theta) q_{-i}(\boldsymbol{v}_i) q_{-i}(\theta)\, \mathrm{d}\boldsymbol{v}_i\, \mathrm{d}\theta$
    3. Find the best marginal posterior approximation for $q_i(\boldsymbol{v}_i)$ and $q_i(\theta_i)$ by
         $\min_{q_i(\boldsymbol{v}_i)} \mathrm{KL}(\hat{Z}_i^{-1} p(y_i|\boldsymbol{v}_i, \theta) q_{-i}(\boldsymbol{v}_i) q_{-i}(\theta) \| q_i(\boldsymbol{v}_i))$
         $\min_{q_i(\theta)} \mathrm{KL}(\hat{Z}_i^{-1} p(y_i|\boldsymbol{v}_i, \theta) q_{-i}(\boldsymbol{v}_i) q_{-i}(\theta) \| q_i(\theta)).$
    4. Update the site approximations $\tilde{t}_i$ by
         $\tilde{t}_i(\boldsymbol{v}_i) \propto q_i(\boldsymbol{v}_i)/q_{-i}(\boldsymbol{v}_i)$
         $\tilde{t}_i(\theta) \propto q_i(\theta)/q_{-i}(\theta)$
    analogously to step 1.
  end for
  5. Update the posterior distributions with (16)–(17).
  6. compute the marginal likelihood with (22).
until $Convergence$
```

where $\boldsymbol{v} = (\tilde{\mathbf{f}}, \boldsymbol{\phi})$ or $\boldsymbol{v} = \mathbf{f}$. Following Cseke & Heskes [2], we define the normalization term (log-partition function) of $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as

$$\log Z(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{n}{2}\log(2\pi). \quad (21)$$

Now the EP marginal likelihood approximation can be computed as

$$\begin{aligned}\log Z_{\mathrm{EP}} &= \log Z(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) + \log Z(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v) - \log Z(\mathbf{0}, \mathbf{K}_{\boldsymbol{\theta}}) \\ &+ \sum_i \Big(\log Z(\mu_{-i,\theta}, \Sigma_{-i,\theta}) + \log Z(\mu_{-i,v}, \Sigma_{-i,v}) + \log \hat{Z}_i \\ &- \log Z(\mu_{i,\theta}, \Sigma_{i,\theta}) - \log Z(\mu_{i,v}, \Sigma_{i,v})\Big) - \log Z(\mathbf{0}, \mathbf{K}_{\boldsymbol{v}}),\end{aligned} \quad (22)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the parameters of the posterior distribution approximation $q(\cdot|D)$, μ_i and Σ_i are the ith marginal terms of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, μ_{-i} and Σ_{-i} are the ith marginal mean and variance parameters of the cavity distributions $q_{-i}(\cdot)$, and $\mathbf{K}_j$ are the GP prior covariances.

Note that for θ the marginal parameters are one-dimensional, but for $\boldsymbol{v}$ they are two-dimensional if we integrate over the signal variance like for the site approximations in (15).

3.5.2. Predictions

For predicting a future observation y^* for input $\mathbf{x}^*$, we need to compute the predictive distribution

$$\begin{aligned}p(y^*|\mathbf{x}^*, D) &= \iint p(y^*, \boldsymbol{v}^*, \theta^*|\mathbf{x}^*, D)\, \mathrm{d}\boldsymbol{v}^*\, \mathrm{d}\theta^* \\ &= \iint p(y^*|\boldsymbol{v}^*, \theta^*) q(\boldsymbol{v}^*|\mathbf{x}^*, D) q(\theta^*|\mathbf{x}^*, D)\, \mathrm{d}\boldsymbol{v}^*\, \mathrm{d}\theta^*.\end{aligned} \quad (23)$$

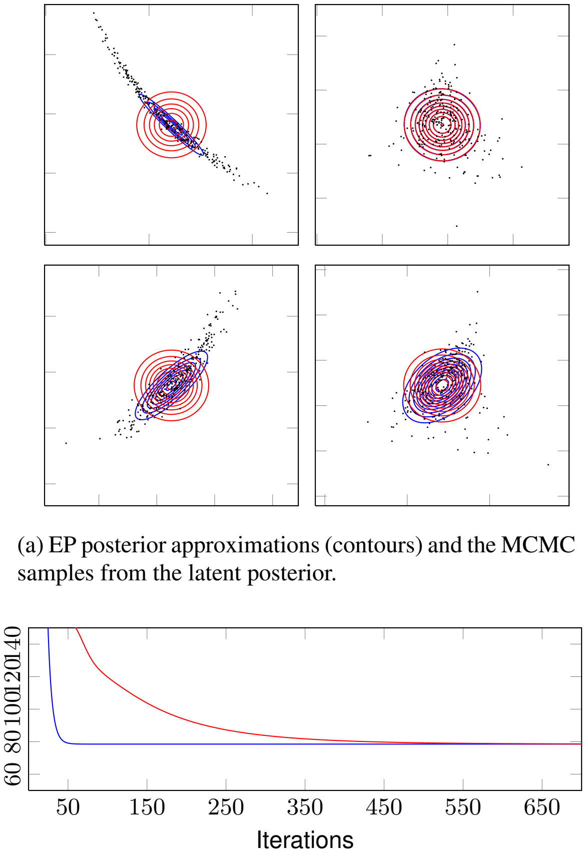
Note that if we assume a stationary signal or noise variance, the respective posterior approximations reduce to one-dimensional Gaussian distributions. This means that $q(\boldsymbol{v}|D)$ becomes $n + 1$ dimensional, and the posterior predictive distribution equals the posterior distribution. Because we approximate the posterior predictive distribution of the latent variables and the predictive distribution of y^* by a Gaussian distribution, we can always compute the predictions analytically, regardless whether we have input-dependent signal or noise variance. For a GP with EP marginalized noise variance, the expected value is given by $\mathbb{E}[y^*] = \mathbb{E}[f^*]$ and the variance by

