Expectation Propagation for Likelihoods Depending on an Inner Product of Two Multivariate Random Variables

Tomi Peltola
tomi.peltola@aalto.fi
Department of Biomedical Engineering and Computational Science
Aalto University, Finland

Pasi Jylänki
pasi.jylanki@ru.nl
Donders Institute for Brain, Cognition and Behaviour
Radboud University Nijmegen, The Netherlands

Aki Vehtari
aki.vehtari@aalto.fi
Department of Biomedical Engineering and Computational Science
Aalto University, Finland

Abstract

We describe how a deterministic Gaussian posterior approximation can be constructed using expectation propagation (EP) for models, where the likelihood function depends on an inner product of two multivariate random variables. The family of applicable models includes a wide variety of important linear latent variable models used in statistical machine learning, such as principal component and factor analysis, their linear extensions, and errors-in-variables regression. The EP computations are facilitated by an integral transformation of the Dirac delta function, which allows transforming the multidimensional integrals over the two multivariate random variables into an analytically tractable form up to one-dimensional analytically intractable integrals that can be efficiently computed numerically. We study the resulting posterior approximations in sparse principal component analysis with Gaussian and probit likelihoods. Comparisons to Gibbs sampling and variational inference are presented.

1 INTRODUCTION

Probability models that contain an inner product of two multivariate random variables are an essential building block of a wide variety of models in probabilistic data analysis and machine learning. Such models include linear latent variable models: principal component [1], factor and canonical correlation analysis [2], which form an important model family, for example, for analysis of high-dimensional data. Extensions of these models have also been applied for biclustering [3], imputation [4] and multi-task learning [5], among others. A perhaps underutilized family of models, linear regression models with uncertainty in the predictors (e.g., measurement error, misclassification or missingness) [6] fall also into this category. More generally, here we consider models, where a likelihood term for the ith observation of the jth variable y_{ij} can be written as $p(y_{ij}|\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i, \theta_{ij})$, where $\boldsymbol{w}_j$ and $\boldsymbol{x}_i$ are two multivariate random variables and θ_{ij} a possible further parameter of the model. The approach can be extended to cases, where y_{ij} is not observed, but is a parameter of the model.

A strength of linear models is their interpretability. Following this, Bayesian sparse versions and other extensions of linear latent variable models are currently a much researched topic as they are well suited for analysis of datasets with a limited number of samples but a large number of variables. A challenge to the application of these models is computation of the analytically intractable posterior distribution. Markov chain Monte Carlo (MCMC) sampling is often applied, but the convergence can be slow with problems caused by multimodality arising from symmetries in the model structure and the sparsity-promoting priors. Already assessing the convergence can be difficult. Furthermore, non-Gaussian observation models and non-conjugate priors for model parameters often require elaborate sampling algorithms.

Many variational Bayes (VB) approaches have been proposed to facilitate the inference with the linear latent variable models in the particular setting of this work (e.g., [2, 4, 7–9]). As an alternative to VB, expectation propagation (EP) [10, 11] has been found to pro-

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

vide accurate posterior approximations in sparse linear models [12–16] and in applications to non-Gaussian observation models [10, 17, 18]. Rattray et al. [19] have proposed a hybrid message passing framework (VB-EP), where a mean-field variational Bayes approximation is formed for the likelihood terms dependent on the inner products $\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i$ and EP is used to approximate sparsity-promoting spike and slab prior terms. The factorized mean-field assumption between $\boldsymbol{w}_j$ and $\boldsymbol{x}_i$ results in tractable and computationally convenient update formulas for the approximate posterior in the VB setting, but to our knowledge, no computationally efficient EP approximations have been proposed for the likelihood terms.

The main contribution of this work is to describe how a deterministic Gaussian posterior approximation can be constructed using EP for models, where the likelihood terms depend on the inner products $\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i$. The challenge in forming an efficient EP algorithm is that it requires computing analytically intractable multidimensional integrals over probability distributions that depend on $\boldsymbol{x}_i$ and $\boldsymbol{w}_j$. We utilize a transformation of the Dirac delta function to transform the integrals into one-dimensional problems that can be solved efficiently using numerical integration. The presented experiments demonstrate that EP can provide significantly more accurate estimates compared to VB in some cases. A transformation of the Dirac delta function was also recently applied by Challis and Barber [20] in variational inference for non-conjugate models, but for a different purpose.

