Approximate Marginals in Latent Gaussian Models

Botond Cseke Tom Heskes

B.Cseke@science.ru.nl T.Heskes@science.ru.nl

Institute for Computing and Information Sciences Faculty of Science, Radboud University Nijmegen Heyendaalseweg 135, 6525 AJ, The Netherlands

Editor: Manfred Opper

Abstract

We consider the problem of improving the Gaussian approximate posterior marginals computed by expectation propagation and the Laplace method in latent Gaussian models and propose 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 method by using a parallel updating scheme. In some cases, expectation propagation gives excellent estimates where the Laplace approximation fails. Inspired by bounds on the correct marginals, we arrive at factorized approximations, which can be applied on top of both expectation propagation and the Laplace method. The factorized approximations can give nearly indistinguishable results from the non-factorized approximations and their computational complexity scales linearly with the number of variables. We experienced that the expectation propagation based marginal approximations we introduce are typically more accurate than the methods of similar complexity proposed by Rue et al. (2009).

Keywords: approximate marginals, Gaussian Markov random fields, Laplace approximation, variational inference, expectation propagation

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 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. 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 (Minka, 2001), 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 method (e.g., Minka, 2001; Kuss and Rasmussen, 2005). Furthermore, expectation propagation can still be applied in cases where the Laplace method is out of the question, for example, 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 propagation is at most a small constant factor slower than the Laplace method 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 method) that yield an order of magnitude speed-up, with hardly any degradation in performance.

The paper is structured as follows. In Sections 1.1 and 2 we specify the model and briefly present the Laplace method and expectation propagation. In Section 3, we introduce and compare several methods for correcting marginals given a fixed setting of the hyper-parameters. In Section 4.6, we discuss the computational complexity of these methods when applied to sparse models. In Section 5, we introduce a method for numerical integration over hyper-parameters and finally in Section 6, we show that the proposed methods are competitive both in computational complexity and accuracy with the methods introduced in Rue et al. (2009).

In order to increase the readability of the paper we include a schematic figure (Figure 13) and an explanatory list (Section D of the Appendix) of the marginal approximation methods we introduce or refer to. In the following we define the model and give a short outline of how we proceed to approximate the marginal densities.

1.1 Latent Gaussian Models

In this section, we introduce notation and define the model under consideration. Let $p(y|x, \theta_l)$ be the conditional probability of the observations $y = (y_1, \dots, y_n)^T$ given the latent variables $x = (x_1, \dots, x_n)^T$ and the hyper-parameters θ_l . We assume that $p(y|x, \theta_l)$ factorizes as

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

The prior $p(x|\theta_p)$ over the latent variables is taken to be Gaussian with canonical parameters $h(\theta_p)$ and $Q(\theta_p)$, that is,

$$p(\boldsymbol{x}|\boldsymbol{\theta}_p) \propto \exp\left(\boldsymbol{x}^T \boldsymbol{h}(\boldsymbol{\theta}_p) - \frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta}_p) \boldsymbol{x}\right).$$

Examples for $p(x|\theta_p)$ include Gaussian process models, where $Q^{-1}(\theta_p)$ is the covariance matrix at the corresponding input and Gaussian Markov random fields, where the elements of $Q(\theta_p)$ are the interactions strengths $Q_{ij}(\theta_p)$ between the latent variables x_i and x_j . The prior $p(\theta_l, \theta_p)$ over the hyper-parameters is typically taken to be non-informative—uniform for location variables and

log-uniform for scale variables—and factorizes w.r.t. θ_l and θ_p . In order to simplify the notation, we use the proxy $\theta = (\theta_l, \theta_p)$ to denote the hyper-parameters of the model.

The joint distribution of the variables in the model we study is

$$p(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}) \propto \prod_{i=1}^{n} p(y_i | x_i, \boldsymbol{\theta}) \exp\left(\boldsymbol{x}^T \boldsymbol{h}(\boldsymbol{\theta}) - \frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q}(\boldsymbol{\theta}) \boldsymbol{x}\right) p(\boldsymbol{\theta}).$$

We take y fixed and we consider the problem of computing accurate approximations of the posterior marginal densities of the latent variables $p(x_i|y,\theta)$, given a fixed hyper-parameter value. Then we integrate these marginals over the approximations of the hyper-parameters posterior density $p(\theta|y)$. The exact quantities are given by the formulas

$$p(x_i|\boldsymbol{y},\boldsymbol{\theta}) = \frac{1}{p(\boldsymbol{y}|\boldsymbol{\theta})} p(y_i|x_i,\boldsymbol{\theta}) \int d\boldsymbol{x}_{\setminus i} p(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{j \neq i} p(y_j|x_j,\boldsymbol{\theta}),$$
(1)

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta}).$$
 (2)

We use the term *evidence* for $p(y|\theta) = \int dx p(y,x|\theta)$. In the following we omit $p(y|x,\theta)$'s and $p(x|\theta)$'s dependence on θ whenever it is not relevant and use $t_i(x_i)$ as an alias of $p(y_i|x_i,\theta)$ and $p_0(x)$ as an alias of $p(x|\theta)$. We use the notation $p(x) = Z_p^{-1} p_0(x) \prod_i t_i(x_i)$, with $Z_p(\theta) \equiv p(y|\theta)$. A Gaussian approximation of p(x) will be denoted by q and Z_q will denote its normalization constant.

1.2 An Outline of the Main Methods Presented in the Paper

In this paper, we will discuss a variety of methods for approximating marginals in latent Gaussian models. To assist the reader, we give an outline of these methods, leaving the technical details for later sections. We advise the reader to consult Figure 13 and Section D in the Appendix for a schematic figure and the corresponding explanatory list.

The posterior probability density p(x) is proportional to a (sparse) multivariate Gaussian distribution over all latent variables and a product of non-Gaussian terms $t_j(x_j)$, each of which depends on just a single latent variable. The first step is to find a global, Gaussian approximation of this posterior. There is obviously no need to approximate the Gaussian prior part, but we then do have to approximate the non-Gaussian terms $t_j(x_j)$ by Gaussian forms $\tilde{t}_j(x_j)$, thus constructing an approximation $q(x) \propto p_0(x) \prod_j \tilde{t}_j(x_j)$. Here we consider two choices, which we refer to as the Laplace method and expectation propagation.

The Laplace method (LM) finds the maximum of the (log) posterior and then makes a second order Taylor approximation around this maximum. It is easy to see that the Hessian, the matrix with second derivatives in this second order Taylor approximation, consists of the (sparse) precision matrix resulting from the Gaussian prior plus a diagonal term consisting of second derivatives of the logarithm of each of the terms $t_j(x_j)$. Hence, the approximation resulting from the Laplace method can always be written as the original prior $p_0(x)$ times a product of so-called term approximations $\tilde{t}_j(x_j)$, each of which has a Gaussian form (not necessarily normalizable) depending on just a single latent variable.

Expectation propagation (EP) aims to iteratively refine these term approximations $\tilde{t}_j(x_j)$. It works as follows. In the product of Gaussian prior times term approximations, we replace the term approximation that we aim to refine by the corresponding original non-Gaussian term. The resulting distribution $t_j(x_j)\tilde{t}_j(x_j)^{-1}q(\boldsymbol{x})$ is referred to as the tilted distribution: a Gaussian form $\tilde{t}_j(x_j)^{-1}q(\boldsymbol{x})$

times a non-Gaussian term $t_j(x_j)$ depending on a single latent variable. We then compute the zeroth, first, and second moments of this tilted distribution (e.g., through one-dimensional numerical integration) and determine the term approximation $\tilde{t}_j^{new}(x_j)$ which results in the same zeroth, first, and second moments. In the Gaussian approximation, we replace the old term approximation $\tilde{t}_j^{new}(x_j)$ by this new term approximation $\tilde{t}_j^{new}(x_j)$. In its original setting, expectation propagation refines term approximations $\tilde{t}_j(x_j)$ sequentially. In this particular setting of sparse models with many latent variables., a tremendous speed-up speed-up can be obtained by using a batch-mode version, that is, updating the term approximations in parallel.

Whichever procedure one prefers, Laplace or expectation propagation, this first step yields a global Gaussian approximation q(x) of the original non-Gaussian posterior. We can then write the exact non-Gaussian posterior as this Gaussian approximation q(x) times a product of correction terms, where each correction term is nothing but the original term $t_j(x_j)$ divided by its term approximation $\tilde{t}_j(x_j)$. Any further approximation is based on the assumption that these correction terms are close to 1 in average w.r.t. q, that is, that the Gaussian term approximation is indeed a sensible approximation of the original non-Gaussian term in the region where the main mass of q lies.

We are interested in accurate approximations of marginals $p(x_i)$ on a single variable, say x_i . For this, we have to integrate out all variables except x_i . Decomposing the global Gaussian approximation q(x) into the product of $q(x_i)$ and the conditional $q(x_{\setminus i}|x_i)$, we can take both $q(x_i)$ and the correction term depending on x_i outside of the integral over $x_{\setminus i}$. The remaining integrand is then the conditional Gaussian $q(x_{\setminus i}|x_i)$ times the product of all correction terms, except the one for x_i . The crucial observation here is that this integrand is of exactly the same form as the problem we started with: a (sparse) Gaussian prior (here the conditional $q(x_{\setminus i}|x_i)$) times a product of non-Gaussian terms (here the correction terms). In principle, we could again use the Laplace method or expectation propagation to approximate the integral. Doing this for the Laplace method yields the Laplace approximation of Tierney and Kadane (1986) (LA-TK) (Section 3.1). Doing the same in conjunction with expectation propagation leads to the approximation in Section 3.2.

However, both easily become very expensive, since we have to apply the Laplace method or run a full expectation propagation for each setting of x_i . Luckily, we now have an additional property that we can try to exploit: the non-Gaussian correction terms in the integrand have been constructed such that they are somehow close to 1.

The first, obvious approximation is to replace these correction terms within the integral by 1, leaving only the product of $q(x_i)$ and the correction term depending on x_i . We will refer to this type of approximation as a local approximation. In the case of expectation propagation it is exactly the corresponding marginal of the tilted distribution and we refer to it by EP-L (Section 3). The same approximation, but then in conjunction with the Laplace method is referred to as LM-L (Section 3).

The method proposed by Rue and co-workers can be viewed as a compromise between applying the expensive Laplace method (LA-TK) and the cheap local approximation (LM-L). Instead of finding the optimum of the integrand (conditional Gaussian times correction terms) and expanding around that, Rue et al. (2009) propose to expand the integrand around the optimum of conditional Gaussian only. Essentially, in the computation of the optimum of the integrand they hereby ignore the correction terms and simply set them to 1. Their method is referred to as LA-CM (see Section 4.1.2), where CM stands for conditional mean. It is straightforward and from the computational point of view relatively inexpensive to correct for the fact that the Taylor expansion is not done at the maximum of the integrand. The method which takes this into account is called LA-CM2 (Section 4.1.2).

In conjunction with EP we can use a similar argument. The term approximations inside the integral are optimized for the global Gaussian approximation, that is, when averaging over x_i . A full run of expectation propagation would give the term approximations that are optimal conditioned upon x_i , instead of marginalized over x_i . This difference is likely to be rather small and hence we expect that the main difference can be picked up by doing just one (parallel) iteration of expectation propagation. This approximation is referred to as EP-1STEP (Section 4.1.1). Iterating EP until convergence would lead to an approximation that will be referred to as EP-FULL (Section 4.1.1).

Another line of reasoning, followed by Opper et al. (2009), is to Taylor expand the correction terms around 1 (or their logarithm around 0). This is referred to as EP-OPW (Section 4.3). In their original work, they apply this Taylor expansion not only for the correction terms inside the integral, but also for the correction term depending on x_i outside of the integral, which is unnecessary in the current context. The interesting observation here is that, in a first-order Taylor expansion, the correlations within $q(x_{\setminus i}|x_i)$ become irrelevant and the integral over $x_{\setminus i}$ factorizes into a product of one-dimensional integrals. This (and also the existence of variational bounds on the marginals) suggests the approximation EP-FACT (Section 4.2), which corresponds to EP-1STEP, but then with $q(\mathbf{x}_{\setminus i}|x_i)$ replaced by its factorization $\prod_i q(x_i|x_i)$. The same replacement for LA-CM gives a method we refer to as LA-FACT (Section 4.2). Both EP-FACT and LA-FACT are an order of magnitude faster than their counterparts based on the non-factorized conditional distribution since they do not require computing the log-determinant of a high-dimensional (sparse) matrix for each setting of x_i . By a recursive application of the factorization principle one can obtain better approximations, which will be detailed in a future report. We use EP-FACTN to denote these approximations. In the following we expand the above mentioned ideas. We start with the presentation of the global approximation methods.

2. Global Gaussian Approximations

A close inspection of (1) and (2) shows that computing $p(x_i|y,\theta)$ leads to computing similar integrals as for $p(y|\theta)$. In this section, we review two approximation schemes that approximate such integrals: the Laplace method and expectation propagation (Minka, 2001). There are other approximation schemes, such as the variational approximation (e.g., Opper and Archambeau, 2009). The marginal approximation methods we propose for expectation propagation in Section 3 can be, under mild conditions, translated to the variational approximation in Opper and Archambeau (2009). For this reason, we will not discuss the details of this method.

2.1 The Laplace Method

The Laplace method approximates the evidence Z_p and, as a side product, it provides Gaussian approximation that is characterized by the local properties of the distribution at its mode $x^* = \operatorname{argmax}_x \log p(x)$. The mean parameter of the corresponding approximating Gaussian density is $m = x^*$ while the inverse of the covariance parameter V is the Hessian of $-\log p$ at x^* .