$$\mathbb{V}[y^*] = \mathbb{V}[f^*] + \exp\big(\mathbb{E}[\theta^*] + \frac{1}{2}\mathbb{V}[\theta^*]\big), \quad (24)$$

where we have omitted dependence on $\{\mathbf{x}^*, D\}$. For a GP with EP marginalized noise and signal variance the results are quite lengthy and are omitted here to save space (see Appendix A).

3.5.3. Factorized Approximation and Converge

Next we discuss certain key properties of the posterior approximations introduced in Sections 3.2-3.4. More precisely, we illustrate the importance of the utilized factorization assumptions in terms of both accuracy and convergence of the resulting EP algorithm.



(a) EP posterior approximations (contours) and the MCMC samples from the latent posterior.

(b) Convergence of EP with factorized (red) and joint (blue) approximations.

Fig. 1. Example comparisons of EP posterior approximations with MCMC samples from the latent posterior and the convergence of the EP algorithm. Red contours correspond to the factorized approximation $q(\tilde{\mathbf{f}}|D)q(\boldsymbol{\theta}|D)$ and the blue contours correspond to the full joint approximation $q(\tilde{\mathbf{f}},\boldsymbol{\phi}|D)$.

Figure 1a visualizes the marginal posterior distributions of the latent values related to both the unscaled function values $\tilde{f}_i$ (x-axis) and the magnitude process ϕ_i (y-axis). Each of the four subplot shows the latent values associated with four different observations (likelihood terms) resulting from a non-trivial simulated data set (see Section 4). MCMC samples from the true posterior distribution are plotted with black dots together with two different EP approximations: the partially coupled approximation $q(\tilde{\mathbf{f}},\boldsymbol{\phi})q(\boldsymbol{\theta})$ introduced in Section 3.4 (blue contours) and a fully factorized approximation of the form $q(\tilde{\mathbf{f}})q(\boldsymbol{\phi})q(\boldsymbol{\theta})$ (red contours). Subplots on the left show strong posterior dependencies between the latent values resulting from the combined effect of the within-observation couplings $f_i = \tilde{f}_i \exp(\phi_i/2)$ and the between-observation correlations controlled by the GP priors. On the other hand, subplots on the right show much weaker couplings indicating that the the within-observation coupling does not necessarily introduce strong posterior dependencies. Comparison of the joint posterior approximations of θ_i with either ϕ_i, $\tilde{f}_i$, or $f_i = \tilde{f}_i \exp(\phi_i/2)$ did not show strong dependencies, which is why we used a factorized approximation for θ to facilitate computations.

According to our experiments, the full factorization does not significantly affect the predictive performance compared to the partially coupled approximation. However, representing these couplings has a significant effect on the convergence properties of the EP algorithm. Figure 1b shows the EP marginal likelihood approximation as a function of EP iterations in both settings. The fully factorized approximation (red line) converges very slowly compared to the partially coupled approximation (blue line); the former requires often hundreds of iterations whereas the partially-coupled approach converges usually in less than 50 iterations. In our experiments the convergence properties of the fully factorized algorithm could not be improved by adjusting damping.

