Improving posterior marginal approximations in latent Gaussian models

Botond Cseke **Tom Heskes**
Radboud University Nijmegen, Institute for Computing and Information Sciences
Nijmegen, The Netherlands
{b.cseke,t.heskes}@science.ru.nl

Abstract

We consider the problem of correcting the posterior marginal approximations computed by expectation propagation and Laplace approximation in latent Gaussian models and propose correction methods that are similar in spirit to the Laplace approximation of Tierney and Kadane (1986). We show that in the case of sparse Gaussian models, the computational complexity of expectation propagation can be made comparable to that of the Laplace approximation by using a parallel updating scheme. In some cases, expectation propagation gives excellent estimates, where the Laplace approximation fails. Inspired by bounds on the marginal corrections, we arrive at factorized approximations, which can be applied on top of both expectation propagation and Laplace. These give nearly indistinguishable results from the non-factorized approximations in a fraction of the time.

1 Introduction

Following Rue et al. (2009), we consider the problem of computing marginal probabilities over single variables in (sparse) latent Gaussian models. Probabilistic models with latent Gaussian variables are of interest in many areas of statistics, such as spatial data analysis (Rue and Held, 2005), and machine learning, such as Gaussian process models (e.g. Kuss and Rasmussen, 2005). The general setting considered in Rue et al. (2009) as well as in this paper is as follows. The prior distribution over the latent variables is a Gaussian random field with a sparse precision (inverse covariance) matrix and the likelihood factorizes into a product of terms depending on just a single latent variable. Both the prior and the likelihood may depend on a small set of hyper-parameters (say at most 6 in total). We are interested in the posterior marginal probabilities over single variables given all observations.

Rue et al. (2009) propose an integrated nested Laplace approximation to approximate these posterior marginal distributions. Their procedure consists of three steps. 1) Approximate the posterior of the hyper-parameters given the data and use this to determine a grid of hyper-parameter values. 2) Approximate the posterior marginal distributions given the data and the hyper-parameters values on the grid. 3) Numerically integrate the product of the two approximations to obtain the posterior marginals of interest. The crucial contribution is the improved marginal posterior approximation in step 2), based on the approach of Tierney and Kadane (1986), that goes beyond the Gaussian approximation and takes into account higher order characteristics of (all) likelihood terms. Comparing their approach with Monte Carlo sampling techniques on several high-dimensional models, they show that their procedure is remarkably fast and accurate.

The main objective of the current paper is to see whether we can improve upon the approach of Rue et al. (2009). Expectation propagation, a method for approximate inference developed and studied mainly in the machine learning community, is then an obvious candidate. It is well-known to yield approximations that are more accurate than the Laplace approximation (e.g. Minka, 2001; Kuss and Rasmussen, 2005). Furthermore, expectation propagation can still be applied in cases where the Laplace approximation is doomed to fail, e.g., when the log-posterior is not twice-differentiable (Seeger, 2008). The typical price to be paid is that of higher computational complexity. However, we will see that, using a parallel instead of a sequential updating scheme, expectation propa-

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

gation is at most a (relatively small) constant factor slower than the Laplace approximation in applications on sparse Gaussian models with many latent variables. Moreover, along the way we will arrive at further approximations (both for expectation propagation and the Laplace approximation) that yield an order of magnitude speed-up, with hardly any degradation of performance.

Section 1.1 specifies the model and introduces notation, Section 2 introduces and compares several methods for correcting marginals given a fixed setting of the hyper-parameters, Section 3 discusses the computational complexity of these methods when applied to sparse models, and Section 4 treats integration over hyper-parameters.

1.1 Sparse latent Gaussian models

In this section we introduce notation and define the models under consideration. Let $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$ be the conditional probability of the observations $\boldsymbol{y} = (y_1,\ldots,y_n)^T$ given the latent variables $\boldsymbol{x} = (x_1,\ldots,x_n)^T$ and the hyper-parameters $\boldsymbol{\theta}$. We assume that this likelihood factorizes over the latent variables:

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|x_i,\boldsymbol{\theta}) \ .$$

The prior $p(\boldsymbol{x}|\boldsymbol{\theta})$ over the latent variables is Gaussian, e.g., a Gaussian process or a so-called thin plate spline mimicking prior on a two-dimensional grid (Rue et al., 2009). We call such a model "sparse", when the precision (inverse covariance) matrix of the Gaussian prior is sparse. Furthermore, we assume that the number of hyper-parameters $\boldsymbol{\theta}$ is relatively small, say at most 6. We will omit $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$'s and $p(\boldsymbol{x}|\boldsymbol{\theta})$'s dependence on $\boldsymbol{\theta}$ whenever it is not relevant, use $p_0(\boldsymbol{x})$ as an alias of the prior $p(\boldsymbol{x}|\boldsymbol{\theta})$, and $q(\boldsymbol{x})$ for an approximating Gaussian distribution.