The idea behind the method is the following. Let $f = \log p$. Expanding f in second order at an arbitrary value \tilde{x} , we get

$$f(\boldsymbol{x}) = f(\tilde{\boldsymbol{x}}) + (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}}) + \frac{1}{2} (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \nabla_{\boldsymbol{x}}^2 f(\tilde{\boldsymbol{x}}) (\boldsymbol{x} - \tilde{\boldsymbol{x}}) + R_2 [f] (\boldsymbol{x}; \tilde{\boldsymbol{x}}),$$
(3)

where $R_2[f](x; \tilde{x})$ is the residual term of the expansion at \tilde{x} with $R_2[f](\tilde{x}; \tilde{x}) = 0$. By using the change of variables $s = x - \tilde{x}$, we have

$$\log \int d\boldsymbol{x} e^{f(\boldsymbol{x})} = f(\tilde{\boldsymbol{x}}) - \frac{1}{2} \nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}})^T \left[\nabla_{\boldsymbol{x}\boldsymbol{x}}^2 f(\tilde{\boldsymbol{x}}) \right]^{-1} \nabla_{\boldsymbol{x}} f(\tilde{\boldsymbol{x}})$$

$$- \frac{1}{2} \log \left| - \nabla_{\boldsymbol{x}\boldsymbol{x}}^2 f(\tilde{\boldsymbol{x}}) \right| + \log \mathbf{E}_{\boldsymbol{s}} \left[e^{R_2[f](\boldsymbol{s} + \tilde{\boldsymbol{x}}; \tilde{\boldsymbol{x}})} \right],$$

$$(4)$$

where $|\cdot|$ denotes the determinant and the expectation w.r.t. s is taken over a normal distribution with canonical parameters $\nabla_{x} f(\tilde{x})$ and $-\nabla_{xx}^{2} f(\tilde{x})$.

A closer look at (3) and (4) suggests that choosing $\tilde{x} = x^*$ and using the approximation $R_2[\log p](x; \tilde{x}) \approx 0$ yields an approximation of the log evidence

$$\log Z_p \approx \log p\left(\boldsymbol{x}^*\right) - \frac{1}{2}\log \left| -\nabla_{\boldsymbol{x}\boldsymbol{x}}^2 \log p\left(\boldsymbol{x}^*\right) \right|. \tag{5}$$

In the meantime, p can be approximated by the Gaussian

$$q\left(\boldsymbol{x}\right) = N\left(\boldsymbol{x}|\boldsymbol{x}^*, -\left[\nabla_{\boldsymbol{x}\boldsymbol{x}}^2\log p\left(\boldsymbol{x}^*\right)\right]^{-1}\right).$$

Note that any reasonably good approximation of $E_s\left[e^{R_2[f](s+\tilde{x};\tilde{x})}\right]$ can improve the accuracy of the approximation in (5).

The Laplace method requires the second order differentiability of $\log p$ at x^* , thus a sufficient condition for the applicability of this approximation scheme is the second order differentiability of $\log p$. The necessary condition is the second order differentiability at the mode x^* . A distribution p for which the method fails to give any meaningful information about the variances is, for example, when $p(y_j|x_j) = \lambda \exp(-\lambda|y_j - x_j|)/2$. In this case, the Hessian of $\log p$ at an arbitrary point \tilde{x} is either equal to the precision Q of the prior or it is undefined. Since the Laplace method captures the characteristics of the modal configuration, it often gives poor estimates of the normalization constant (e.g., Kuss and Rasmussen, 2005). The example in Section 4.1 shows how this behavior influences the approximation of the marginals in case of a two dimensional toy model. However, compared to other methods, the main advantage of the Laplace method is its speed. The optimization of $\log p$ w.r.t. x for computing $m = x^*$ requires only a few Newton steps.

2.2 Expectation Propagation

Expectation propagation (EP) approximates the integral for computing the evidence in the following way. Let us assume that q is a Gaussian approximation of p constrained to have the form $q(x) = Z_q^{-1} p_0(x) \prod_j \tilde{t}_j(x_j)$. Then the evidence can be approximated as

$$Z_{p} = \int dx \, p_{0}(x) \prod_{j} t_{j}(x_{j}),$$

$$= Z_{q} \int dx \, q(x) \prod_{j} \frac{t_{j}(x_{j})}{\tilde{t}_{j}(x_{j})},$$

$$\approx Z_{q} \prod_{j} \int dx_{j} \, q(x_{j}) \frac{t_{j}(x_{j})}{\tilde{t}_{j}(x_{j})}$$
(6)

and we are left with choosing the appropriate $\tilde{t}_j(x_j)$ s that yield both a good approximation of the evidence and of p(x). EP computes the terms $\tilde{t}_i(x_i)$ by iterating

$$\tilde{t}_j^{new}(x_j) \propto \frac{\text{Collapse}\left(t_j(x_j)\tilde{t}_j(x_j)^{-1}q(\boldsymbol{x})\right)}{q(\boldsymbol{x})}\tilde{t}_j(x_j), \text{ for all } j = 1,\dots,n,$$
(7)

where $\operatorname{Collapse}(r) = \operatorname{argmin}_{r' \in \mathcal{N}} \operatorname{D}[r||r']$ is the Kullback-Leibler (KL) projection of the distribution r into the family of Gaussian distributions \mathcal{N} . In other words, it is the Gaussian distribution that matches the first two moments of r. Using the properties of the KL divergence, one can check that when the terms t_j depend only on the variables x_j then $\operatorname{Collapse}\left(t_j(x_j)\tilde{t}_j(x_j)^{-1}q(x)\right)/q(x) = \operatorname{Collapse}\left(t_j(x_j)\tilde{t}_j(x_j)^{-1}q(x_j)\right)/q(x_j)$, therefore, the iteration in (7) is well defined. At any fixed point of this iteration, we have a set of $\tilde{t}_j(x_j)$ terms for which $\operatorname{Collapse}\left(t_j(x_j)\tilde{t}_j(x_j)^{-1}q(x)\right)=q(x)$ for any $j\in\{1,\ldots,n\}$. By defining the cavity distribution $q^{\setminus j}(x)\propto \tilde{t}_j(x_j)^{-1}q(x)$ and scaling the terms t_j , the above fixed point condition can be rewritten as

$$\int dx_j \left\{ 1, x_j, x_j^2 \right\} q^{\setminus j}(x_j) \tilde{t}_j(x_j) = \int dx_j \left\{ 1, x_j, x_j^2 \right\} q^{\setminus j}(x_j) t_j(x_j), \ j = 1, \dots, n,$$

and so, the approximation for Z_p has the form

$$Z_ppprox \int\! dm{x}\, p_0(m{x}) \prod_j ilde{t}_j(x_j).$$

Expectation propagation, can be viewed as a generalization of loopy belief propagation (e.g., Murphy et al., 1999) to probabilistic models with continuous variables and also as an iterative application of the assumed density filtering procedure (e.g., Csató and Opper, 2001). An equivalent algorithm for Gaussian process classification based on statistical physics methods was derived in Opper and Winther (2000). A close inspection of the parametric form of the iteration in Section C of the Appendix shows that the convexity of $\log \int dx N(x|m,V)t_j(x_j)$ w.r.t. m or the concavity of $\log t_j(x_j)$ (Seeger, 2008) is a sufficient condition for the terms \tilde{t}_j s to be normalizable and thus for the existence of q^{new} . However, this alone does not guarantee convergence. To our knowledge, the issue of EP's convergence in case of the models we study in this paper is still an open question. The iteration in (7) can also be derived by using variational free energies (e.g., Heskes et al., 2005; Minka, 2005). It can be relaxed such that the projections are taken on $t_j(x_j)^{\alpha} \tilde{t}_j(x_j)^{-\alpha} q(x)$, with $\alpha \in (0,1]$. The limit $\alpha \to 0$ corresponds to the variational approximation of Opper and Archambeau (2009).

In a personal correspondence, H. Rue emphasized that in many real world models, linear constraints of the form Ax = b have to be considered and expressed the concern that EP might not be suited to handle such constraints. Incorporating these constraints into EP would require to define updates for terms of the form $\delta_0(Ax - b)$. In the following we propose a way to deal with such terms. First we start out by deriving a sampling distributions for the Gaussian random variables x|Ax = b, where we assume that A is a $k \times n$ matrix with k < n. Let $x \sim \mathcal{N}(m, V)$ and $y = Ax - b + \epsilon$ with $\epsilon \sim \mathcal{N}(0, vI)$. Then the conditional density of x given y is a Gaussian with parameters $m + VA^T(AVA^T + vI)^{-1}(y - Am + b)$ and $V - VA^T(AVA^T + vI)^{-1}AV$. Setting y = 0 and taking the limit $v \to 0$ we find that

$$x|Ax = b \sim \mathcal{N}\left(m - VA^{T}(AVA^{T})^{-1}(Am - b), V - VA^{T}(AVA^{T})^{-1}AV\right). \tag{8}$$

As a consequence, we propose the following procedure: 1) first we perform the term updates for all "regular" terms, then we project the new moment parameters of q according to (8), 2) the value of the corresponding factor in (6) is $N(\mathbf{0}|\mathbf{A}\mathbf{m}-\mathbf{b},\mathbf{A}\mathbf{V}\mathbf{A}^T)$ and it corresponds to a Bayesian update in the limit $v \to 0$.

3. Approximation of the Posterior Marginals

The global approximations provide Gaussian approximations q of p and approximations of the evidence Z_p . The Gaussian approximation q can be used to compute Gaussian approximations of posterior marginals. In case of the Laplace method this only requires linear algebraic methods (computing the diagonal elements of the Hessian's inverse), while in the case of EP, the approximate marginals are a side product of the method itself. We refer to the corresponding Gaussian marginal approximations by LM-G (Laplace method) and EP-G (EP). Moreover, one can make use of the approximation method at hand in order to improve the Gaussian approximate marginals.

In case of the Laplace method, one can easily check that the residual term in (3) decomposes as $R_2 [\log p](x; \tilde{x}) = \sum_j R_2 [\log t_j](x_j; \tilde{x}_j)$, thus, when approximating the marginal of x_i it is sufficient to assume $R_2 [\log t_j](x_j; \tilde{x}_j) \approx 0$ only for $j \neq i$. This yields a locally improved approximation $q(x_i) \times \exp R_2 [\log t_i](x_i; x_i^*)$ to which we refer by LM-L.

As shown in Section 2.2, EP is built on exploiting the low-dimensionality of $t_i(x_i)$ and approximating the *tilted* marginals $t_i(x_i)q^{\setminus i}(x_i)$. These are known to be better approximations of the marginals $p(x_i)$ than $q(x_i)$ (e.g., Opper and Winther, 2000; Opper et al., 2009). We refer to this approximation by EP-L.

These observations show that there are ways to improve the marginals of the global approximation q by exploiting the properties of the methods. For the moment, however, we postpone this to Section 4 and first try to compute the marginals from scratch. This gives us some insight into where to look for further improvements.

The exact marginals can be computed as

$$p(x_i) = \frac{1}{Z_p} t_i(x_i) \int d\mathbf{x}_{\setminus i} \, p_0\left(\mathbf{x}_{\setminus i}, x_i\right) \prod_{j \neq i} t_j\left(x_j\right),\tag{9}$$

thus, as mentioned earlier, computing the marginal for a fixed x_i leads to computing the normalization constant of the distribution $p_0\left(\boldsymbol{x}_{\setminus i}|x_i\right)\prod_{j\neq i}t_j\left(x_j\right)$. Therefore, we can use our favorite method to approximate it. In the following, we present the details of these procedures for the Laplace method and EP.

3.1 Laplace Approximation

We use the same line of argument as in Section 2.1, but now we fix x_i and expand $\log p$ w.r.t. $\boldsymbol{x}_{\setminus i}$ at an arbitrary $\tilde{\boldsymbol{x}}_{\setminus i}$. The expression is identical to (3) with $\boldsymbol{x} = (x_i, \boldsymbol{x}_{\setminus i}^T)^T$ and $\tilde{\boldsymbol{x}} = (x_i, \tilde{\boldsymbol{x}}_{\setminus i}^T)^T$. Let $\boldsymbol{x}_{\setminus i}^*(x_i) = \operatorname{argmax}_{\boldsymbol{x}_{\setminus i}} \log p\left(x_i, \boldsymbol{x}_{\setminus i}\right)$ and let $\tilde{\boldsymbol{x}}_{\setminus i} = \boldsymbol{x}_{\setminus i}^*(x_i)$. Then the approximation of (4) simplifies to a form similar to (5), that is, the approximation of the marginal density, up to the constant $\log Z_p$, is given by

$$\log \int d\boldsymbol{x}_{\setminus i} p\left(\boldsymbol{x}\right) \approx \log p\left(x_{i}, \boldsymbol{x}_{\setminus i}^{*}(x_{i})\right) - \frac{1}{2}\log \left|-\nabla_{\boldsymbol{x}_{\setminus i}\boldsymbol{x}_{\setminus i}}^{2}\log p\left(x_{i}, \boldsymbol{x}_{\setminus i}^{*}(x_{i})\right)\right|. \tag{10}$$

This approximation is known in statistics as the Laplace approximation (Tierney and Kadane, 1986) and we will refer to it as $\tilde{p}_i^{\text{LA-TK}}(x_i)$.

The error of the approximation can be characterized in terms of the residual of the second order expansion. The residual decomposes as

$$R_2\left[\log p\right](\boldsymbol{x}; \tilde{\boldsymbol{x}}) = \sum_{j \neq i} R_2\left[\log t_j\right] \left(s_j + x_j^*(x_i); x_j^*(x_i)\right)$$

and the expectation (see Equation (4)) is taken w.r.t. $s \in \mathbb{R}^{(n-1)}$ having a normal density with mean $\mathbf{0}$ and inverse covariance $-\nabla^2_{\boldsymbol{x}_{\setminus i}\boldsymbol{x}_{\setminus i}}\log p(x_i,\boldsymbol{x}^*_{\setminus i}(x_i))$. This means that in principle we have exact estimates of the error and that any reasonable approximation of the integral can improve the quality of the approximation in (10).