This behavior can be explained by slow propagation of information between the latent values from different likelihood terms with the fully factorized approximation. Because each likelihood term is updated separately from the others, information on the posterior dependencies in other site terms is not available during the update. These findings are fully congruent with the convergence differences in multi-class GP classification when between-class dependencies are omitted [16].

4. EXPERIMENTS

In this section we go through the different data sets we use for experiments, different methods and the assessment criteria for the results. All the experiments were done with modified GPstuff toolbox [18].

Simulated data 1. The first simulated data was generated with:

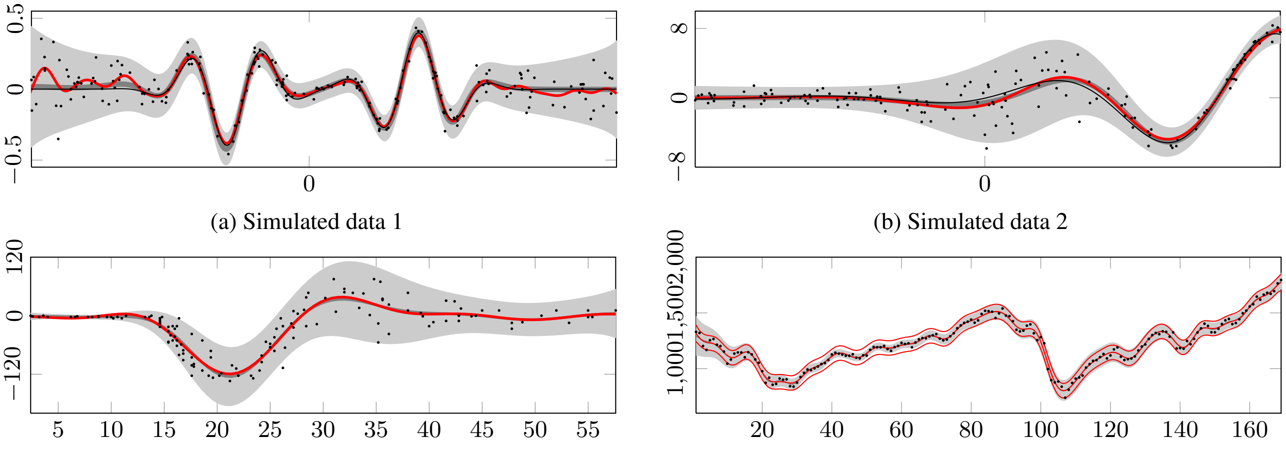
$$\begin{aligned}
\tilde{f}(x) &= \sin(x),\\
\sigma_f(x) &= \mathrm{N}(x|-2.5,1) + \mathrm{N}(x|2.5,1),\\
\sigma(x) &= 0.08 + \mathrm{N}(x|-8,3) + \mathrm{N}(x|8,3),\\
y(x) &= \sigma_f(x)\tilde{f}(x) + \epsilon,
\end{aligned} \tag{25}$$

where $\epsilon \sim \mathrm{N}(0,\sigma(x))$. The training data was generated for 200 random inputs from $\mathrm{U}(-8,8)$. For the test set we used uniform grid of 1000 points in the interval $(-8,8)$ and computed the function values analogously to training set, without adding noise. The experiment was repeated 100 times for different realizations of the training data set to assess the variation in the final predictions of the test set.

Simulated data 2. The second simulated dataset was generated with

$$\begin{aligned}
\tilde{f}(x) &= \sin(x),\\
\sigma_f(x) &= \exp(2\sin(0.2x)),\\
\sigma(x) &= \exp(0.75\sin(0.5x+1)) + 0.1.
\end{aligned} \tag{26}$$

The training and test data were generated analogously to the first experiment. We used 150 training points and the differ-



(a) Simulated data 1

(b) Simulated data 2

(c) Motorcycle

(d) SP500

Fig. 2. One-dimensional data sets and the EP predictions with uncertainty intervals. Thin black lines correspond to the true signal in the simulated data sets, and the thick gray lines are the GP predictions with EP. The grey area is the 95% credible interval of the prediction. Red lines correspond to the standard GP prediction with MAP values for the signal and noise variance (credible intervals only shown for SP500).

ent generating signals for the observations. The second experiment was also repeated 100 times as in the first experiment.

Motorcycle. The motorcycle data [17] consists of 133 accelerometer readings in a simulated motorcycle crash.