2 Posterior marginals conditioned upon the hyper-parameters

2.1 Global approximations

In this section we will focus on approximating posterior marginal distributions given a fixed setting of the hyper-parameters $\boldsymbol{\theta}$, which is omitted from the notation. That is, our goal is to approximate

$$p(x_i|\boldsymbol{y}) = \frac{1}{Z} t_i(x_i) \int d\boldsymbol{x}_{\setminus i}\, p_0(\boldsymbol{x}) \prod_{j\neq i} t_j(x_j) \,, \qquad (1)$$

where we used shorthand notation $t_i(x_i) \equiv p(y_i|x_i)$ and with normalization constant

$$Z = \int d\boldsymbol{x}\, p_0(\boldsymbol{x}) \prod_i t_i(x_i) \,, \qquad (2)$$

which in fact corresponds to the "evidence" $p(\boldsymbol{y}|\boldsymbol{\theta})$ that we need in order to compute the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. In the following we will describe several approximation procedures. Discussion of the corresponding computational complexities is postponed until Section 3.

As a first step, we construct a global Gaussian approximation $q(\boldsymbol{x})$ of $p(\boldsymbol{x})$, e.g., through expectation propagation (EP) or using Laplace's method. The approximation obtained through EP is of the form

$$q(\boldsymbol{x}) = \frac{1}{Z_q} p_0(\boldsymbol{x}) \prod_i \tilde{t}_i(x_i) \,, \qquad (3)$$

where $\tilde{t}(x_i)$ are so-called Gaussian term proxies and where Z_q ensures proper normalization. A Gaussian term proxy has the form of a Gaussian, but need not be normalized nor normalizable, i.e., may have a negative precision. Expectation propagation iteratively improves the term proxies one by one[1]. When updating the ith term proxy given all other term proxies, the new term proxy $\tilde{t}_i(x_i)$ is chosen such that

$$\int dx_i \, \{1, x_i, x_i^2\} q^{\setminus i}(x_i) \tilde{t}_i(x_i) = \int dx_i \, \{1, x_i, x_i^2\} q^{\setminus i}(x_i) t_i(x_i) \,, \qquad (4)$$

with the "cavity" distribution, the Gaussian approximation with the ith term proxy left out,

$$q^{\setminus i}(\boldsymbol{x}) \propto p_0(\boldsymbol{x}) \prod_{j\neq i} \tilde{t}_j(x_j) \,.$$

That is, we choose the new term proxy $\tilde{t}_i(x_i)$ such that the moments (up to second order) of "cavity times term proxy" equal those of "cavity times actual term". The solution of this "moment matching" operation is typically found through numerical integration. We refer to (Minka, 2005; Kuss and Rasmussen, 2005; Seeger, 2008) for more information on (how to use) EP for approximate inference in Gaussian processes and other models.

The global Gaussian approximation based on Laplace's method is obtained by first finding the mode $\boldsymbol{m} = \operatorname{argmax}_{\boldsymbol{x}} \log p(\boldsymbol{x},\boldsymbol{y})$, and then setting the covariance matrix to the negative inverse of the Hessian, $\boldsymbol{H}(\boldsymbol{x}) = \frac{\partial^2}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} \log p(\boldsymbol{x},\boldsymbol{y})$, evaluated at $\boldsymbol{m}$.

[1] Below we will describe a parallel updating scheme which, for sparse models, is a lot faster than the standard sequential scheme.

It is easy to see that this Hessian amounts to the (sparse) precision matrix from the prior $p_0(\boldsymbol{x})$ plus diagonal terms corresponding to second derivatives of the $\log t_i(x_i)$ terms. Consequently, also the Gaussian approximation resulting from Laplace's method can be written in the form (3) and, if desired, the corresponding term proxies can be used for initialization of the EP algorithm. The marginal $q(x_i)$ of the global Gaussian approximation (3) can be considered our lowest order approximation of the posterior marginal distribution of interest. We will write $\tilde{p}_i^{\text{EP}}(x_i)$ for the Gaussian marginal following from the EP approximation and $\tilde{p}_i^{\text{LA}}(x_i)$ for the marginal following from Laplace's method. In the following, we will discuss how to improve upon these global approximations.

2.2 Marginal corrections

Given a global Gaussian approximation $q(\boldsymbol{x})$ of the form (3) with corresponding term proxies, we can rewrite (1) as

$$\begin{aligned} p\left(x_i|\boldsymbol{y}\right) &= \frac{Z_q}{Z}\frac{t_i\left(x_i\right)}{\tilde{t}_i\left(x_i\right)}\int d\boldsymbol{x}_{\setminus i}\, q\left(\boldsymbol{x}\right)\prod_{j\neq i}\frac{t_j\left(x_j\right)}{\tilde{t}_j\left(x_j\right)} \qquad (5)\\ &= \frac{Z_q}{Z}\frac{t_i\left(x_i\right)}{\tilde{t}_i\left(x_i\right)}q\left(x_i\right)\int d\boldsymbol{x}_{\setminus i}\, q\left(\boldsymbol{x}_{\setminus i}|x_i\right)\prod_{j\neq i}\frac{t_j\left(x_j\right)}{\tilde{t}_j\left(x_j\right)}\\ &= \frac{Z_q}{Z}\epsilon_i(x_i)q(x_i)\int d\boldsymbol{x}_{\setminus i}\, q\left(\boldsymbol{x}_{\setminus i}|x_i\right)\prod_{j\neq i}\epsilon_j(x_j)\,, \end{aligned}$$

where we defined $\epsilon_i(x_i) = t_i(x_i)/\tilde{t}_i(x_i)$.