3.2 Expectation Propagation

The integral in (9) can also be approximated using EP. As mentioned above EP typically provides better approximations of $\log Z_p$ than the Laplace method. For this reason, the marginals computed by approximating (9) using EP are expected to be more accurate. The procedure is as follows: (1) fix x_i and compute the canonical parameters of $p_0(\boldsymbol{x}_{\setminus i}|x_i)$ given by $\boldsymbol{h}_{\setminus i} - \boldsymbol{Q}_{\setminus i,i}x_i$ and $\boldsymbol{Q}_{\setminus i,\setminus i}$ and (2) use EP to approximate the integral in (9). Thus we approximate the integral by leaving out $p_0(x_i)$ and $t_i(x_i)$ and applying EP using the prior $p_0(\boldsymbol{x}_{\setminus i}|x_i)$ and the terms $t_i(x_i)$, $i \neq i$.

4. Approximation of the Posterior Marginals by Correcting the Global Approximations

As we have seen in the previous section, computing the marginal for a given fixed x_i value can be as expensive as the global procedure itself. On the other hand, however, there are ways to improve the marginals of the global approximation. In this section, we start from the "direct" approach and try to re-use the results of the global approximation to improve on the locally improved marginals LM-L and EP-L.

We start with the observation that for all the presented approximation methods, we can write the approximating distribution q as

$$q(\mathbf{x}) = \frac{1}{Z_q} p_0(\mathbf{x}) \prod_j \tilde{t}_j(x_j). \tag{11}$$

In case of the Laplace method, the canonical parameters of the Gaussian functions \tilde{t}_j are defined by the parameters of the Taylor expansion of $\log t_j$ at x_i^* , while in case of EP, they are the parameters corresponding to EP's fixed point.

In the following, we do not keep track of the normalization constants that are independent of x_i . In order to avoid overloading the notation and to express that a distribution is approximated as proportional to an expression on the right hand side of the \approx relation, we occasionally use Z as a proxy for unknown normalization constants. One can keep track of these constants, but in most cases, from the practical point of view, it is easier to perform a univariate numerical interpolation followed by numerical quadrature and (re)normalization.

4.1 Improving the Marginals of the Global Approximations

Given a global Gaussian approximation q(x) of the form (11) with corresponding term approximations $\tilde{t}_i(x_i)$, we can rewrite $p(x_i)$ as

$$p(x_{i}) = \frac{Z_{q}}{Z_{p}} \frac{t_{i}(x_{i})}{\tilde{t}_{i}(x_{i})} \int d\boldsymbol{x}_{\setminus i} q(\boldsymbol{x}) \prod_{j \neq i} \frac{t_{j}(x_{j})}{\tilde{t}_{j}(x_{j})},$$

$$= \frac{Z_{q}}{Z_{p}} \frac{t_{i}(x_{i})}{\tilde{t}_{i}(x_{i})} q(x_{i}) \int d\boldsymbol{x}_{\setminus i} q(\boldsymbol{x}_{\setminus i}|x_{i}) \prod_{j \neq i} \frac{t_{j}(x_{j})}{\tilde{t}_{j}(x_{j})},$$

$$= \frac{Z_{q}}{Z_{p}} \varepsilon_{i}(x_{i}) q(x_{i}) \int d\boldsymbol{x}_{\setminus i} q(\boldsymbol{x}_{\setminus i}|x_{i}) \prod_{j \neq i} \varepsilon_{j}(x_{j}),$$

$$= \frac{Z_{q}}{Z_{p}} \varepsilon_{i}(x_{i}) q(x_{i}) \int d\boldsymbol{x}_{\setminus i} q(\boldsymbol{x}_{\setminus i}|x_{i}) \prod_{j \neq i} \varepsilon_{j}(x_{j}),$$

$$(12)$$

where we define $\varepsilon_i(x_i) \equiv t_i(x_i)/\tilde{t}_i(x_i)$. In case of EP, the term approximations $\tilde{t}_i(x_i)$ are chosen to be close to the terms $t_i(x_i)$ in average w.r.t. $q(x_i)$. For this reason, we expect the $\varepsilon_i(x_i)$'s to be close to 1 in average w.r.t $q(x_i)$.

Equation (12) is still exact and it shows that there are two corrections to the Gaussian approximation $q(x_i)$: one direct, local correction through $\varepsilon_i(x_i)$ and one more indirect correction through the (weighted integral over) $\varepsilon_j(x_j)$ s for $j \neq i$. The direct, local correction comes without additional cost and suggests the above-mentioned (Section 3) local approximation

$$p(x_i) \approx \frac{1}{Z} \varepsilon_i(x_i) q(x_i)$$
.

We use the notations $\tilde{p}_i^{\text{EP-L}}(x_i)$ and $\tilde{p}_i^{\text{LM-L}}(x_i)$ for the approximations following the global Gaussian approximations by EP and Laplace method, respectively.

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

$$c_i(x_i) \equiv \int d\mathbf{x}_{\setminus i} \, q\left(\mathbf{x}_{\setminus i} | x_i\right) \prod_{j \neq i} \varepsilon_j(x_j) \,. \tag{13}$$

Again, for each x_i , we are in fact back to the form (9): we have to estimate the normalization constant of a latent Gaussian model, where $q\left(x_{\setminus i}|x_i\right)$ now plays the role of an (n-1)-dimensional Gaussian prior and the $\varepsilon_j(x_j)$ s are terms depending on a single variable. Running a complete procedure, be it EP or Laplace, for each x_i —as described in Sections 3.1 and 3.2—is often computationally too intensive and further approximations are needed to reduce the computational burden.

4.1.1 IMPROVING THE MARGINALS RESULTING FROM EP

Let us write $\tilde{\epsilon}_j(x_j;x_i)$ for the term approximation 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 too expensive, so instead we propose to perform just one simultaneous EP step for all $j \neq i$. Since the term approximations of the global EP approximation are tuned to make $\tilde{t}_j(x_j)$ close to $t_j(x_j)$ w.r.t. $q(x_i)$, it is plausible to initialize $\tilde{\epsilon}_j(x_j;x_i)$ to 1. Following EP, computing the new term approximation for term j then amounts to choosing $\tilde{\epsilon}_j(x_i;x_i)$ such that

$$\int dx_j \{1, x_j, x_j^2\} q(x_j | x_i) \tilde{\varepsilon}_j(x_j; x_i) = \int dx_j \{1, x_j, x_j^2\} q(x_j | x_i) \varepsilon_j(x_j), \tag{14}$$

that is, we get $\tilde{\epsilon}_j(x_j;x_i)$ by *collapsing* $\epsilon_j(x_j;x_i)q(x_j|x_i)$ into a Gaussian and dividing it by $q(x_j|x_i)$. As we have seen in Section 2.2, EP computes \tilde{t}_j such that

$$\int dx_j \{1, x_j, x_j^2\} q(x_j) = \int dx_j \{1, x_j, x_j^2\} q(x_j) \varepsilon_j(x_j),$$

thus, the difference here is made by the conditioning on x_i and $\tilde{\varepsilon}_j(x_j;x_i)$ can be viewed as an update $\tilde{t}_j(x_j;x_i)$ of $\tilde{t}_j(x_j)$ that accounts "locally" for this difference—up to second order. Replacing the terms $\varepsilon_j(x_j)$ in (13) by their term approximations $\tilde{\varepsilon}_j(x_j;x_i)$ yields an estimate for $c_i(x_i)$. The corresponding approximation

$$p(x_i) pprox rac{1}{Z} \mathbf{\hat{\epsilon}}_i(x_i) q(x_i) \int dm{x}_{\setminus i} \, q\left(m{x}_{\setminus i} | x_i
ight) \prod_{j
eq i} ilde{\mathbf{\hat{\epsilon}}}_j(x_j; x_i)$$

is referred to as $\tilde{p}_i^{\text{EP-1STEP}}(x_i)$. By performing further EP steps, one can refine the term approximations $\tilde{\epsilon}_j(x_j;x_i)$. Iterating the EP steps until convergence (as mentioned above) leads to a similar (costly) approximation as in Section 3.2. We refer to the resulting approximation as EP-FULL.

4.1.2 IMPROVING THE MARGINALS RESULTING FROM THE LAPLACE METHOD

According to the Laplace approximation presented in Section 3.1 one has to recompute the conditional mode $\boldsymbol{x}_{\backslash i}^*(x_i)$ for every choice of x_i . In order to lessen the computational burden, Rue et al. (2009) propose to re-use the global approximation by approximating the conditional mode with the conditional mean, that is, $\boldsymbol{x}_{\backslash i}^*(x_i) \approx \boldsymbol{m}_{\backslash i} + \boldsymbol{V}_{\backslash i,i}\boldsymbol{V}_{i,i}^{-1}(x_i - m_i)$, where $\boldsymbol{m} = \boldsymbol{x}^*(= \operatorname{argmax}_{\boldsymbol{x}} \log p(\boldsymbol{x}))$. This approximation often performs reasonably well when p is close to a Gaussian.

In our setting, the approximation proposed by Rue et al. (2009) can be understood as follows. The error terms ε_j can be identified with the residual terms, that is, $\log \varepsilon_i(x_i) = R_2 [\log t_i](x_i; m_i)$. In order to assess $c_i(x_i)$, one could, in principle, apply the Laplace method to

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

This would be identical to the direct method of Tierney and Kadane (1986) presented in Section 3.1. Using the conditional mean as an approximation of the conditional mode leads to ignoring the terms $\varepsilon_j(x_j)$ and using the mode of $q(\boldsymbol{x}_{\setminus i}|x_i)$. The corresponding approximation is of the form (4.1.1), where now $\tilde{\varepsilon}_j(x_j;x_i)$ follows from a second-order Taylor expansion of $\log \varepsilon_j(x_j)$ around the mode or mean of $q(x_j|x_i)$ instead of the mode of $f(\boldsymbol{x}_{\setminus i};x_i)$. We refer to this approximation as $\tilde{p}_i^{\text{LA-CM}}(x_i)$.

Taking a closer look at (4) and using our assumptions in Section 3.1, we can easily see that when we are not evaluating the normalization constant at the conditional mode, we can refine the approximation by adding $-\frac{1}{2}\nabla_{\boldsymbol{x}_{\backslash i}}f(\tilde{\boldsymbol{x}}_{\backslash i})[\nabla^2_{\boldsymbol{x}_{\backslash i},\boldsymbol{x}_{\backslash i}}f(\tilde{\boldsymbol{x}}_{\backslash i})]^{-1}\nabla_{\boldsymbol{x}_{\backslash i}}f(\tilde{\boldsymbol{x}}_{\backslash i})$, which is not identical to zero when the expansion in not made at the mode, that is, $\tilde{\boldsymbol{x}}_{\backslash i}\neq\boldsymbol{x}^*_{\backslash i}(x_i)$. As we will see in Section 4.7, this correction adds no significant computational burden to the method proposed in Rue et al. (2009). We refer to this approximation as $\tilde{p}_i^{\text{LA-CM2}}(x_i)$.

In order to further reduce computational effort, Rue et al. (2009) suggest additional approximations. Because they can only be expected to reduce the accuracy of the final approximation, we will not consider them in our experiments in Sections 4.5 and 6. Below we propose another EP-related approximation, motivated by theoretical bounds on the corrections $c_i(x_i)$.

4.2 Bounds and Factorized Approximations

The computational bottleneck in the above procedures for approximating the correction $c_i(x_i)$ is not computing appropriate approximations of the terms $\varepsilon_j(x_j)$, either through EP or Laplace, but instead computing the normalization of the resulting Gaussian form in (4.1.1), which leads to the

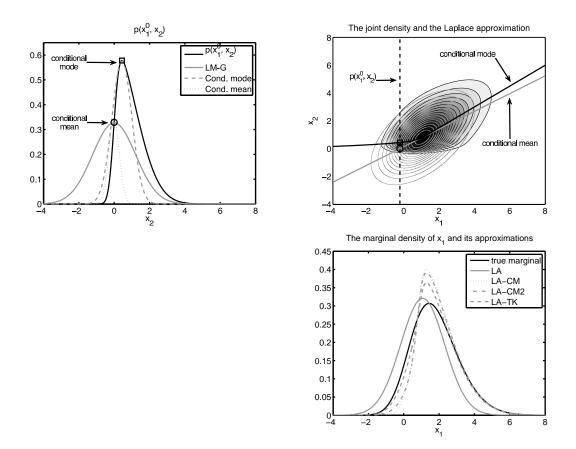


Figure 1: A two-dimensional example, illustrating how the Laplace approximation works and why it can fail. In the top-right panel, the black contour curves show the true distribution, the gray contour curves stand for the global Laplace approximation, and the black and gray curves show the conditional modes and the conditional means w.r.t. x_1 . The square and circle outline these quantities for a fixed x_1^0 . The dashed vertical line emphasizes the "slice" $p(x_1^0, x_2)$ at x_1^0 . The top-left panel shows $p(x_1^0, x_2)$ and the approximations for computing its area under the curve. The areas under the Gaussian curves corresponding to the conditional mode (square) and the conditional mean (circle) are the approximations of $p(x_1^0) = \int dx_2 \, p(x_1^0, x_2)$. The bottom-right panel shows the marginal of $p(x_1)$ and its approximations. The conditional mean can severely underestimate the mass for $x_1 = x_1^0$.

computation of a Gaussian normalization constant. 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 on (13)

$$c_i(x_i) \ge \exp\left[\sum_{j \ne i} \int dx_j \, q(x_j|x_i) \log \varepsilon_j(x_j)\right] \equiv c_i^{\text{lower}}(x_i).$$

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

$$c_i(x_i) \leq \prod_{i \neq i} \left[\int dx_j \, q(x_j|x_i) \varepsilon_j(x_j)^{n-1} \right]^{1/(n-1)} \equiv c_i^{\text{upper}}(x_i) \,.$$

This upper bound will in many cases be useless because the integral often 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

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

It is easy to show that

$$c_i^{\text{lower}}(x_i) \le c_i^{(\alpha)}(x_i) \le c_i^{\text{upper}}(x_i) \quad \forall \ 0 \le \alpha \le n-1$$

where $\alpha=0$ is interpreted as the limit $\alpha\to 0$. Furthermore, for any α we obtain exactly the same Taylor expansion in terms of $\varepsilon_j(x_j)-1$ (see Opper et al., 2009 and Section 4.3 below). The most sensible choice seems to be $\alpha=1$, because it gives exact results when n=2 as well as in the case when all x_j s are indeed conditionally independent given x_i . We refer to the corresponding approximation as $\tilde{p}_i^{\text{EP-FACT}}(x_i)$. Note that when EP converges, this approximation always exists, because $q(x_j|x_i)\varepsilon_j(x_j)$ is proportional to the conditional marginal of the so called *tilted* distributions $t_j(x_j)\tilde{t}_j(x_j)^{-1}q(x)$.