Concrete. The second empirical experiment uses concrete quality data [19, 6], where the output is volume percentage of air in concrete, air-%, with 27 different input variables. The input variables depend on the properties of the stone materials, additives and the amount of cement and water.

SP500. The last empirical experiment is concerned with predicting the SP500 index. The data set consists of monthly averages of the index between years 2001–2014, with a total of 169 observations. We demonstrate on this data how a GP with input-dependent noise variance works as a stochastic volatility model.

We compare 8 different methods: **GP** (Standard GP regression), **EP(n)** and **MCMC(n)** (integration over input-dependent noise variance with EP and MCMC), **EP(n+m)** and **MCMC(m+n)** (integration over input-dependent signal and noise variance with EP and MCMC), **EP-MC(n)** and **EP-MC(m+n)** (EP optimized hyperparameters for covariance functions and sampling of the posterior of the latent variables).

In standard GP regression we use *maximum a posteriori* (MAP) values for all the model parameters (signal variance, noise variance, length-scales). In the EP methods, when integrating over input-dependent noise variance, we use MAP values for signal variance and length-scales, and when integrating over input-dependent signal and noise variance, we use MAP values for the length-scales.

We also ran the experiments by integrating over stationary (not input-dependent) signal and noise variances. However, results coincided with standard GP regression, and thus they are not reported here to save space.

The performance of the different methods was assessed by computing the mean log-predictive density (MLPD) for N test data points. For the three empirical datasets, we computed the approximate MLPD of the n training data points with 10-fold cross-validation.

Figure 2 presents the behaviour of the EP (m+n) for the one-dimensional experiments. MLPD values from the experiments are shown in Table 1. We can conclude from the results that integrating over the input-dependent noise variance increases predictive capability greatly in our experiments compared to standard GP regression. Furthermore, integrating over the input-dependent signal variance tends to enhance the predictions even more. In some cases integration over the signal variance is not needed prediction wise, but our results show that even in these cases, it does not harm the predictive quality. The results show that our EP implementation is comparable to the MCMC methods.

The predictive distribution with the SP500 data in Figure 2d illustrates the practical benefits of the input-dependent noise: The period of steady growth between samples 40-80 has clearly lower signal variance compared to the more volatile periods related to financial crisis of 2008 (samples 90-110) and the subsequent shaky growth characterized by debt crises and monetary interventions (samples 110-140).

With our implementation, MCMC was roughly two orders of magnitude slower than EP. This depends highly on the implementation and number of MCMC draws required for convergence. For example, with the SP500 and Concrete data using ARD lengthscales for $\tilde{f}$, the state-of-the-art MCMC methods based on elliptical slice sampling had convergence issues even after thousands of samples, as the results indicate.

Table 1. The table shows MLPD values for different methods, where higher values correspond to better predictions. For the concrete data *ISO* means that we have an isotropic covariance functions for all the latent variables, and *ARD* denotes automatic relevance determination for $\mathbf{f}$ and $\tilde{\mathbf{f}}$, and an isotropic covariance function for the rest of the latent variables.

Method	Simulated 1	Simulated 2	Motorcycle	Concrete (ISO)	Concrete (ARD)	SP500
GP	0.95 ± 0.026	-1.70 ± 0.034	-0.71	0.06	0.11	0.27
EP (n)	1.22 ± 0.025	-1.49 ± 0.032	-0.41	0.13	0.21	0.42
EP (m+n)	1.23 ± 0.028	-1.47 ± 0.029	-0.42	0.22	0.26	0.41
EP-MC (n)	1.22 ± 0.025	-1.49 ± 0.032	-0.40	0.11	0.23	0.43
EP-MC (m+n)	1.24 ± 0.023	-1.47 ± 0.029	-0.41	0.21	0.28	0.42
MCMC	0.95 ± 0.020	-1.70 ± 0.025	-0.71	0.07	0.13	0.28
MCMC (n)	1.22 ± 0.021	-1.55 ± 0.150	-0.39	0.10	0.22	0.19
MCMC (m+n)	1.24 ± 0.019	-1.49 ± 0.030	-0.40	0.20	0.19	0.26

5. DISCUSSION

In this work we have introduced a straightforward but an easily implementable and computationally efficient way to integrate over the uncertainty of the noise and signal variance in Gaussian process regression. Our implementation is easy to apply also for input-dependent noise and signal variance, and it further extends the well-known nonstationary GP models. We have tested our EP implementation on several different data sets and showed that the EP results are on par with state-of-the-art MCMC methods. Furthermore, our results show that EP can be used in complex problems where even the state-of-the-art MCMC methods have convergence problems.