Equation (5), which is still exact, shows that there are two corrections to the Gaussian approximation $q(x_i)$: one direct, local correction through $\epsilon_i(x_i)$ and one more indirect correction through the (weighted integral over) $\epsilon_j(x_j)$'s for $j \neq i$. The direct, local correction comes without additional cost and suggests a second approximation,

$$p\left(x_i|\boldsymbol{y}\right) \approx \epsilon_i(x_i)q(x_i)\,,$$

which will be denoted $\tilde{p}_i^{\text{EP-L}}(x_i)$ and $\tilde{p}_i^{\text{LA-L}}(x_i)$ for the approximations following the global EP and Laplace approximation, respectively. The approximation $\tilde{p}_i^{\text{EP-L}}(x_i)$ is the marginal of EP's "tilted" distribution $q_i(\boldsymbol{x}) \propto t_i(x_i)q^{\setminus i}(\boldsymbol{x})$ (e.g. Minka, 2001; Opper et al., 2009).

To improve upon this approximation, we somehow have to get a handle on the indirect correction

$$c_i(x_i) \equiv \int d\boldsymbol{x}_{\setminus i}\, q\left(\boldsymbol{x}_{\setminus i}|x_i\right)\prod_{j\neq i}\epsilon_j(x_j)\,. \qquad (6)$$

The observation here is that, for each x_i, we are in fact back to the form (2): we have to estimate the normalization constant of a sparse latent Gaussian model, where $q\left(\boldsymbol{x}_{\setminus i}|x_i\right)$ now plays the role of a sparse $(n-1)$-dimensional Gaussian prior and the $\epsilon_j(x_j)$ are terms depending on a single variable. The idea is to choose a grid of x_i values, compute $c_i(x_i)$ for each value of x_i using our favorite method for computing normalization constants, and numerically interpolate between the resulting approximations. Running a complete procedure, be it EP or Laplace's method, for each x_i is often computationally too intensive and further approximations are needed to reduce the computational burden.

2.2.1 EP corrections

Let us write $\tilde{\epsilon}_j(x_j; x_i)$ for the term proxy of $\epsilon_j(x_j)$ in the context of approximating $c_i(x_i)$. A full run of EP for each x_i may be way too expensive, so instead we propose to make just one parallel step. Since the term proxies of the global EP approximation are tuned to make $\tilde{t}_j(x_j)$ close to $t_j(x_j)$, it makes sense to initialize $\tilde{\epsilon}_j(x_j; x_i)$ to 1. Following the same procedure as in (4), computing the new term proxy for term j then amounts to choosing $\tilde{\epsilon}_j(x_j; x_i)$ such that

$$\int dx_j\, \{1, x_j, x_j^2\} q(x_j|x_i)\tilde{\epsilon}_j(x_j; x_i) = \int dx_j\, \{1, x_j, x_j^2\} q(x_j|x_i)\epsilon_j(x_j)\,. \qquad (7)$$

Replacing the terms $\epsilon_j(x_j)$ in (6) by their term proxies $\tilde{\epsilon}_j(x_j; x_i)$ yields an estimate for $c_i(x_i)$. The corresponding approximation

$$p(x_i|\boldsymbol{y}) \approx \epsilon_i(x_i)q(x_i)\int d\boldsymbol{x}_{\setminus i}\, q\left(\boldsymbol{x}_{\setminus i}|x_i\right)\prod_{j\neq i}\tilde{\epsilon}_j(x_j; x_i) \qquad (8)$$

is referred to as $\tilde{p}_i^{\text{EP-1STEP}}(x_i)$.

2.2.2 Laplace corrections

In our setting, the approximation proposed by Rue et al. (2009) can be understood as follows. In principle, one could, following Tierney and Kadane (1986), run a Laplace approximation on

$$f(\boldsymbol{x}_{\setminus i}; x_i) \equiv q\left(\boldsymbol{x}_{\setminus i}|x_i\right)\prod_{j\neq i}\epsilon_j(x_j)\,.$$

To do this, one would need to compute, for each value of x_i, the mode of $f(\boldsymbol{x}_{\setminus i}; x_i)$ as well as (the determinant of minus) the Hessian of $\log f(\boldsymbol{x}_{\setminus i}; x_i)$, evaluated at this mode. We will refer to the corresponding approximation as $\tilde{p}_i^{\text{LA-TK}}(x_i)$. Because finding the optimum of $f(\boldsymbol{x}_{\setminus i}; x_i)$ is computationally rather expensive, Rue et al. (2009) propose to replace the mode of $f(\boldsymbol{x}_{\setminus i}; x_i)$ by the mode of $q(\boldsymbol{x}_{\setminus i}|x_i)$, i.e., the conditional mean of the Laplace approximation, and to evaluate the Hessian at this conditional mean. The corresponding approximation, which we will refer to as $\tilde{p}_i^{\text{LA-CM}}(x_i)$,