Using (14), it is easy to see that $\tilde{p}_i^{\text{EP-FACT}}(x_i)$ corresponds to $\tilde{p}_i^{\text{EP-ISTEP}}(x_i)$ if in (4.1.1) we would replace $q(\boldsymbol{x}_{\setminus i}|x_i)$ by the factorization $\prod_{j\neq i}q(x_j|x_i)$, that is, as if the variables x_j in the global Gaussian approximation are conditionally independent given x_i . A similar replacement in the Laplace approximation yields the approximation referred to as $\tilde{p}_i^{\text{LA-FACT}}(x_i)$. Here, we compute the univariate integrals with the Laplace method and using the approximation $x_j^*(x_i) \approx \mathrm{E}_q[x_j|x_i]$, with $q(\boldsymbol{x})$ being the global approximation resulting from the Laplace method.

The factorization principle can be applied to groups of variables x_I by factorizing $q(x_{\setminus I}|x_I)$. Another way to make use of the factorization is by applying it recursively. In this way, we can obtain higher order corrections of the approximate marginals and the evidence approximation. We will detail these methods in a future report.

An advantage of the bounding arguments is that we can extend the factorized approximation to cases when t_j depends on more variables, say, x_{I_j} , with $I_j \in \{1, ..., n\}$. In this case, the factorization is unfeasible since $\prod_j t_j(x_{I_j})$ may not factorize w.r.t. x_j . By using the bounding argument (Minka, 2005), we can still compute a "factorized" approximation

$$c_i^{(lpha)}(x_i) = \prod_{j \neq i} \left[\int \! dm{x}_{I_j} \, q(m{x}_{I_j}|x_i) \pmb{arepsilon}_j(m{x}_{I_j})^{lpha}
ight]^{1/lpha} \, .$$

An example illustrating this idea is the logistic regression model presented in Section 4.5.

4.3 Connection to the Taylor Expansion in Opper et al. (2009)

The line of argument in Opper et al. (2009) when applied to approximating the marginals can be explained in our notation as follows. By expanding $p(\mathbf{x}) = Z_q Z_p^{-1} q(\mathbf{x}) \prod_j \varepsilon_j(x_j)$ in first order w.r.t. all $\varepsilon_j(x_j) - 1$, they obtain a first order approximation of the exact p in terms of the global approximation q and the tilted distributions $t_j(x_j)q^{\setminus i}(\mathbf{x})$. The marginalization of this expansion yields the marginal approximation

$$\tilde{p}_i^{\text{EP-OPW}}(x_i) \equiv \frac{Z_q}{Z_p} q(x_i) \left[1 + \sum_i \int dx_j q(x_j|x_i) [\varepsilon_j(x_j) - 1] \right].$$

Since the goal of Opper et al. (2009) was to provide improved approximations of the posterior distribution p(x), and not only of its marginals, a natural adaptation of their approach would be to expand w.r.t. to all $j \neq i$ and not i itself. This leads to the approximation

$$p(x_i) \approx q(x_i)\varepsilon_i(x_i) \left[1 + \sum_{j \neq i} \int dx_j q(x_j|x_i) [\varepsilon_j(x_j) - 1]\right],$$

which is also the first order expansion of $\tilde{p}_i^{\text{EP-FACT}}(x_i)$ w.r.t. $\varepsilon_j(x_j) - 1$, $j \neq i$. A further expansion w.r.t $\varepsilon_i(x_i) - 1$ leads to $\tilde{p}_i^{\text{EP-OPW}}(x_i)$, thus the two approximations are equal in first order. An advantage of $\tilde{p}_i^{\text{EP-FACT}}(x_i)$ is that it is non-negative by construction, while $\tilde{p}_i^{\text{EP-OPW}}(x_i)$ can take on negative values.

4.4 Approximating Predictive Densities in Gaussian Process models

In many real-world problems, the prior $p_0(x)$ is defined as a Gaussian process—most often in terms of moment parameters—and besides marginals, one is also interested in computing accurate approximations of the predictive densities

$$p(x_*|y) = Z_p^{-1} \int dx \, p_0(x_*|x) p_0(x) \prod_j t_j(x_j),$$

where x_* is a set of latent variables of which distribution we want to approximate. By defining the $\hat{q}(x,x_*) \propto p_0(x_*|x)q(x)$ and using the same line of argument as in (12), one can derive similar approximations as EP-FACT or EP-1STEP. For example, $\tilde{p}^{\text{EP-FACT}}$ has the form

$$\tilde{p}^{\text{EP-FACT}}(\boldsymbol{x}_*) \propto \hat{q}(\boldsymbol{x}_*) \prod_{j} \int dx_j \hat{q}(x_i | \boldsymbol{x}_*) \boldsymbol{\varepsilon}_j(x_j).$$

One can check that the marginalization and the conditioning of \hat{q} leads to rank k updates, where k is the dimensionality of x_* . For k=1, the complexity $\tilde{p}^{\text{EP-FACT}}(x_*)$ roughly scales with the complexity of $\tilde{p}_i^{\text{EP-FACT}}(x_i)$.

4.5 Comparisons on Toy Models

In the following, we compare the performance of the marginal approximations on a few low-dimensional toy models; complex real-world models are considered in Section 6. For most of the models presented below, we use a prior p_0 with a symmetric covariance matrix

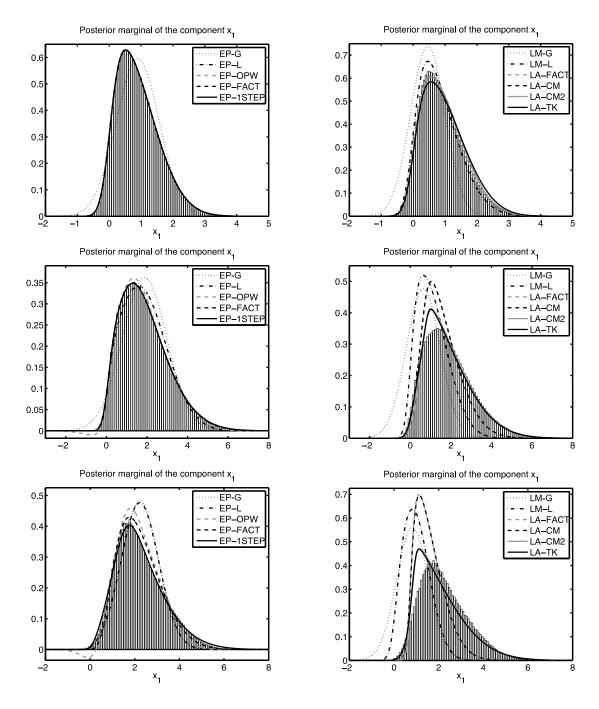
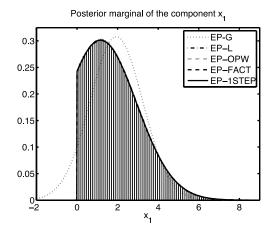


Figure 2: Various marginal corrections for a probit model with $t_i(x_i) = \Phi(4x_i)$ and identical variances and correlations in the prior p_0 , using expectation propagation (left column) and Laplace approximations (right column). The panels show the corrections for a 3-dimensional model with prior variances and correlations (v,c) = (1,0.25) (top), (v,c) = (4,0.9) (center) and for a 32-dimensional model (v,c) = (4,0.95) (bottom). Note how, the accuracy of the approximations decreases as the correlation, the prior variance and the dimension of the model increases.



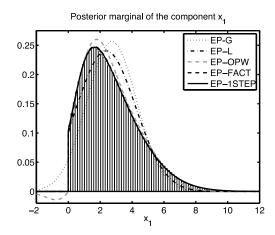
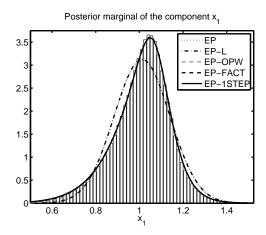


Figure 3: The posterior marginals of the first components of a 3-dimensional model with Heaviside terms with (v,c)=(4,0.5) (left) and (v,c)=(9,0.95) (right). The EP based approximations perform well even when the Laplace method is not applicable. The approximations have a similar behavior as in case of the probit model.

 $V = v[(1-c)I + c\mathbf{1}\mathbf{1}^T]$, where we vary the variance v and the correlation c. We have chosen the models below, because they are often used in practice, and they lead to sufficiently non-Gaussian posterior marginals.

Probit terms. The terms t_j are defined as $t_j(x_j) = \Phi(y_j x_j)$, where Φ is the standard Gaussian cumulative density function. This choice of terms is typically made in binary classification models, where $y_j \in \{-1,1\}$. In order to obtain skewed marginals, in this example we set $y_j = 4$. The top and center panels in Figure 2 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 (top), but is clearly insufficient when they are strong (center). 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 use strong prior correlations with moderate prior variances (v,c)=(4,0.95), we can see small differences (top-right). As we can see in Figure 2, EP-OPW performs slightly worse than EP-FACT and can indeed turn negative.

It is known that the Laplace method does not perform well on this model (e.g., Kuss and Rasmussen, 2005). The approximations it yields tend to be acceptable for weak correlations (top), with LA-CM and LA-FACT clearly outperforming LM-G and LM-L, but are far off when the correlations are stronger (center, bottom). These corrections suffer from essentially the same problems as the global Gaussian approximation based on Laplace's method: the mode and the inverse Hessian represent the mean and the covariance badly and fail to sufficiently improve it. It is interesting to see that LA-CM2 can be almost as accurate as LA-TK, while its computational complexity scales with LA-CM. The examples suggest that, at least in case of this model, LA-CM2 has the best accuracy/complexity tradeoff when compared to LA-CM and LA-TK.



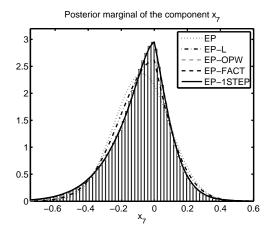
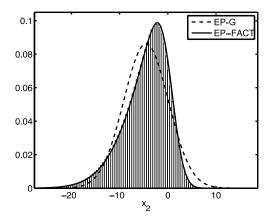


Figure 4: The posterior densities of a non-zero and a zero coefficient in a toy linear regression model with double exponential prior on the coefficients. It is interesting to compare the effects of the double exponential prior terms centered a zero on the quality of the local approximation EP-L. The effect is insignificant in the case the non-zero coefficient while in the case of the zero coefficient it has a strong effect, but the EP-L might still be quite inaccurate. We considered n=8 coefficients the first two being 1 and the rest 0 and we generated m=8 observables according to the model.

Step-function terms. Expectation propagation can still be applied when the Laplace method is not applicable. One such example is when the terms t_j are defined as $t_j(x_j) = \Theta(y_j x_j)$, where Θ is the step-function $\Theta(z) = \operatorname{sign}(z)$ for $z \neq 0$ and $\Theta(0) = 1$. We chose $y_i = 1$. The plots on the left of Figure 3 show the marginals of the first component of a three dimensional model with (v,c) = (4,0.5) (left) and (v,c) = (9,0.95) (right). The performance of the approximations is similar to those of the previous model, except that in this case, we are dealing with discontinuous marginals.

Linear regression with sparsifying prior. Another model where the Laplace method is not applicable is the linear regression model with double exponential prior on the coefficients. We choose a model with n=8 coefficients and m=8 observations—m being close to n led to the most interesting posterior marginals. The elements of the design matrix U are sampled according to the standard normal density and renormalized such that every column vector has unit length. The regression coefficients are chosen as $x=[1,1,0,\ldots,0]^T$ and the observations y_j are generated by $y=Ux+\epsilon$, where ϵ_j is normal with variance v=0.01. We take zero centered independent double exponential priors on the x_j coefficients. The panels of Figure 4 show a few posterior marginals of the regression coefficients x_j given the maximum a posteriori (MAP) hyper-parameters v and λ . The priors on the hyper-parameters are taken as independent and log-uniform. The approximations are accurate but in this case, the local approximations EP-L fail dramatically when the mass of the distribution is not close to zero.

A logistic regression model. We can try to use EP-FACT to approximate the marginal probability densities even when the terms t_i , $i \in \{1, ..., m\}$ depend on more than one variable or a linear transformation of the variables. As an example, we define the terms as $t_i(x) = \Phi(u_i^T x)$. In this case, the factorization principle does not apply, but we can still use the line of argument in Section 4.2 and



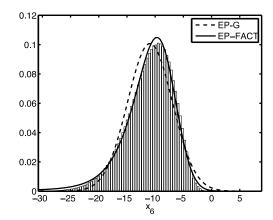


Figure 5: The posterior marginal approximation EP-FACT of the coefficients in a toy logistic regression model with Gaussian prior on the coefficients and moderate posterior correlations.. The panels show that even when the non-Gaussian terms depend on more than one variable and the posterior the approximation EP-FACT might still be accurate. We generated n = 8 coefficients and m = 8 observable variables.

evaluate how EP-FACT performs. The panels of Figure 5 show a few marginals of a model where we have chosen $u_i^j \sim \mathcal{N}(0,10)$ and an independent Gaussian prior $p_0(\boldsymbol{x}) = \prod_j N(x_j|0,v^{-1})$ with v=0.01. We used n=8 and m=8. Although one would expect that the factorization might lead to poor approximations, EP-FACT seems to approximate the marginals significantly better than the global approximation EP-G.