A reference Matlab/Octave implementation of the method is available at http://becs.aalto.fi/en/research/bayes/gpstuff/.

6. REFERENCES

[1] R P Adams and O Stegle, "Gaussian process product models for nonparametric nonstationarity," *ICML*, 2008, pp. 1–8.

[2] B Cseke and T Heskes, "Approximate marginals in latent Gaussian models," *JMLR*, vol. 12, pp. 417–454, 2011.

[3] P J Diggle and P J Ribeiro, *Model-based Geostatistics*, Springer, 2007.

[4] M N Gibbs, *Bayesian Gaussian Processes for Regression and Classification*, Ph.D. thesis, 1997.

[5] P W Goldberg, C K I Williams, and C M Bishop, "Regression with input-dependent noise: A Gaussian process treatment," *NIPS*, vol. 10, pp. 493–499, 1997.

[6] P Jylänki, J Vanhatalo, and A Vehtari, "Gaussian process regression with a student-t likelihood," *JMLR*, vol. 12, pp. 3227–3257, 2011.

[7] K Kersting, C Plagemann, P Pfaff, and W Burgard, "Most likely heteroscedastic Gaussian process regression," *ICML*, 2007, pp. 393–400.

[8] M. Lazaro-Gredilla, and M. Titsias, "Variational heteroscedastic Gaussian process regression," *ICML*, 2011, pp. 841–848.

[9] Q V Le, A J Smola, and S Canu, "Heteroscedastic Gaussian process regression," *ICML*, 2005, pp. 489–496.

[10] T P Minka, "Expectation propagation for approximate Bayesian inference," *UAI*, vol. 17, pp. 362–369, 2001.

[11] A Naish-Guzman and S Holden, "Robust regression with twinned Gaussian processes," *NIPS*, 2007, pp. 1065–1072.

[12] R M Neal, "Regression and classification using Gaussian process priors," in *Bayesian Statistics*. 1998, vol. 6, pp. 475–501, Oxford University Press.

[13] H Nickisch and C E Rasmussen, "Approximations for binary Gaussian process classification," *JMLR*, vol. 9, pp. 2035–2078, 2008.

[14] C J Paciorek and M J Schervish, "Nonstationary covariance functions for Gaussian process Regression," *NIPS*, vol. 16, pp. 273–280, 2004.

[15] C E Rasmussen and C K I Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[16] J Riihimäki, P Jylänki, and A Vehtari, "Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood," *JMLR*, vol. 14, pp. 75–109, 2013.

[17] B W Silverman, "Some aspects of the spline smoothing approach to non-parametric regression curve fitting," *JRSSB*, pp. 1–52, 1985.

[18] J Vanhatalo, J Riihimäki, J Hartikainen, P Jylänki, V Tolvanen, and A Vehtari, "GPstuff: Bayesian modeling with Gaussian processes," *JMLR*, vol. 14, pp. 1175–1179, 2013.

[19] A Vehtari and J Lampinen, "Bayesian model assessment and comparison using cross-validation predictive densities," *Neural Computation*, vol. 14, no. 10, pp. 2439–2468, 2002.