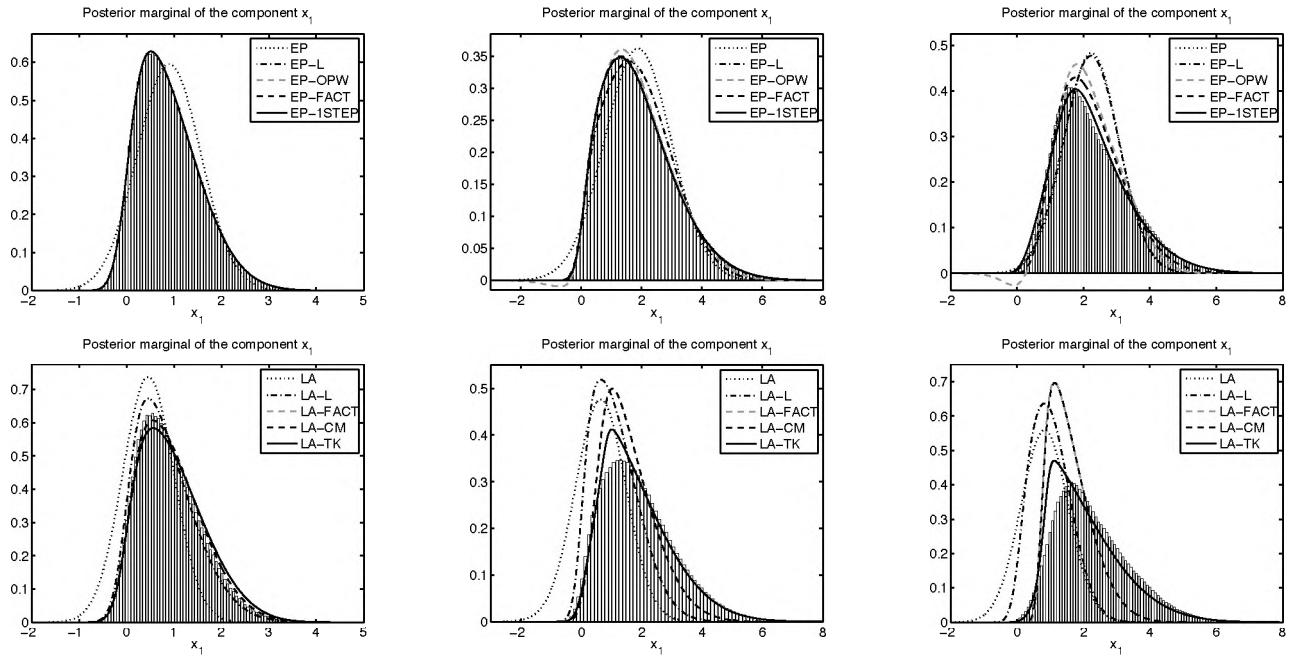

Figure 1: Various marginal corrections for a probit model with $t_i(x_i) = \Phi(4x_i)$ and identical variances and correlations in the prior p_0, using expectation propagation (top row) and Laplace-type approximations (bottom row). The panels show the corrections for a three-dimensional model with prior variances and correlations $(v, c) = (1, 0.25)$ (left), $(v, c) = (4, 0.9)$ (middle) and for a 32-dimensional model $(v, c) = (4, 0.95)$ (right).

is of the form (8), where now $\tilde{\epsilon}_j(x_j; x_i)$ follows from a second-order Taylor expansion of $\log \epsilon_j(x_j)$ around the mode (and thus mean) of $q(x_j|x_i)$.

In order to further reduce computational effort, Rue et al. (2009) suggest additional approximations that, because they can only be expected to reduce the accuracy of the final approximation, will not be considered in our experiments in Sections 2.3 and 4.

2.2.3 Bounds and factorized approximations

As we will discuss below, the computational bottleneck in the above procedures for approximating the correction $c_i(x_i)$ is not computing appropriate approximations of the terms $\epsilon_j(x_j)$, either through EP or using Laplace's method, but instead computing the normalization of the resulting Gaussian form which boils down to the computation of the determinant of a sparse matrix. Here we propose a simplification, which we motivate through its connection to bounds on the marginal correction $c_i(x_i)$.

Using Jensen's inequality, we obtain the lower bound

$$c_i(x_i) \geq \exp\left[\sum_{j\neq i} \int dx_j \; q(x_j|x_i) \log \epsilon_j(x_j)\right] \equiv c_i^{\mathrm{l}}(x_i) \; .$$

Following Minka (2005), we can also get an upper bound:

$$c_i(x_i) \leq \prod_{j\neq i} \left[\int dx_j \; q(x_j|x_i)\epsilon_j(x_j)^{n-1}\right]^{1/(n-1)} \equiv c_i^{\mathrm{u}}(x_i).$$

This upper bound will in many cases be useless because the integral does not exist. The lower bound, which corresponds to a mean-field-type approximation, does not have this problem, but may still be somewhat conservative. We therefore propose the general family of approximations

$$c_i^{(\alpha)}(x_i) = \prod_{j\neq i} \left[\int dx_j \; q(x_j|x_i)\epsilon_j(x_j)^{\alpha}\right]^{1/\alpha} . \qquad (9)$$

It is easy to show that

$$c_i^{\mathrm{l}}(x_i) \leq c_i^{(\alpha)}(x_i) \leq c_i^{\mathrm{u}}(x_i) \quad \forall\; 0 \leq \alpha \leq n-1 \; ,$$

where $\alpha = 0$ is interpreted as the limit $\alpha \to 0$. The choice $\alpha = 1$ makes the most sense: it gives exact results for $n = 2$ as well as when all x_j's (indeed) happen to be conditionally independent given x_i. We refer to the corresponding approximation as $\tilde{p}_i^{\text{EP-FACT}}(x_i)$.

Using (7), it is easy to see that $\tilde{p}_i^{\text{EP-FACT}}(x_i)$ corresponds to $\tilde{p}_i^{\text{EP-1STEP}}(x_i)$ if in (8) we would replace $q(\boldsymbol{x}_{\setminus i}|x_i)$ by the factorization $\prod_{j\neq i} q(x_j|x_i)$, i.e., as if the variables x_j in the global Gaussian approximation are conditionally independent given x_i. The same replacement in the Laplace approximation yields the approximation referred to as $\tilde{p}_i^{\text{LA-FACT}}(x_i)$.