4.6 Computational Complexities of the Global Approximations in Sparse Gaussian Models

In this section, we review the computational complexities of the Laplace method and expectation propagation when applied to sparse Gaussian models, that is, models for which the n-dimensional precision matrix Q of the Gaussian prior is sparse. This is common in many practical applications in which the prior p_0 can be defined as a Gaussian Markov random field (e.g., van Gerven et al., 2009, 2010). We explore whether EP is indeed orders of magnitude slower, as suggested in Rue et al. (2009).

The computational complexity for both the (global) Laplace method and expectation propagation is dominated by several operations. 1) Computing the *Cholesky factor*, \tilde{L} of a matrix \tilde{Q} , for example, corresponding to the posterior approximation $\tilde{p}^{\text{EP-G}}$ or $\tilde{p}^{\text{LM-G}}$, with the same sparsity structure as the prior precision matrix Q. The computational complexity, denoted c_{chol} , scales typically with $\text{nnzeros}(Q)^2/n$, with nnzeros(Q) being the number of non-zeros in the precision matrix Q. 2) Computing the *diagonal elements of the inverse* of \tilde{Q} . For sparse matrices, these can be computed efficiently by solving the Takahashi equations (Takahashi et al., 1973; Erisman and Tinney, 1975), which take the Cholesky factor \tilde{L} as input. A detailed description of solving the Takahasi equations can be found in Section A of the Appendix. The computational complexity, denoted c_{taka} , scales with n^3 in the worst case, but typically scales with $\text{nnzeros}(L)^2/n$. In practice, we experienced that it is significantly more expensive than the Cholesky factorization, possibly due to the additional

covariance values one has to compute during the process.¹ 3) Solving a *triangular system* of the form $\tilde{L}a = b$, with corresponding computational complexity $c_{\text{tria}} \propto \text{nnzeros}(L)$.

The complexity of the latter two operations strongly depends on the number of non-zeros in the Cholesky factor, which should be kept to a minimum. There are various methods to achieve this by reordering the variables of the model. The approximate minimum degree reordering algorithm (Amestoy et al., 1996) seems to be the one with the best average performance (Ingram, 2006). Since the sparsity structure is fixed, the reordering algorithm has to be run only once, prior to running any other algorithm.

4.6.1 THE LAPLACE METHOD

To compute the global Gaussian approximation using the Laplace method, we first have to find the maximum a-posteriori solution. This can be done using, for example, the Newton method. Each Newton step requires one Cholesky factorization and solving two triangular systems. The off-diagonal elements of the posterior precision matrix \tilde{Q} are by construction equal to the off-diagonal elements of the prior precision matrix, so we only have to compute the n diagonal elements. To arrive at the lowest-order marginals $\tilde{p}_i^{\text{LM-G}}$ for all nodes i, we need the diagonal elements of the covariance matrix, the inverse of the precision matrix. These can be computed by solving the Takahashi equations, for which we can use the Cholesky factor computed in the last Newton step. Thus, computing the lowest order (Gaussian) marginals $\tilde{p}_i^{\text{LM-G}}$ for all variables x_i , $i=1,\ldots,n$ by the Laplace method scales in total with $n_{\text{steps}}^{\text{Newton}} \times (c_{\text{chol}} + 2 \times c_{\text{tria}}) + c_{\text{taka}}$.

4.6.2 EXPECTATION PROPAGATION

In order to update a term approximation $\tilde{t}_j(x_j)$, we compute $q^{\setminus j}(x_j)$ using the marginals $q(x_j)$ from the current global approximation q(x) and re-estimate the normalization constant and the first two moments of $t_j(x_j)q^{\setminus j}(x_j)$. In standard practice, the term approximations \tilde{t}_j are updated sequentially and all marginal means and variances are recomputed using rank one updates after each term 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}_j , $j=1,\ldots,n$.

A parallel EP step consists of: 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 moments of $\varepsilon_j(x_j)q(x_j)$ for all $j=1,\ldots,n$. This adds up to a computational complexity that scales with $n_{\text{steps}}^{\text{EP}} \times \left(c_{\text{chol}} + 2 \times c_{\text{tria}} + c_{\text{taka}} + n \times n_{\text{quad}}\right)$. After convergence, EP yields the lowest order marginals $\tilde{p}_i^{\text{EP-G}}$ for all variables x_i , $i=1,\ldots,n$.

Because of the parallel schedule, we can make use of exactly the same computational tricks as with the Laplace method (Cholesky, Takahashi). Since solving the Takahashi equations for large n dominates all other operations, the main difference between the Laplace method and EP is that for EP we have to solve these equations a number of times, namely the number of EP steps, yet for Laplace only once. Initializing the term approximations in EP to the terms obtained by the Laplace method and then performing a few EP steps to obtain better estimates of the probability mass, makes EP just a (small) constant factor slower than Laplace. For efficient sequential updating

^{1.} 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
$q(x_j x_i)$	$c_{\text{tria}} + n \times n_{\text{grid}}$	$c_{\text{tria}} + n \times n_{\text{grid}}$	$c_{\text{tria}} + n \times n_{\text{grid}}$	$c_{\text{tria}} + n \times n_{\text{grid}}$
$\tilde{\varepsilon}(x_j;x_i)$	$n \times n_{\mathrm{grid}}$	$n \times n_{\text{grid}}$	$n \times n_{\text{grid}} \times n_{\text{quad}}$	$n \times n_{\text{grid}} \times n_{\text{quad}}$
Norm. or dets	$c_{ m chol} imes n_{ m grid}$	$n \times n_{\mathrm{grid}}$	$c_{ m chol} imes n_{ m grid}$	$n \times n_{\mathrm{grid}}$

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 computation 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 and n_{quad} to the number of Gauss-Hermite quadrature nodes for x_i .

of EP, we would need a fast one-rank Takahashi update (or something similar), which, to the best of our knowledge, does not exist yet.

It is interesting to realize that since for any $Q_{ij} \neq 0$ the Takahashi equations also provide $[Q^{-1}]_{ij}$, we can run EP using the factors $t_{ij}(x_i,x_j) = t_i(x_i)^{1/n_i}t_j(x_j)^{1/n_j}$ where n_k is the number of neighbors of node k according to the adjacency matrix defined by the structure of Q. This increases the amount of computation, but the approximation might be more accurate.

4.7 Computational Complexities of Marginal Approximations

After running the global approximation to obtain the lowest order approximation, we are left with some Gaussian q(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 $\tilde{p}_i^{\text{LM-L}}$ and $\tilde{p}_i^{\text{EP-L}}$ we get more or less for free. All other correction methods require the computation of the conditional densities $q(x_j|x_i)$. The conditional variance is independent of x_i , the conditional mean is a linear function of x_i . Computing $q(x_j|x_i)$ at all grid points for each j then amounts to solving two sparse triangular systems and $(n-1)\times n_{\text{grid}}$ evaluations. To arrive at the term approximations $\tilde{\mathbf{e}}(x_j;x_i)$, 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-OPW and EP-FACT, we then simply have to compute a product or sum of n normalization terms. For LA-CM and EP-1STEP, we need to compute the determinant of an (n-1)-dimensional sparse matrix, which costs a Cholesky factorization. For LA-CM2 an additional c_{tria} has to be added for each x_i .

5. 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 density of the hyper-parameters given the observations, which is approximated by $\tilde{p}(\theta|y) \propto \tilde{p}(y|\theta) p(\theta)$, where $\tilde{p}(y|\theta)$ is the evidence

approximation provided by the Laplace method or expectation propagation. For the moment we assume that the approximate posterior density of the hyper-parameters is unimodal.

We propose a slight modification of the method used by Rue et al. (2009). Their method explores the space of the hyper-parameters in the eigen-space corresponding to the modal configuration and can be described briefly as: (1) compute the modal configuration (μ, Σ) of $\log \tilde{p}(\theta|\boldsymbol{y})$, (2) starting from the mode μ , select a set of uniformly spaced nodes \mathcal{X}_i along the scaled eigenvectors $\sqrt{\lambda_i}\boldsymbol{u}_i$ —here $\Sigma = \boldsymbol{U}\Lambda\boldsymbol{U}^T$ —by thresholding at both ends according to $\log \tilde{p}(\mu|\boldsymbol{y}) - \log \tilde{p}(\mu + k_i \Delta \sqrt{\lambda_i}\boldsymbol{u}_i|\boldsymbol{y}) < \delta, k_i \in \mathbb{Z}$, and finally (3) use all hyper-parameters corresponding to the nodes of the product grid $\mathcal{X}_1 \times \ldots \times \mathcal{X}_d, d = \dim(\theta)$ and satisfying the latter thresholding condition, to perform numerical quadrature using the rectangle rule.

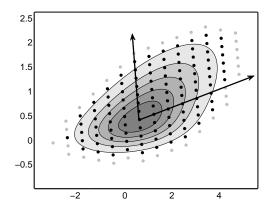
Since the computational bottleneck of the procedure is the evaluation of the approximate evidence, we propose to improve this method by selecting the nodes—step (2) from above—in a different way: we keep the thresholding condition but we do a breadth-first search with regard to (k_1,\ldots,k_d) on the grid graph \mathbb{Z}^d . We start from the origin and the hyper-parameter values that do not satisfy the thresholding condition are not included in the set of nodes whose neighbors we search. This simple modification proves to be very economical, since when exploring the volume around the mode, only the hyper-parameters that form the boundary surface are explored, but not selected. Thus, the proportion of useless computational time is the ratio of surface to volume. Although the boundary nodes do not satisfy the thresholding conditions, we can still use them in the numerical procedure. The number of grid points to be evaluated grows exponentially, as it does for the method in Rue et al. (2009). The difference is that in our method it roughly grows proportional to the volume of a d-dimensional sphere, whereas in the case of the method in Rue et al. (2009) it relates to the (larger) volume of a d-dimensional cube. Figure 6 illustrates the methods on a two-dimensional example. When the posterior density is not unimodal then we suggest to use a d-dimensional uniformly spaced grid, that is, $\Sigma = I$ and choose a well suited μ and threshold δ which allows the exploration of the most significant modes. Once the hyper-parameters $\{\theta_1, \dots, \theta_m\}$ are selected, the integration of the corrected approximate marginals over the hyper-parameter's approximate posterior density can be written as

$$\tilde{p}\left(x_{i}|\boldsymbol{y}\right) = \frac{\sum_{j=1}^{m} \tilde{p}\left(x_{i}|\boldsymbol{y}, \theta_{j}\right) \tilde{p}\left(\theta_{j}|\boldsymbol{y}\right)}{\sum_{j=1}^{m} \tilde{p}\left(\theta_{j}|\boldsymbol{y}\right)},$$

implying that the proposed procedure is similar to a reasonably efficient sampling procedure.

6. Examples

As real-world examples, we chose four models: a stochastic volatility model (Zoeter and Heskes, 2005; Rue et al., 2009), a log Gaussian Cox process model (Rue et al., 2009), a Gaussian process binary classification model (Kuss and Rasmussen, 2005) and a ranking model (Birlutiu and Heskes, 2007). Our aim is to show that the EP based correction methods can be as accurate as the Laplace approximation based ones and given that we have a sparse Gaussian prior, EP can be considered as an alternative to the Laplace method even when the number of variables is of the order of tens of thousands.



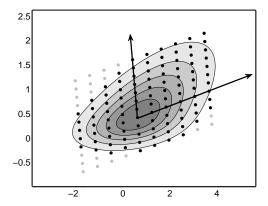


Figure 6: A comparison of the points selected by the thresholding breadth-first search procedure (left panel) and the method proposed by Rue et al. (2009) (right panel) when exploring in the eigen-space corresponding to the modal configuration. The black dots show the selected points while the gray ones stand for the ones that do not satisfy the thresholding condition. The principal axes on the figure are not perpendicular because of the different scaling of the axes. The number of evaluations in our method roughly grows proportional to the volume of a *d*-dimensional sphere, whereas the method of Rue et al. (2009) relates to the (larger) volume of a *d*-dimensional cube.

6.1 A Stochastic Volatility Model

As a first example for a sparse Gaussian model, we implemented the stochastic volatility model presented in Zoeter and Heskes (2005) where the authors used a sequential (global) EP algorithm to approximate the posterior density. The same model was used by Rue et al. (2009) to show that the global Laplace approximation is by magnitudes faster in sparse models than a sequential EP algorithm. They also showed that their marginal approximations work well on this model.

The data set consists of 945 samples of the daily difference of the pound-dollar exchange rate from October 1st, 1981, to June 28th, 1995. 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 quantity η_t governing the volatility is a linear predictor defined 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 with a prior $\mu \sim N(\mu|0,1)$. Thus the prior on (f_1,\ldots,f_T,μ) is a sparse latent Gaussian field. 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 joint density of the stochastic volatility model is

$$p(y, f, \mu, \tau, \phi) = \prod_{t=1}^{T} N(y_{t}|0, e^{f_{t}+\mu}) N(f_{1}|0, 1) \prod_{t=2}^{T} N(f_{t}|\phi f_{t-1}, 1/\tau) \times N(\mu|0, 1) \Gamma(\tau|1, 10) N\left(\log\left(\frac{1+\phi}{1-\phi}\right)|0, 3\right) \left(\frac{2}{1-\phi^{2}}\right),$$

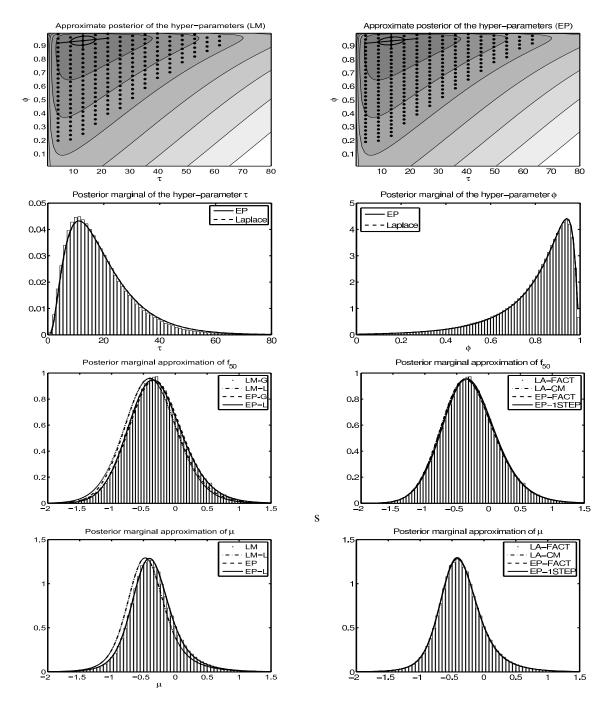


Figure 7: Plots of the posterior densities in the stochastic volatility model in Section 6.1. Figure panels show the logarithm of the approximate posterior density of the hyper-parameters using EP (top-right) and the Laplace method (top-left), their marginals (second row) and the posterior marginal approximations of f_{50} and μ (bottom rows) when integrated over the corresponding approximations of the hyper-parameters' posterior density. Dots show the hyper-parameters used for numerical integration; ellipses visualize the Hessian at the approximate posterior density's mode. The rest of the panels show the posterior density approximations of f_{50} and μ .

where $\Gamma(\cdot|k,\theta)$ denotes the Gamma density with mean value $k\theta$. Rue et al. (2009) propose to use the first 50 observations, both because of using the whole data set makes the approximation problem easier and because of comparison to Zoeter and Heskes (2005). For comparison, we used the same number of observations.