A. DERIVATION OF PREDICTIVE DISTRIBUTION

Here we denote $D = (X, \mathbf{y})$ and $\boldsymbol{v}^* = (\tilde{f}^*, \phi^*)$. Now $p(y^* \mid \tilde{f}^*, \phi^*, \theta^*) = \mathrm{N}(y^* \mid e^{\frac{1}{2}\phi^*}\tilde{f}^*, e^{\theta^*})$ and the Gaussian predictive distribution of y^* can be computed analytically with

$$
\begin{aligned}
&\mathbb{E}[y^* \mid x^*, D] = \iiint y^* p(y^* \mid \boldsymbol{v}^*, \theta^*) q(\boldsymbol{v}^* \mid \mathbf{x}^*, D) q(\theta^* \mid \mathbf{x}^*, D) \mathrm{d}y^* \mathrm{d}\boldsymbol{v}^* \mathrm{d}\theta^* \\
&= \iint e^{\frac{1}{2}\phi^*} \tilde{f}^* q(\tilde{f}^*, \phi^* \mid \mathbf{x}^*, D) \mathrm{d}\tilde{f}^* \mathrm{d}\phi^* \\
&= \iint e^{\frac{1}{2}\phi^*} \tilde{f}^* q(\tilde{f}^* \mid \phi^*, \mathbf{x}^*, D) q(\phi^* \mid \mathbf{x}^*, D) \mathrm{d}\tilde{f}^* \mathrm{d}\phi^* \\
&= \int e^{\frac{1}{2}\phi^*} (\mu_{\tilde{f}^*} + \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} (\phi^* - \mu_{\phi^*})) q(\phi^* \mid \mathbf{x}^*, D) \mathrm{d}\phi^* \\
&= (\mu_{\tilde{f}^*} - \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} \mu_{\phi^*}) e^{\frac{1}{2}\mu_{\phi^*} + \frac{1}{8}\sigma^2_{\phi^*}} + \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} e^{\frac{1}{2}\mu_{\phi^*} + \frac{1}{8}\sigma^2_{\phi^*}} (\mu_{\phi^*} + \frac{1}{2}\sigma^2_{\phi^*}), \\
&\mathbb{E}[(y^*)^2 \mid x^*, D] = \iiint (e^{\phi^*} (\tilde{f}^*)^2 + e^{\theta^*}) q(\tilde{f}^*, \phi^* \mid \mathbf{x}^*, D) q(\theta^* \mid \mathbf{x}^*, D) \mathrm{d}\tilde{f}^* \mathrm{d}\phi^* \mathrm{d}\theta^* \\
&= \iint e^{\phi^*} (\tilde{f}^*)^2 q(\tilde{f}^*, \phi^* \mid \mathbf{x}^*, D) \mathrm{d}\tilde{f}^* \mathrm{d}\phi^* + \int e^{\theta^*} q(\theta^* \mid \mathbf{x}^*, D) \mathrm{d}\theta^* \\
&= \int e^{\phi^*} \left((\mu_{\tilde{f}^*} + \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} (\phi^* - \mu_{\phi^*}))^2 + \sigma^2_{\tilde{f}^*} - \frac{\sigma^2_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} \right) q(\phi^* \mid \mathbf{x}^*, D) \mathrm{d}\phi^* \\
&+ e^{\mu_{\theta^*} + \frac{1}{2}\sigma^2_{\theta^*}} \\
&= \int e^{\phi^*} \big((\mu_{\tilde{f}^*} - \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} \mu_{\phi^*})^2 + \sigma^2_{\tilde{f}^*} - \frac{\sigma^2_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} + 2(\mu_{\tilde{f}^*} - \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} \mu_{\phi^*}) \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} \phi^* \\
&+ \frac{\sigma^2_{\tilde{f}^*\phi^*}}{\sigma^4_{\phi^*}} (\phi^*)^2 \big) q(\phi^* \mid \mathbf{x}^*) \mathrm{d}\phi^* + e^{\mu_{\theta^*} + \frac{1}{2}\sigma^2_{\theta^*}} \\
&= \Big((\mu_{\tilde{f}^*} - \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} \mu_{\phi^*})^2 + \sigma^2_{\tilde{f}^*} - \frac{\sigma^2_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} \Big) e^{\mu_{\phi^*} + \frac{1}{2}\sigma^2_{\phi^*}} \\
&+ 2(\mu_{\tilde{f}^*} - \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} \mu_{\phi^*}) \frac{\sigma_{\tilde{f}^*\phi^*}}{\sigma^2_{\phi^*}} e^{\mu_{\phi^*} + \frac{1}{2}\sigma^2_{\phi^*}} (\mu_{\phi^*} + \sigma^2_{\phi^*}) \\
&+ \frac{\sigma^2_{\tilde{f}^*\phi^*}}{\sigma^4_{\phi^*}} e^{\mu_{\phi^*} + \frac{1}{2}\sigma^2_{\phi^*}} (\mu_{\phi^*} + \sigma^2_{\phi^*}) + e^{\mu_{\theta^*} + \frac{1}{2}\sigma^2_{\theta^*}},
\end{aligned}
$$

when

$$
q(\theta^* \mid \mathbf{x}^*, D) = \mathrm{N}(\mu_{\theta^*}, \sigma^2_{\theta^*}), \quad q(\tilde{f}^*, \phi^* \mid \mathbf{x}^*, D) = \mathrm{N}\Big(\begin{bmatrix} \mu_{\tilde{f}^*} \\ \mu_{\phi^*} \end{bmatrix}, \begin{bmatrix} \sigma^2_{\tilde{f}^*} & \sigma_{\tilde{f}^*\phi^*} \\ \sigma_{\tilde{f}^*\phi^*} & \sigma^2_{\phi^*} \end{bmatrix} \Big).
$$