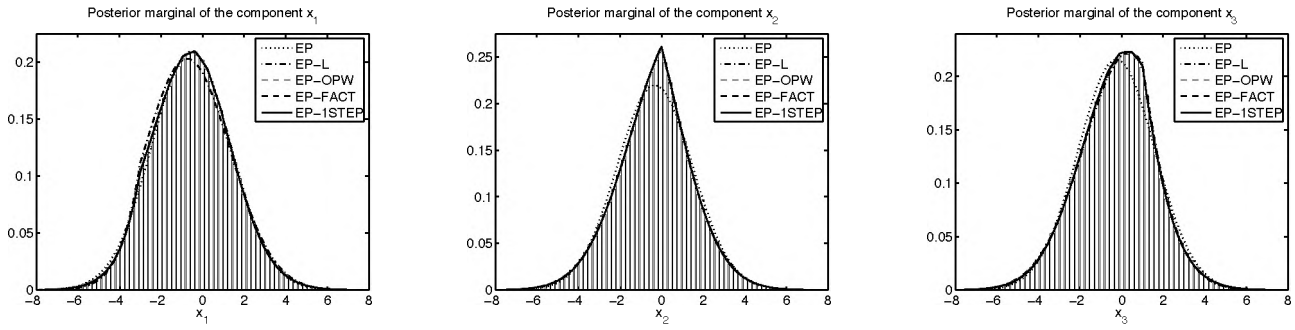

Figure 2: Marginal corrections for a three-dimensional model with $p\left(y_i|x_i,\lambda\right) = \lambda e^{-\lambda|y_i-x_i|}/2$ ($\lambda = 0.25, [y_1, y_2, y_3] = [-3, 0, 1]$) and identical variances and correlations in p_0, corresponding to a prior variance and correlation $(v, c) = (9, 0.9)$.

2.2.4 Taylor expansions

To make the connection to the earlier work in (Opper et al., 2009), we expand the exact $c_i(x_i)$ of (6) in $\epsilon_j(x_j) - 1$ for all $j \neq i$. Keeping only lowest order terms, we obtain

$$c_i(x_i) \approx 1 + \sum_{j\neq i} \int dx_j q(x_j|x_i) \left[\epsilon_j(x_j) - 1\right] \equiv c_i^{\text{TAYLOR}}(x_i),$$

which coincides with the Taylor expansion of $c_i^{(\alpha)}(x_i)$ of (9) for any α. An obvious approximation would be

$$p_i(x_i) \approx q_i(x_i)\epsilon_i(x_i)c_i^{\text{TAYLOR}}(x_i)\,. \tag{10}$$

The approximation proposed in (Opper et al., 2009) goes one step further by Taylor expanding not only $\epsilon_j(x_j)$ for $j \neq i$, but also $\epsilon_i(x_i)$ up to the same order, which boils down to

$$p_i(x_i) \approx q(x_i)\left[\epsilon_i(x_i) + c_i^{\text{TAYLOR}}(x_i) - 1\right] \equiv \tilde{p}_i^{\text{EP-OPW}}(x_i)\,. \tag{11}$$

Computing $\tilde{p}_i^{\text{EP-OPW}}(x_i)$ is as expensive as computing $\tilde{p}^{\text{EP-FACT}}(x_i)$. Where $\tilde{p}_i^{\text{EP-OPW}}(x_i)$ can yield negative probabilities, $\tilde{p}^{\text{EP-FACT}}(x_i)$ is nonnegative by construction. Furthermore, $\tilde{p}^{\text{EP-FACT}}(x_i)$ appears to be more accurate (see below), if only because it prevents the unnecessary step from (10) to (11).

2.3 Comparisons on toy models

To illustrate the correction methods, we take a probit model with $t\left(x_i\right) = \Phi\left(4x_i\right)$, with Φ the Gaussian cumulative density function, and a zero-mean prior p_0 with covariance matrix $\boldsymbol{Q}^{-1} = v[(1-c)\boldsymbol{I} + c\mathbf{1}\mathbf{1}^T]$. The left and middle panels in Figure 1 show the marginal corrections of the first component for a three-dimensional model with $(v, c) = (1, 0.25)$ and $(v, c) = (4, 0.9)$, respectively. The bars, in this and all other figures, correspond to a large number of Monte Carlo samples, either obtained through Gibbs or Metropolis sampling, and are supposed to represent the gold standard. The local correction EP-L yields sufficiently accurate approximations when the correlations are weak (left-top), but is clearly insufficient when they are strong (middle-top). The corrections EP-1STEP and EP-FACT yield accurate estimates and are almost indistinguishable even for strong prior correlations. Only when we increase the number of dimensions (here from 3 to 32) and with strong correlations $(v, c) = (4, 0.95)$, we can see small differences (right-top). As we can see on Figure 1, EP-OPW does slightly worse than EP-FACT and can indeed go negative.

It is known that the Laplace-type approximations does not perform well on this model (e.g. Kuss and Rasmussen, 2005). The approximations tend to be acceptable for weak correlations (bottom-left), with LA-CM and LA-FACT clearly outperforming LA and LA-L, but are far off when the correlations are strong (bottom-middle). The Laplace corrections suffer from essentially the same problems as the global Gaussian approximation based on Laplace's method: the mode and the inverse Hessian badly represent the mean and the covariance and fail to sufficiently improve it.