The results are shown in Figure 7. The Laplace and EP approximation of the evidence are nearly indistinguishable (top-row), as are the posterior marginals of the hyper-parameters (second row). Here EP is around a factor 5 slower than Laplace. The posterior marginals of f_{50} and μ obtained using the more involved methods (bottom rows) are practically indistinguishable from each other and the gold (sampling) standard. This is not the case for the cheaper variants LM-G, EP-G, and LM-L, but *is* the case for EP-L (third row): apparently to obtain excellent posterior marginals on this model, there is no need for (computationally expensive) corrections, but it suffices to compute a single global EP approximation per hyper-parameter setting and correct this for the (non-Gaussian) local term.

6.2 A log-Gaussian Cox Process Model

As a large sized example, we implemented the Laplace approximation and expectation propagation for the log-Gaussian Cox process model applied to the tropical rainforest bio-diversity data as presented in Rue et al. (2009). The observational data used in Rue et al. (2009) is the number of trees y_{ij} form a certain species in a small rectangular rainforest area indexed by i = 1, ..., 201 and j = 1, ..., 101 with mean altitude a_{ij} and gradient g_{ij} . The data is modeled by a discretized Poisson point process in two dimensions and the log of the mean parameter η_{ij} is defined as a Gaussian field. This means that the observations y_{ij} are taken to be Poisson distributed with mean $w_{ij}e^{\eta_{ij}}$, where the parameters w_{ij} are proportional to the size of the area where y_{ij} is measured. Since Rue et al. (2009) consider rectangular areas of the equal size, in their model w_{ij} is constant.

The latent Gaussian field η_{ij} modeling the log of the mean is defined as

$$\eta_{ij} = \beta_a a_{ij} + \beta_g g_{ij} + \beta_0 + f_{ij}^{(s)} + f_{ij}^{(u)}$$

where a_{ij} and g_{ij} are scalar quantities specifying altitude and gradient data, β_a and β_g are the corresponding linear coefficients and β_0 is a bias parameter. The latent fields $f^{(s)}$ and $f^{(u)}$ are defined as follows: $f^{(s)}$ is a second-order polynomial intrinsic Gaussian Markov random field with precision parameter e^{ν_s} constructed to mimic a thin plate spline on a uniform two dimensional grid, while $f^{(u)}$ is an independent field with $f^{(u)}_{ij} \sim \mathcal{N}(0, e^{-\nu_{\eta}})$ included to model the noise. The fields $f^{(s)}$ and $f^{(u)}$ are modeling the unobserved spatially structured or unstructured covariates. Independent wide priors $\mathcal{N}(0, \nu_{\beta}^{-1})$ are taken on β_a, β_g and β_0 , with $\nu_{\beta}^{-1} = 10^3$, thus the field $f^{(s)}$ explains the assumed a-priori correlation in η . We worked with the data set used in the INLA software package (Martino and Rue, 2009). The data set contains the corresponding a_{ij}, g_{ij}, w_{ij} and y_{ij} for a grid size of 101×201 . We also used the same modeling approach, that is, we have taken $(\eta^T, f^{(s)^T}, \beta_a, \beta_g, \beta_0)^T$ as latent variable, thus having an inference problem of dimension 40605. The joint density of the

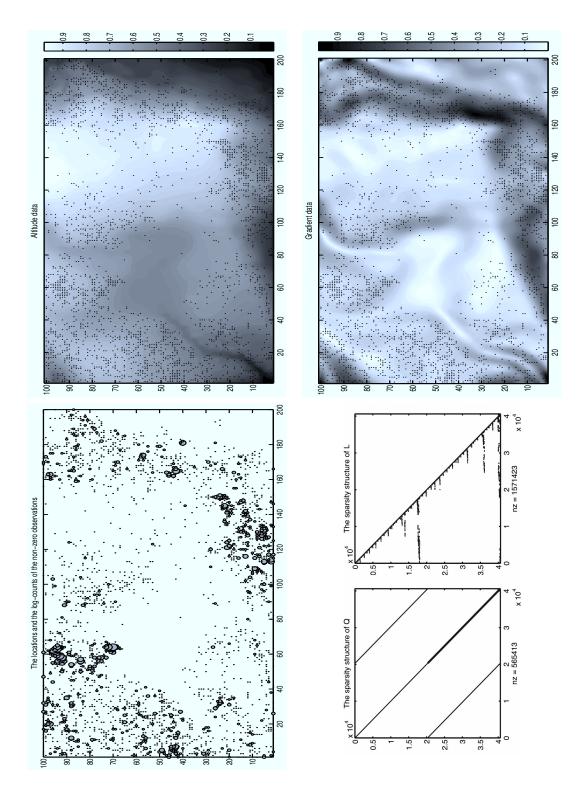


Figure 8: The panels show the altitude a_{ij} , gradient g_{ij} and the non-zero observation y_{ij} data for the log-Gaussian Cox process model in Section 6.2 together with the sparsity structure of Q and the Cholesky factor L of its approximate minimum degree reordering.

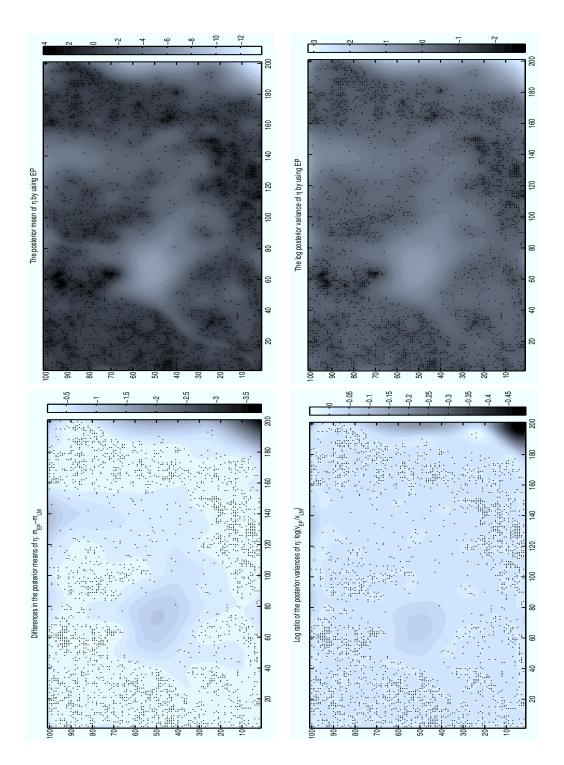


Figure 9: The approximate posterior mean and variance of the Gaussian random field η from the log-Gaussian Cox process model in Section 6.2. The top figures show the approximation obtained by the EP algorithm. The bottom panels show the comparison of the former to the approximation obtained by the Laplace method. The black contour curve in the bottom-left panel corresponds to the zero value.

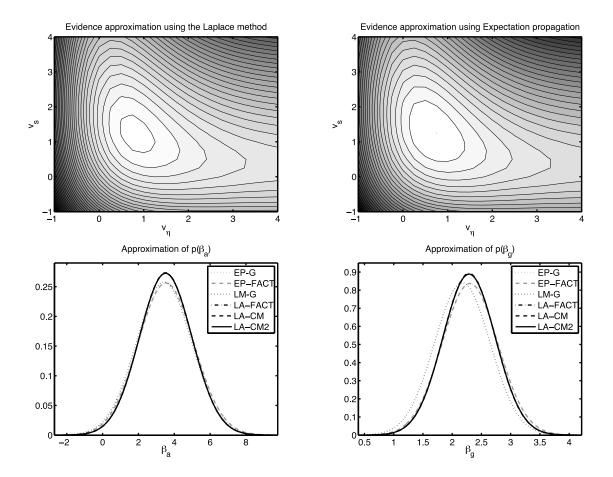


Figure 10: The posterior approximations of the evidence (top) and β_a and β_g (bottom). The Laplace method results in similar evidence estimates as EP (the level curves on the top panels show identical levels). The marginal approximations show marginals for the approximate MAP hyper-parameters.

log-Gaussian Cox process model is

$$p(\boldsymbol{y}, \boldsymbol{\eta}, \boldsymbol{f}^{(s)}, \boldsymbol{\beta}_{a}, \boldsymbol{\beta}_{g}, \boldsymbol{\beta}_{0} | \boldsymbol{\nu}_{\boldsymbol{\eta}}, \boldsymbol{\nu}_{s}, \boldsymbol{a}, \boldsymbol{g}, \boldsymbol{w}) =$$

$$= \prod_{ij} \operatorname{Poisson}(y_{ij} | w_{ij} e^{\boldsymbol{\eta}_{ij}}) N\left(\boldsymbol{\eta}_{ij} | f_{ij}^{(s)} + a_{ij} \boldsymbol{\beta}_{a} + g_{ij} \boldsymbol{\beta}_{g} + \boldsymbol{\beta}_{0}, e^{-\boldsymbol{\nu}_{\boldsymbol{\eta}}}\right)$$

$$\times \left(\frac{\boldsymbol{\nu}_{s}}{2\pi}\right)^{N/2} |\boldsymbol{S}|_{*}^{1/2} \exp\left\{-\frac{1}{2} e^{\boldsymbol{\nu}_{s}} \boldsymbol{f}^{(s)}^{T} \boldsymbol{S} \boldsymbol{f}^{(s)}\right\} N\left(\boldsymbol{\beta}_{a}, \boldsymbol{\beta}_{g}, \boldsymbol{\beta}_{0} | \boldsymbol{0}, 10^{3} \boldsymbol{I}\right),$$

where $|S|_*$ is the generalized determinant—an irrelevant constant—of the structure matrix S consisting of the finite difference coefficients of a second order improper polynomial Gaussian Markov random field on a uniform two dimensional grid—with the corresponding boundary conditions (Rue and Held, 2005). We used uninformative priors for v_{η} and v_s . The bottom-right panels of Figure 8 show the sparsity structure of the precision matrix Q corresponding to the Gaussian random vector $(\eta^T, f^{(s)^T}, \beta_a, \beta_g, \beta_0)^T$ and the sparsity structure of its Cholesky factor L when Q is reordered with the AMD algorithm.

Expectation propagation was initialized using the term approximations corresponding to the Laplace method. Figure 8 shows the data we used and Figure 9 shows the mean values and standard deviations of the log intensity η when using the EP algorithm and the Laplace method with the hyper-parameter fixed to their corresponding approximate a posteriori (MAP) value.

The top panels of Figure 10 show the evidence approximations while the bottom panels show the marginal approximations for the corresponding MAP hyper-parameters. For β_a , there is a slight difference in variance between the Laplace approximation and the EP based methods, while for β_g , besides a similar effect, the approximation methods also improve on the mean of LM-G. It seems that EP-G is a sufficiently good approximation and EP-FACT does not really improve on it.

6.3 A Gaussian Process Model for Binary Classification

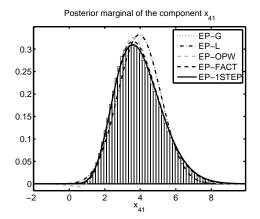
In this section we revisit and detail the probit model presented in Section 4.5. We use it in a binary classification problem with a Gaussian process prior on the latent variables. The data consists of the inputs $u_j \in \mathbb{R}^d$, j = 1, ..., n and the corresponding binary outputs $y_j \in \{-1, 1\}$, j = 1, ..., n.

The model is defined as follows. The binary observables y_i are assumed to be Bernoulli distributed and conditionally independent given a set of latent variables $x_j \in \mathbb{R}$, j = 1, ..., n which are controlling the parameters of the distribution through the cumulative density function Φ of the standard normal distribution, that is, $p(y_j|x_j) = \Phi(x_j)^{(1+y_j)/2}(1-\Phi(x_j))^{(1-y_j)/2}$. The latent variables $x_j = x(u_j)$ are modeled as the values at locations u_i of a zero mean Gaussian process x with a (positive definite) covariance function $c(\cdot, \cdot)$. The joint density of the model can then be written as

$$p(\boldsymbol{y}, \boldsymbol{x} | \{\boldsymbol{u}_j\}_j, c) \propto \prod_j \Phi(y_j x_j) N(\boldsymbol{x} | \boldsymbol{0}, [c(\boldsymbol{u}_i, \boldsymbol{u}_k)]_{i,k}),$$

where $[c(\boldsymbol{u}_i, \boldsymbol{u}_k)]_{i,k}$ denotes the matrix formed by the covariance values $c(\boldsymbol{u}_i, \boldsymbol{u}_k)$, $i, k = 1, \ldots, n$. We chose the Ionosphere² data set and a zero mean Gaussian process with covariance function $c(\boldsymbol{u}_i, \boldsymbol{u}_k) = \exp(a - e^{v}||\boldsymbol{u}_i - \boldsymbol{u}_k||^2)$. This setting yields a non-sparse precision matrix, therefore, the speed-up arguments do not apply. However, the parallel updating scheme is still applicable and it does not have higher complexity than the serial one.