Expectation propagation can still be applied when the Laplace approximation is doomed to fail. An example is Bayesian linear regression with a double-exponential prior (Seeger, 2008). Direct application of the Laplace approximation makes no sense, because there is no local curvature information available that properly represents the behavior of the function $|x|$. Figure 2 describes a toy model with the same characteristics. It can be seen that the lowest order (Gaussian) EP approximation gets the mass right, but not the shape. Local corrections already help a lot, and both factorized and one-step EP corrections are practically indistinguishable from the sampling results.

We compared the various methods on several other toy models (not shown due to lack of space), leading to similar observations. It is relatively easy to come

up with models on which (all) Laplace-type approximations fail and expectation propagation, in particular EP-1STEP and EP-FACT are still fine. It is a lot harder to find cases where the factorized approximations EP-FACT and LA-FACT give quite different results than the non-factorized and computationally more expensive EP-1STEP and LA-CM: for this we really need to go to high dimensions and strong correlations.

3 Inference in sparse models

In this section we review the computational complexities of the Laplace approximation and expectation propagation when applied to sparse Gaussian models, i.e., models for which the n-dimensional precision matrix $\boldsymbol{Q}$ of the Gaussian prior is sparse. Is expectation propagation indeed orders of magnitude slower as suggested in (Rue et al., 2009)?

3.1 Global approximations

The computational complexity for the Gaussian approximation based on both Laplace's method and expectation propagation is dominated by several operations. 1) Computing the *Cholesky factor*, say $\tilde{\boldsymbol{L}}$ of a matrix $\tilde{\boldsymbol{Q}}$, e.g., corresponding to the posterior approximation $\tilde{p}^{\text{EP}}$ or $\tilde{p}^{\text{LA}}$, with the same sparsity structure as the prior precision matrix $\boldsymbol{Q}$. The computational complexity, denoted c_{chol}, in the worst case scales with n^3, but typically with $\texttt{nnzeros}(\boldsymbol{Q})^2/n$, with $\texttt{nnzeros}(\boldsymbol{Q})$ the number of non-zeros in the precision matrix $\boldsymbol{Q}$. 2) Computing the *diagonal elements of the inverse* of $\tilde{\boldsymbol{Q}}$. For sparse matrices, these can be computed efficiently by solving the Takahashi equations (e.g. Erisman and Tinney, 1975; Rue et al., 2009), which take the Cholesky factor $\tilde{\boldsymbol{L}}$ as input. The computational complexity, denoted c_{taka}, in the worst case scales with n^3, but typically scales with $\texttt{nnzeros}(\boldsymbol{L})^2/n$. In practice, we experienced that it is significantly more expensive than the Cholesky factorization, possibly due to our implementation[2]. 3) Solving a *triangular system* of the form $\tilde{\boldsymbol{L}}\boldsymbol{a} = \boldsymbol{b}$, with corresponding computational complexity $c_{\text{tria}} \propto \texttt{nnzeros}(\boldsymbol{L})$.

To keep the number of non-zeros in the Cholesky factor to a minimum, we apply the approximate minimum degree reordering algorithm (Amestoy et al., 1996), which is claimed to have the best average performance (Ingram, 2006). Since the sparsity structure is fixed, this reordering algorithm has to be run only once, prior to running any other algorithm.

Laplace's method. The maximum a-posteriori solution required for Laplace's method can be found, for example, through a Newton method. Each Newton step requires one Cholesky factorization and the solution of two triangular systems. To arrive at the lowest-order marginals $\tilde{p}_i^{\text{LA}}$ for all nodes i, we need the diagonal elements of the covariance matrix, which can be computed by solving the Takahashi equations using the Cholesky factor from the last Newton step. So, in total, computing the lowest order marginals $\tilde{p}_i^{\text{LA}}$ for all nodes i using Laplace's method scales with $n_{\text{steps}}^{\text{Newton}} \times (c_{\text{chol}} + 2 \times c_{\text{tria}}) + c_{\text{taka}}$.

Expectation propagation. To update a term approximation $\tilde{t}_i(x_i)$ according to Equation (4), we compute $q^{\backslash i}(x_i) \propto q(x_i)/\tilde{t}_i(x_i)$ using the marginals $q(x_i)$ from the current global approximation $q(\boldsymbol{x})$ and re-estimate the normalization constant and the first two moments of $t_i(x_i)\, q^{\backslash i}(x_i)$. In standard practice, term approximations $\tilde{t}_i$ are updated sequentially and all marginal means and variances are recomputed using rank one updates after term each update. Instead, we adopt a parallel strategy, that is, we recompute marginal means and variances only after we have updated *all* term approximations $\tilde{t}_i$, $i = 1, \ldots, n$.

A parallel EP step boils down to: 1) compute the Cholesky factorization of the current precision matrix, 2) solve two triangular systems to compute the current posterior mean and solve the Takahashi equations to compute the diagonal elements of the covariance matrix, and 3) if necessary, use univariate Gauss-Hermite numerical quadrature with n_{quad} nodes to compute the quantities in Equation (4). This adds up to a computational complexity that scales with $n_{\text{steps}}^{\text{EP}} \times (c_{\text{chol}} + 2 \times c_{\text{tria}} + c_{\text{taka}} + n \times n_{\text{quad}})$. After convergence, EP yields the lowest order marginals $\tilde{p}_i^{\text{EP}}$ for all nodes i.