^{2.} The data set is publicly available at http://archive.ics.uci.edu/ml/datasets/Ionosphere.



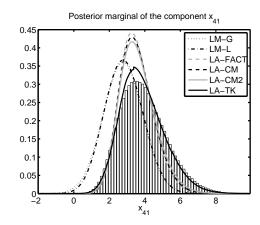


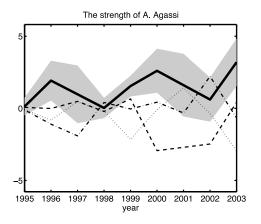
Figure 11: The approximate posterior marginal approximations of a Gaussian process binary classification model, with the hyper-parameters set to the approximate MAP values yielded by EP. The behavior of the marginals is similar to that in Figure 2, however, in this case, the correction provided by LA-CM2 is not that significant as in Figure 2.

Kuss and Rasmussen (2005) showed that on this model and data set EP leads to accurate approximations of the evidence while the Laplace method is substantially less accurate. We propose to illustrate how this behaviour manifests itself when approximating marginals. We use the whole set of n = 351 data points and compare the resulting marginals with the histograms obtained from 1.5×10^6 samples by using elliptical slice sampling (Murray et al., 2010). The hyper-parameters are set to the approximate MAP values obtained from EP's evidence approximation with uniform priors on a and v.

It turned out that many posterior marginal densities are skewed, however, most of the skewed marginals are well approximated by EP-L (the marginal of EP's tilted distribution). The panels of Figure 11 show the approximate posterior marginal densities of the latent variable with j=41. These approximate posterior marginals exhibit a similar behavior like the ones in Figure 2.

6.4 A Ranking Model

To show that we can implement linear constraints with EP and that the factorization principle might work even in is cases when the non-Gaussian terms depend on more than one variable, we use a ranking model for rating players in sports competitions. The model is a simplified version of the models presented in Dangauthier et al. (2008) and Birlutiu and Heskes (2007) and we only consider it as an example to support the above mentioned claims. We assume that a player j is characterized by his/her strength which at time t is $x_t^{(j)}$. The prior on the evolution of the players' strength $x_t = (x_t^{(1)}, \dots, x_t^{(n)})$ is taken to be a factorizing AR(1) model. Each game between two players is represented by the triple (i, j, t) and the collection of these triples is denoted by G. We assume that the outcomes of the games are a binary variables $y_{i,j,t} \in \{-1,1\}$, the games are conditionally independent given the players strengths and the probability of player i winning the game against player j at time t is $\Phi(x_t^{(i)} - x_t^{(j)})$, where Φ is the standard normal cumulative density function. To



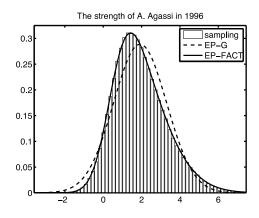


Figure 12: The left panel shows the mean strengths of the players A. Agassi (cont.), Y. Kafelnikov (dashed), C. Moya (dashed-dotted), and T. Henman (dotted) with the standard deviations of A. Agassi's strength based on the ranking model presented in Section 6.4. The data set consists of the games played by these players against each other in the years 1995-2003. We implemented linear constraints such that the players strength sum to zero in every year. The left panel shows that this indeed holds for the means. The right panel shows A. Agassi's strength distribution in 1996 which is a non-Gaussian density and can be well approximated using EP-FACT.

implement linear constraints, we constrain the players' strength to sum to zero at any given time t. These constraints are purely artificial and are only considered for illustration purposes.

The joint posterior density of the players' strength is given by

$$p(x^{1},...,x^{T}|y,v_{1},v,a) \propto \prod_{t=1}^{T} \delta_{0}(\mathbf{1}^{T}x_{t}) \prod_{(i,j,t)\in G} \Phi(y_{i,j,t}(x_{t}^{(i)}-x_{t}^{(j)})) \times \prod_{j=1}^{n} N(x_{t}^{(j)}|0,v_{1}) \prod_{t=1}^{T-1} N(x_{t+1}^{(j)}|ax_{t}^{(j)},v).$$

We approximate this density with a Gaussian density using EP and we use the factorized corrections EP-FACT, to improve on the Gaussian marginals. The prior on the players strengths is a sparse Gaussian Markov random field, thus we can apply the methods presented in Section 4.6.2.

We have chosen a data set consisting of four³ tennis players and their ATP tournament games played against each other form 1995 to 2003. There was a total of 45 games. To obtain reasonably skewed marginals, we chose $v_1 = 1$, a = 1 and v = 9. The left panel in Figure 12 shows the evolution of the players' mean strengths and the corresponding standard deviations for the best player. Note that the players' mean strengths average to zero at all times. The right panel shows that the factorized approximations EP-FACT, can indeed improve on the Gaussian marginal approximations computed by EP even in models where non-Gaussian terms depend on more than one variable. This might be due to the relatively sparse interaction between the variables $x_t^{(j)}$, t = 1, ..., T, j = 1, ..., n.

^{3.} We have chosen A. Agassi, Y. Kafelnikov, C. Moya and T Henman.

7. Discussion

We introduced several methods to improve on the marginal approximations obtained by marginalizing the global approximations. The approximation denoted by EP-FACT seems to be, in most cases, both accurate and fast. An improvement in accuracy can be achieved with some additional computational cost by using EP-1STEP. We showed that by using a parallel EP scheduling the computational complexity of EP in sparse Gaussian model can scale with the computational complexity of the Laplace method.

There are many options for further improvement, in particular with respect to 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 , are applicable to expectation propagation. However, if the computation of the determinant in EP-1STEP dominates the computation time, the factorized approximation EP-FACT may be a faster but less accurate alternative.

One of the main problems of expectation propagation is that it is not guaranteed to converge and may run into numerical problems. There were no problems with the convergence of EP in the problems considered in this paper, but even when there are, it can still be useful to start from the Laplace solution and perform a few EP steps to get closer to the main mass of the probability instead of relying on the mode and the curvature.

For models with weak correlations and smooth nonlinearities, any approximation method gives reasonably good results. However, it is possible to come up with cases (strong correlations, hard nonlinearities), where any deterministic approximation method fails. The most interesting problems are somewhere in between, and for those we can hardly tell how advanced and computationally intensive an approximation method we need. The heuristic suggested in Rue et al. (2009), to systematically increase the complexity and stop when no further changes can be obtained, appears to be 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 far off. It would be interesting to obtain a better theoretical understanding of the (asymptotic) approximation errors implied by the different approaches.

Acknowledgments

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

Appendix A. Solving the Takahashi Equations

The Takahashi equations (Takahashi et al., 1973) aim to compute certain elements of the inverse of a positive definite matrix from its Cholesky factor. The derivation of the equations or the algorithm can be found in many papers (e.g., Erisman and Tinney, 1975; Rue et al., 2009). In the following we present the line of arguments in Rue et al. (2009). Let $Q = LL^T$, $z \sim N(\mathbf{0}, I)$ and $L^Tx = z$. Then using the notation $V = Q^{-1}$ we find that $x \sim N(\mathbf{0}, V)$. The equations $L^Tx = z$ can be rewritten as $L_{ii}x_i = z_i - L_{ii}^{-1}\sum_{k=i+1}^n L_{ki}x_k$. Multiplying both sides with x_j , $j \geq n$, using $z = L^{-T}x$ and taking expectations we arrive at the Takahashi equations $V_{ij} = \delta_{ij}L_{ii}^{-2} - L_{ii}^{-1}\sum_{k=i+1}^n L_{ki}V_{kj}$. Since we only

want to compute the diagonal of V or the elements V_{ij} for which $L_{ij} \neq 0$, the algorithm can be written in the following MATLAB friendly form

```
1: function V = SolveTakahashi(L)
2: for i = n : -1 : 1
3: I = \left\{ j : L_{ij} \neq 0, \ j > i \right\}
4: V_{I,i} = -V_{I,I}L_{I,i}/L_{i,i}
5: V_{i,I} = V_{I,i}^T
6: V_{i,i} = 1/L_{i,i}^2 - V_{i,I}L_{I,i}/L_{ii}
7: end
```

The complexity of this algorithm scales with nonzeros(Q)²/n.

Appendix B. Gaussian Formulas

The first and second moments of a distribution $p(\mathbf{x}) = Z^{-1}(\mathbf{m}, \mathbf{V}) f(\mathbf{x}) q(\mathbf{x})$ with $q(\mathbf{x}) = N(\mathbf{x} | \mathbf{m}, \mathbf{V})$ are given by

$$\begin{aligned} \mathbf{E}_p[\mathbf{x}] &= \mathbf{m} + \mathbf{V} \nabla_{\mathbf{m}} \log Z(\mathbf{m}, \mathbf{V}), \\ \mathbf{V}_p[\mathbf{x}] &= \mathbf{V} + \mathbf{V} \nabla_{\mathbf{mm}}^2 \log Z(\mathbf{m}, \mathbf{V}) \mathbf{V}. \end{aligned}$$

Applying integration by parts, one can show that the moments of p can also be written in the form

$$E_{p}[\boldsymbol{x}] = \boldsymbol{m} + \frac{1}{Z} \boldsymbol{V} E_{q} [\nabla_{\boldsymbol{x}} f],$$

$$V_{p}[\boldsymbol{x}] = \boldsymbol{V} + \frac{1}{Z^{2}} \boldsymbol{V} \left[Z E_{q} [\nabla_{\boldsymbol{x}}^{2} f] - E_{q} [\nabla_{\boldsymbol{x}} f] E_{q} [\nabla_{\boldsymbol{x}} f]^{T} \right] \boldsymbol{V},$$

provided that $f(x)e^{-x^Tx}$ and $\frac{\partial f(x)}{\partial x}e^{-x^Tx}$ vanish at infinity and the required differentials and integrals exist.

Appendix C. Details of EP in Latent Gaussian Models

Assume the distribution has the form

$$p(\boldsymbol{x}) \propto p_0(\boldsymbol{x}) \prod_i t_i(\boldsymbol{U}_i \boldsymbol{x}),$$

where U_i are linear transformations. This formulation includes both the representations when t_j depend only on a subset of parameters, that is, $t_i(\mathbf{x}) = t_i(\mathbf{x}_{I_i})$ with $U_i = I_{\cdot,I_i}$ and the representation used in logistic regression, where U_i is the i^{th} row of the design matrix. Here we present the details of the α -fractional or power EP where the updates are performed on $\tilde{t}_i^{\alpha}(\mathbf{x})$.

C.1 Computing \tilde{t}_i^{new}

First we compute the form of the term approximations, and show that \tilde{t}_i has a low rank representation. Let $q(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{m},\boldsymbol{V})$ and let $\tilde{\boldsymbol{h}} = \boldsymbol{V}^{-1}\boldsymbol{m}$, $\tilde{\boldsymbol{Q}} = \boldsymbol{V}^{-1}$ the canonical parameters of $q(\boldsymbol{x})$. We use $q^{\setminus i}(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{m}^{\setminus i},\boldsymbol{V}^{\setminus i})$ to denote the distribution $q^{\setminus i}(\boldsymbol{x}) \propto q(\boldsymbol{x})/\tilde{t}_i^{\alpha}(\boldsymbol{x})$. After

some calculus one can show that the moment matching Gaussian $q^{new}(x) = N(x|m^{new}, V^{new})$ of $q_i(x) \propto t_i^{\alpha}(x)q^{i}(x)$ is given by

$$egin{aligned} oldsymbol{m}^{new} &=& oldsymbol{m}^{\setminus i} + oldsymbol{V}^{\setminus i} oldsymbol{U}_i^T \left[oldsymbol{U}_i oldsymbol{V}^{\setminus i} oldsymbol{U}_i^T \right]^{-1} \left[\operatorname{E}\left[oldsymbol{z}_i\right] - oldsymbol{U}_i oldsymbol{m}^{\setminus i} \right], \ oldsymbol{V}^{new} &=& oldsymbol{V}^{\setminus i} + oldsymbol{V}^{\setminus i} oldsymbol{U}_i^T \left[oldsymbol{U}_i oldsymbol{V}^{\setminus i} oldsymbol{U}_i^T \right]^{-1} \left[\operatorname{V}\left[oldsymbol{z}_i\right] - oldsymbol{U}_i oldsymbol{V}^{\setminus i} oldsymbol{U}_i^T \right] \left[oldsymbol{U}_i oldsymbol{V}^{\setminus i} oldsymbol{U}_i^T \right]^{-1} oldsymbol{U}_i oldsymbol{V}^{\setminus i}, \end{aligned}$$

where z_i is a random variable distributed as $z_i \sim t(z_i)^{\alpha} N(z_i | U_i m^{\setminus i}, U_i V^{\setminus i} U_i^T)$. The update for the term approximation $\tilde{t}_i(x)$ is given by $(\tilde{t}_i^{new}(x))^{\alpha} \propto q^{new}(x)/q^{\setminus i}(x)$. The latter division yields

$$[\boldsymbol{V}^{new}]^{-1} - [\boldsymbol{V}^{\setminus i}]^{-1} = \boldsymbol{U}_{i}^{T} \left[\boldsymbol{V} \left[\boldsymbol{z}_{i} \right]^{-1} - \left[\boldsymbol{U}_{i} \boldsymbol{V}^{\setminus i} \boldsymbol{U}_{i}^{T} \right]^{-1} \right] \boldsymbol{U}_{i}, \tag{15}$$

$$[\boldsymbol{V}^{new}]^{-1} \boldsymbol{m}^{new} - [\boldsymbol{V}^{\setminus i}]^{-1} \boldsymbol{m}^{\setminus i} = \boldsymbol{U}_{i}^{T} \left[\boldsymbol{V} \left[\boldsymbol{z}_{i} \right]^{-1} \boldsymbol{E} \left[\boldsymbol{z}_{i} \right] - \left[\boldsymbol{U}_{i} \boldsymbol{V}^{\setminus i} \boldsymbol{U}_{i}^{T} \right]^{-1} \boldsymbol{U}_{i} \boldsymbol{m}^{\setminus i} \right] \tag{16}$$

leading to

$$\tilde{t}_i^{new}(\boldsymbol{x}) \propto \exp\left(\left(\boldsymbol{U}_j \boldsymbol{x}\right)^T \tilde{\boldsymbol{h}}^j - \frac{1}{2} \left(\boldsymbol{U}_j \boldsymbol{x}\right)^T \tilde{\boldsymbol{Q}}^j \left(\boldsymbol{U}_j \boldsymbol{x}\right)\right),$$

where \tilde{h}^i and \tilde{Q}^i are given by the corresponding quantities in (15) and (16). The approximating distribution q is defined by the canonical parameters

$$egin{array}{lll} ilde{m{h}} &=& m{h} + \sum_i m{U}_i^T ilde{m{h}}^i, \ & ilde{m{Q}} &=& m{Q} + \sum_i m{U}_i^T ilde{m{Q}}^i m{U}_i, \end{array}$$

that is, the sum over the parameters of \tilde{t}_i and the parameters of the prior $p_0(x) \propto \exp(h^T x - x^T Q x/2)$.