Summarizing, because of the parallel scheme, we use exactly the same computational tricks as with Laplace's method (Cholesky, Takahashi). Initializing the term approximations in EP from the Laplace solution and then doing a few EP steps to obtain better estimates of the probability mass, makes EP just a (small) constant factor slower than Laplace.

3.2 Marginal corrections

After running the global approximation, we are left with some Gaussian $q(\boldsymbol{x})$ with known precision matrix, a corresponding Cholesky factor and single-node marginals $q(x_i)$. We now consider the complexity of computing a corrected marginal through the various methods for a single node i, using n_{grid} grid points (see the summary in Table 1).

The local corrections $\tilde{p}_i^{\text{LA-L}}$ and $\tilde{p}_i^{\text{EP-L}}$ we get more or less for free. All other correction methods require

[2] We used the Matlab implementation of the sparse Cholesky factorization and a C implementation for solving the Takahashi equations.

steps \ methods	LA-CM	LA-FACT	EP-1STEP	EP-FACT
$q(x_j\|x_i)$	$c_{\text{tria}} + n \times n_{\text{grid}}$	$c_{\text{tria}} + n \times n_{\text{grid}}$	$c_{\text{tria}} + n \times n_{\text{grid}}$	$c_{\text{tria}} + n \times n_{\text{grid}}$
$\tilde{\epsilon}(x_j; x_i)$	$n \times n_{\text{grid}}$	$n \times n_{\text{grid}}$	$n \times n_{\text{grid}} \times n_{\text{quad}}$	$n \times n_{\text{grid}} \times n_{\text{quad}}$
Norm. or det.-s	$c_{\text{chol}} \times n_{\text{grid}}$	$n \times n_{\text{grid}}$	$c_{\text{chol}} \times n_{\text{grid}}$	$n \times n_{\text{grid}}$

Table 1: Computational complexities of the steps for computing an improved marginal approximation for a particular node i using the various methods. The frames highlight the complexities that typically dominate the computational time. c_{tria}, c_{chol}, and c_{taka} refer to solving a sparse triangular system, a Cholesky factorization, and Takahashi equations, respectively. n_{grid} refers to the number of grid points for x_i and n_{quad} to the number of quadrature points for x_j.

the computation of the conditional densities $q(x_j|x_i)$, which amounts to solving two sparse triangular systems and $(n-1) \times n_{\text{grid}}$ evaluations. To arrive at the term approximations $\tilde{\epsilon}(x_j; x_i)$, we need to compute second order derivatives for the Laplace approximation and numerical quadratures for EP, which is about n_{quad} times more expensive. For LA-FACT, EP-FACT, and EP-OPW, we then simply have to compute a product/sum of n normalization terms. For LA-TK, LA-CM and EP-1STEP, we need to compute the determinant of an $(n-1)$-dimensional sparse matrix, which costs a Cholesky factorization.

4 Inference of the hyper-parameters

Until now, we considered estimating single-node marginals conditioned upon the hyper-parameters. In this section, we consider the estimation of the posterior marginals that follow by integrating over the hyper-parameters. For this we need the posterior of the hyper-parameters given the observations, which is approximated by $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \tilde{p}(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$, where $\tilde{p}(\boldsymbol{y}|\boldsymbol{\theta})$ is the marginal likelihood approximation provided by Laplace's method or expectation propagation.

The basic idea is to compute the posterior mode of $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$ as well as the Hessian at this mode (using finite differences), select a set of uniformly spaced grid points along the scaled eigenvectors of this Hessian, and use these to perform numerical quadrature using the rectangle rule. We implemented a slight modification of the method used by Rue et al. (2009), which selects the grid points more efficiently (details to be given in an expanded report).

Example. As an example for a sparse Gaussian model we implemented the stochastic volatility model presented in (Rue et al., 2009). The data set consists of 945 samples of the daily difference of the pound-dollar exchange rate from October 1^{st}, 1981, to June 28^{th}, 1995. Similarly to Rue et al. (2009), we used the first 50 observations. The observations y_t given the latent variables η_t are taken to be distributed independently according to $p(y_t|\eta_t) = N(y_t|0, e^{\eta_t})$. The latent field η_t is assumed to be the sum $\eta_t = f_t + \mu$ of a first-order auto-regressive Gaussian process $p(f_t|f_{t-1}, \phi, \tau) = N(f_t|\phi f_{t-1}, 1/\tau)$, with $|\phi| < 1$, and an additional Gaussian bias term $p(\mu) = N(\mu|0, 1)$. The prior on the hyper-parameter τ is taken to be $p(\tau) = \Gamma(\tau|1, 10)$ and a Gaussian prior $\mathcal{N}(0, 3)$ is taken over $\phi' = \log((1+\phi)/(1-\phi))$.

The results are shown in Figure 3. The Laplace and EP approximation of the evidence are nearly indistinguishable (left), as are the posterior marginals of the hyper-parameters (middle-left). Here EP is about a factor 5 slower than Laplace. The posterior marginals of f_{50} and μ obtained using the more involved methods (right half, bottom row) are practically indistinguishable from each other and the gold (sampling) standard. This is not the case for the cheaper variants LA, EP, and LA-L, but *is* the case for EP-L (right half, top row): apparently to obtain excellent posterior marginals on this model, there is no need for (computationally expensive) higher-order corrections, but it suffices to compute a single global EP approximation per hyper-parameter setting and correct this for the (non-Gaussian) local term.

5 Discussion

There are many options for further improvement, in particular w.r.t. efficiency. The ideas behind the simplified Laplace approximation of (Rue et al., 2009), which aims to prevent the expensive computation of a determinant for each x_i, may well be applicable to expectation propagation. However, if this indeed dominates the computation times, the factorized approximation proposed in this paper may well be a better alternative. Incorporation of linear constraints on the latent variables, although not considered in this paper, should be relatively straightforward.

One of the main problems of expectation propagation is that it is not guaranteed to converge and may run into numerical problems. EP converged fine on the problems considered in this paper, but even when it does not, it can still be beneficial to start from the

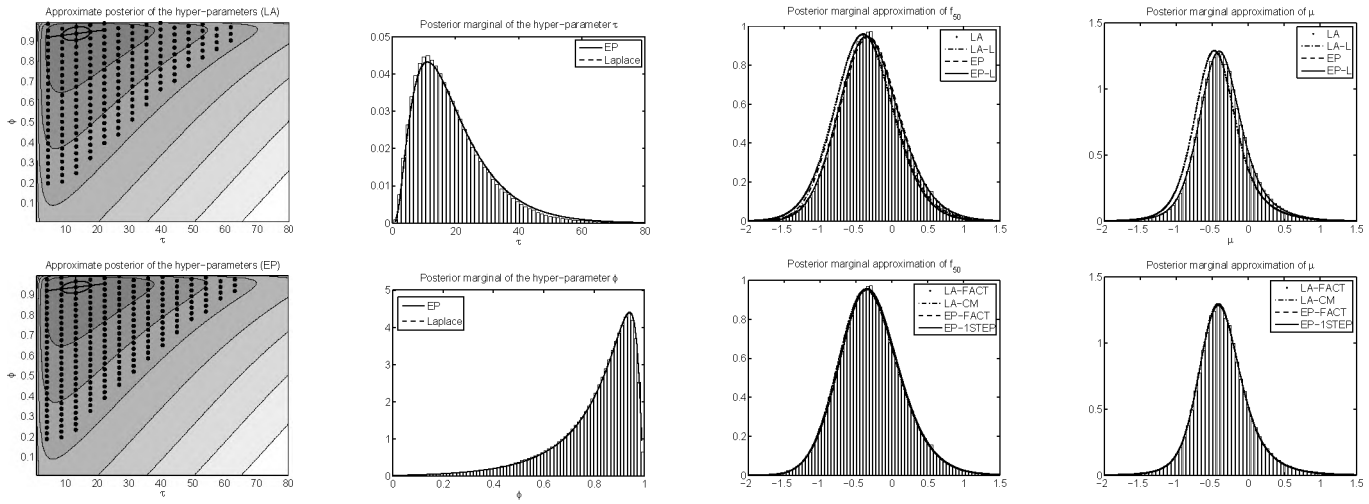

Figure 3: Plots of the posteriors for the stochastic volatility model in Section 4. The logarithm of posterior approximation of the hyper-parameters with EP and Laplace's method (left), their marginals (middle-left) and the posterior marginal approximations of f_{50} and μ (right half) when integrated over the corresponding approximations of the hyper-parameters' posterior. Dots show the hyper-parameters used for numerical integration; ellipses visualize the Hessian at the posterior mode.

Laplace solution and make just a few steps to get a better grip on the probability mass instead of relying on the mode and the curvature.

For models with weak correlations and smooth nonlinearities, any approximation method gives decent results. It may well be possible to come up with cases (strong correlations, hard nonlinearities), where any deterministic approximation method fails. Most interesting problems are somewhere in between, and for those we can hardly tell how advanced and computationally intensive approximation method we need. The heuristic suggested in (Rue et al., 2009), systematically increase the complexity and stop when you do not obtain further changes, appears risky. In particular when going from the factorized to the non-factorized approximations, it is often hard to see changes, but still both approximations can be pretty far off. It would be interesting to obtain a better theoretical understanding of the (asymptotic) approximation errors implied by the different approaches.

Acknowledgements

This research was supported by VICI grant 639.023.604 from the Netherlands Organization for Scientific Research (NWO). We would like to thank the anonymous reviewers and Håvard Rue for valuable comments on an earlier version.

References

P. R. Amestoy, T. A. Davis, and Iain S. D. An approximate minimum degree ordering algorithm. *SIAM J. Matrix Anal. Appl.*, 17(4):886–905, October 1996.

A. M. Erisman and W. F. Tinney. On computing certain elements of the inverse of a sparse matrix. *Commun. ACM*, 18(3):177–179, 1975. ISSN 0001-0782.

S. Ingram. Minimum degree reordering algorithms: A tutorial, 2006. URL `http://www.cs.ubc.ca/~sfingram/cs517\_final.pdf`.

M. Kuss and C. E. Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704, 2005. ISSN 1533-7928.

T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.

T. P. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd., Cambridge, UK, December 2005.

M. Opper, U. Paquet, and O. Winther. Improving on expectation propagation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1241–1248. MIT, Cambridge, MA, US, 2009.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, UK, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal Of The Royal Statistical Society Series B*, 71(2):319–392, 2009.

M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008. ISSN 1533-7928.

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