C.2 Computing the Cavity Distribution $q^{\setminus i}$

Now, we turn our attention to the computation of the distribution $q^{\setminus i}$. The quantities we are interested in are $U_i m^{\setminus i}$ and $U_i V^{\setminus i} U_i^T$. After some calculus, one can show that these are given by

$$\begin{aligned} \boldsymbol{U}_{i}\boldsymbol{V}^{\setminus i}\boldsymbol{U}_{i}^{T} &= \boldsymbol{U}_{i}\left(\tilde{\boldsymbol{Q}} - \alpha\boldsymbol{U}_{i}^{T}\tilde{\boldsymbol{Q}}^{i}\boldsymbol{U}_{i}\right)^{-1}\boldsymbol{U}_{i}^{T} \\ &= \left(\boldsymbol{U}_{i}\boldsymbol{V}\boldsymbol{U}_{i}^{T}\right)\left(\boldsymbol{I} - \alpha\tilde{\boldsymbol{Q}}^{i}\left(\boldsymbol{U}_{i}\boldsymbol{V}\boldsymbol{U}_{i}^{T}\right)\right)^{-1}, \\ \boldsymbol{U}_{i}\boldsymbol{m}^{\setminus i} &= \boldsymbol{U}_{i}\left(\tilde{\boldsymbol{Q}} - \alpha\boldsymbol{U}_{i}^{T}\tilde{\boldsymbol{Q}}^{i}\boldsymbol{U}_{i}\right)^{-1}\left(\tilde{\boldsymbol{h}} - \alpha\boldsymbol{U}_{i}^{T}\tilde{\boldsymbol{h}}^{i}\right) \\ &= \left(\boldsymbol{I} - \alpha\tilde{\boldsymbol{Q}}^{i}\left(\boldsymbol{U}_{i}\boldsymbol{V}\boldsymbol{U}_{i}^{T}\right)\right)^{-1}\left(\boldsymbol{U}_{i}\boldsymbol{m} - \alpha\left(\boldsymbol{U}_{i}\boldsymbol{V}\boldsymbol{U}_{i}^{T}\right)\tilde{\boldsymbol{h}}^{i}\right). \end{aligned}$$

Therefore, the computational bottleneck of EP reduces to the computation of the quantities $U_i m$ and $U_i V U_i^T$. These can be computed from the canonical representation of q by $U_i \tilde{Q}^{-1} h$ and $U_i \tilde{Q}^{-1} U_i^T$.

C.3 Computing EP's Evidence Approximation

Let us define

$$\log Z(\boldsymbol{m}, \boldsymbol{V}) \equiv \frac{1}{2} \boldsymbol{m}^T \boldsymbol{V}^{-1} \boldsymbol{m} + \frac{1}{2} \log \det \boldsymbol{V} + \frac{n}{2} \log (2\pi)$$

and

$$\log Z_i(\boldsymbol{m}, \boldsymbol{V}) \equiv \log \int d\boldsymbol{x} N(\boldsymbol{x} | \boldsymbol{m}, \boldsymbol{V}) t_i^{\alpha}(\boldsymbol{U}_i \boldsymbol{x}).$$

Expectation propagation approximates the evidence $p(y|\theta)$ by $Z_{ep} = Z^{1-n/\alpha} \prod_i Z_i^{\alpha}$. Using the above introduced notation this can be written as

$$\begin{split} \log Z_{EP} = & \log Z(\boldsymbol{m}, \boldsymbol{V}) \\ & + \frac{1}{\alpha} \sum_{i} \left[\log Z_{j} \left(\boldsymbol{m}^{\setminus i}, \boldsymbol{V}^{\setminus i} \right) + \log Z \left(\boldsymbol{m}^{\setminus i}, \boldsymbol{V}^{\setminus i} \right) - \log Z(\boldsymbol{m}, \boldsymbol{V}) \right], \end{split}$$

which in the case when t_i depends on $U_i x$ leads to

$$\begin{split} \log Z_{EP} = & \log Z(\boldsymbol{m}, \boldsymbol{V}) + \frac{1}{\alpha} \sum_{i} \log Z_{j} \left(\boldsymbol{U}_{i} \boldsymbol{m}^{\setminus i}, \boldsymbol{U}_{i} \boldsymbol{V}^{\setminus i} \boldsymbol{U}_{i}^{T} \right) \\ & + \frac{1}{\alpha} \sum_{i} \left[\log Z \left(\boldsymbol{U}_{i} \boldsymbol{m}^{\setminus i}, \boldsymbol{U}_{i} \boldsymbol{V}^{\setminus i} \boldsymbol{U}_{i}^{T} \right) - \log Z \left(\boldsymbol{U}_{i} \boldsymbol{m}, \boldsymbol{U}_{i} \boldsymbol{V} \boldsymbol{U}_{i}^{T} \right) \right]. \end{split}$$

Appendix D. A Summary of the Marginal Approximations

An explanatory list of the approximation methods in Figure 13.

- LA-TK. The Laplace approximation of Tierney and Kadane (1986). The approximation $\tilde{p}^{\text{LA-TK}}(x_i)$ is computed by using the Laplace method to approximate $c_i(x_i)$ (Section 3.1).
- EP-FULL. The full EP approximation of the marginal. This approximation is computed by using EP to approximate $c_i(x_i)$ (Section 4.1.1).
- EP-L. EP local. The approximation $\tilde{p}^{\text{EP-L}}(x_i) \propto \varepsilon_i(x_i) q(x_i)$ is obtained from $c_{x_i}(x_j) \approx 1$, where $\varepsilon_i(x_i) = t_i(x_i)/\tilde{t}_i(x_i)$ and $q(\boldsymbol{x})$ are computed by EP (Section 3).
- LM-L. Lapace method local. EP local. The approximation $\tilde{p}^{\text{EP-L}}(x_i) \propto \varepsilon_i(x_i) q(x_i)$ is obtained from $c_{x_i}(x) \approx 1$, where $\varepsilon_i(x_i) = t_i(x_i)/\tilde{t}_i(x_i)$ and $q(\boldsymbol{x})$ are computed by the Laplace method (Section 3). In this case $\log \varepsilon_i(x_i) = R_2[\log t_i](x_i)$.
- LA-CM. The Laplace approximation with the conditional mode approximated by the conditional mean. The approximation $\tilde{p}^{\text{LA-CM}}(x_i)$ is computed as proposed in Rue et al. (2009), that is, by using the approximation $\boldsymbol{x}_{\backslash i}^*(x_i) \approx \mathrm{E}_q\left[\boldsymbol{x}_{\backslash i}|x_i\right]$ where $q(\boldsymbol{x})$ is given by the Laplace method (Section 4.1.2).
- LA-CM2. The similar approximation as LA-CM, but with an additional term added to account for $x_{\setminus i}^*(x_i) \approx \mathbb{E}_q \left[x_{\setminus i} | x_i \right]$ (Section 4.1.2).

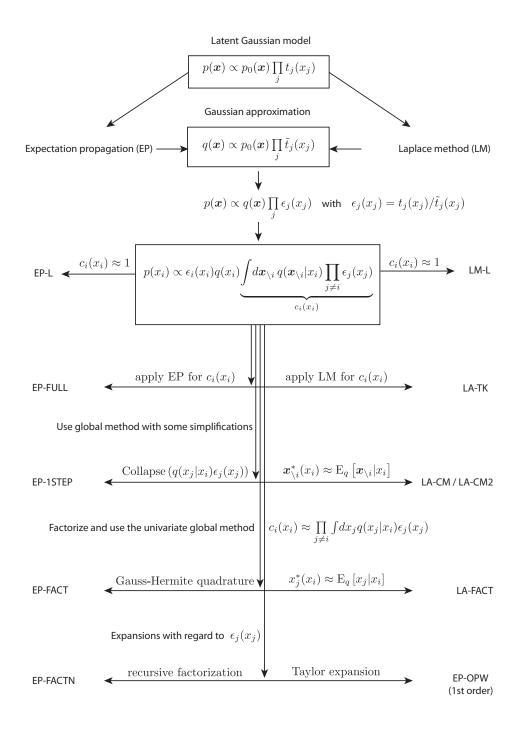


Figure 13: A schematic view of the approximation methods introduced or referred to in this paper. For details see Section D of the Appendix.

- EP-1STEP. The one step EP approximation. The approximation $\tilde{p}^{\text{EP-1STEP}}(x_i)$ is computed by defining $\tilde{\varepsilon}_j(x_j;x_i) \equiv \text{Collapse}(q(x_j|x_i)\varepsilon_j(x_j))/q(x_j|x_i)$ and using the approximation $c_i(x_i) \approx \int dx_{\setminus i} q(x_{\setminus i}|x_i) \prod_{j\neq i} \tilde{\varepsilon}_j(x_j;x_i)$ (see Section 4.1.1). This corresponds to one EP step for computing $c_i(x_i)$ with the initialization $\tilde{\varepsilon}_j(x_j;x_i) = 1$.
- EP-OPW. The Taylor expansion of Opper et al. (2009). The approximation $\tilde{p}^{\text{EP-OPW}}(x_i)$ is computed by expanding $p(x) \propto p_0(x) \prod_j \varepsilon_j(x_j)$ in first order with regard to $\varepsilon_j(x_j) 1$ for all $j = 1, \ldots, n$ and integrating with regard to $x_{\setminus i}$. When expanding only for $j \neq i$ the approximation is equal in first order to $\tilde{p}^{\text{EP-FACT}}(x_i)$ (Section 4.3).
- EP-FACT. The factorized EP approximation. The approximation $\tilde{p}^{\text{EP-FACT}}(x_i)$ is computed using the approximation $c_i(x_i) \approx \prod_{j \neq i} \int dx_j \, q(x_j|x_i) \varepsilon_j(x_j)$, where the univariate integrals are computed numerically or analytically, if it is the case. For further details see Section 4.2.
- LA-FACT. A similar approximation as EP-FACT, but here, the univariate integrals are computed with the Laplace method and using the approximation $x_j^*(x_i) \approx \mathbb{E}_q[x_j|x_i]$, with q(x) being the global approximation resulting from the Laplace method. For further details see Section 4.2.
- EP-FACTN. Higher order approximations obtained by using the factorization recursively. For further details see Section 4.2.

References

- P. R. Amestoy, T. A. Davis, and Iain S. D. An approximate minimum degree ordering algorithm. *SIAM Journal on Matrix Analysis and Applications.*, 17(4):886–905, October 1996.
- A. Birlutiu and T. Heskes. Expectation propagation for rating players in sports competitions. In Joost N. Kok, Jacek Koronacki, Ramon López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron, editors, *Proceedings ECML/PKDD*, volume 4702 of *Lecture Notes in Computer Science*, pages 374–381. Springer, 2007.
- L. Csató and M. Opper. Sparse representation for Gaussian process models. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, Cambridge, MA, USA, 2001. MIT Press.
- P. Dangauthier, R. Herbrich, T. Minka, and T. Graepel. Trueskill through time: Revisiting the history of chess. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems* 20, pages 337–344. MIT Press, Cambridge, MA, 2008.
- A. M. Erisman and W. F. Tinney. On computing certain elements of the inverse of a sparse matrix. *Communications of the ACM*, 18(3):177–179, 1975. ISSN 0001-0782.
- T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter. Approximate inference techniques with expectation constraints. *Journal of Statistical Mechanics: Theory and Experiment*, 2005: P11015, 2005.
- 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.
- S. Martino and H. Rue. Implementing approximate Bayesian inference using integrated nested Laplace approximation: a manual for the INLA program. Technical report, Department of Mathematical Sciences, NTNU, Norway, 2009.
- 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.
- K. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, volume 9, pages 467–475, San Francisco, USA, 1999. Morgan Kaufman.
- I. Murray, R. P. Adams, and D. J.C. MacKay. Elliptical slice sampling. In Y. W. Teh and M. Titterington, editors, *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 541–548. 2010.
- M. Opper and C. Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.
- M. Opper and O. Winther. Gaussian processes for classification: Mean-field algorithms. *Neural Computation*, 12(11):2655–2684, 2000.
- 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.
- K. Takahashi, J. Fagan, and M.-S. Chin. Formation of a sparse impedance matrix and its application to short circuit study. In *Proceedings of the 8th PICA Conference*, 1973.
- 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.
- M. van Gerven, B. Cseke, R. Oostenveld, and T. 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.

CSEKE AND HESKES

- M. van Gerven, B. Cseke, F. de Lange, and T. Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *Neuroimage*, 50(1):150–161, March 2010.
- O. Zoeter and T. Heskes. Gaussian quadrature based expectation propagation. In Z. Ghahramani and R. Cowell, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 445–452. Society for Artificial Intelligence and Statistics, 2005.