This article is structured as follows. In Section 2, the EP algorithm as applied in this work is introduced. In Section 3, the details of the EP computation are then presented. We also briefly describe some essential implementation issues. In Section 4, the proposed approach is applied in sparse principal component analysis with Gaussian and probit likelihoods. Comparisons to a message passing algorithm, VB and MCMC are presented.

1.1 Notation

We write column vectors as bold face lower-case symbols and matrices as bold face upper-case symbols. $\boldsymbol{x}^{\mathrm{T}}$ is the transpose of $\boldsymbol{x}$. $\boldsymbol{I}$ is the identity matrix. $p(a|b)$ is an unspecified density or mass function for the random variable a given the parameter b. $\mathrm{N}(\boldsymbol{a}|\boldsymbol{b}, \boldsymbol{C}^{-1}) = |2\pi\boldsymbol{C}^{-1}|^{-\frac{1}{2}} \exp(-\frac{1}{2}(\boldsymbol{a}-\boldsymbol{b})^{\mathrm{T}}\boldsymbol{C}(\boldsymbol{a}-\boldsymbol{b}))$ is a Gaussian density with the mean $\boldsymbol{b}$ and precision matrix $\boldsymbol{C}$, where $|\cdot|$ denotes the determinant. $t(\boldsymbol{a}|\boldsymbol{d}, \boldsymbol{C}) = \exp(-\frac{1}{2}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{a} + \boldsymbol{a}^{\mathrm{T}}\boldsymbol{d}) \propto \mathrm{N}(\boldsymbol{a}|\boldsymbol{C}^{-1}\boldsymbol{d}, \boldsymbol{C}^{-1})$ is an unnormalized Gaussian density for $\boldsymbol{a}$ with the precision-adjusted mean $\boldsymbol{d}$ and precision matrix $\boldsymbol{C}$. $\imath$ is the imaginary unit.

2 EXPECTATION PROPAGATION

This section presents a suitable approximating family for the linear latent variable models together with a general EP algorithm for determining the parameters of the approximation.

2.1 Form of the Models

The posterior distribution for the models considered here can be written as

$$p(\boldsymbol{w}, \boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) = \frac{1}{Z}\prod_j p(\boldsymbol{w}_j)\prod_i p(\boldsymbol{x}_i)\prod_{i,j} p(y_{ij}|\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i, \theta_j), \tag{1}$$

where $\boldsymbol{w}$ is the collection of K-dimensional coefficient vectors $\boldsymbol{w}_j$ with $j = 1, \ldots, m$, $\boldsymbol{x}$ is the collection of K-dimensional latent variable vectors $\boldsymbol{x}_i$, $i = 1, \ldots, n$, $\boldsymbol{y}$ is the collection of observations y_{ij} for n samples and m variables. $\boldsymbol{\theta}$ are possible parameters of the observation model (e.g., residual variance), which are for now assumed given. Z is the normalization constant.

2.2 Form of the Approximation

The EP approximation has approximate terms for each of the factors in Equation 1. For example, $p(\boldsymbol{w}_j)$ is replaced by the *site term* $t_j(\boldsymbol{w}_j|\tilde{\boldsymbol{\mu}}_{w,j}, \tilde{\boldsymbol{\Gamma}}_{w,j})$ in the approximation. The full EP approximation is written as

$$q(\boldsymbol{w}, \boldsymbol{x}) = \prod_j \mathrm{N}(\boldsymbol{w}_j|\boldsymbol{m}_{w,j}, \boldsymbol{\Gamma}_{w,j}^{-1})\prod_i \mathrm{N}(\boldsymbol{x}_i|\boldsymbol{m}_{x,i}, \boldsymbol{\Gamma}_{x,i}^{-1}),$$

where the factor for $\boldsymbol{w}_j$ decomposes as

$$\mathrm{N}(\boldsymbol{w}_j|\boldsymbol{m}_{w,j}, \boldsymbol{\Gamma}_{w,j}^{-1}) \propto t_j(\boldsymbol{w}_j|\tilde{\boldsymbol{\mu}}_{w,j}, \tilde{\boldsymbol{\Gamma}}_{w,j})\prod_i t_{ij}(\boldsymbol{w}_j|\tilde{\boldsymbol{\mu}}_{w,ij}, \tilde{\boldsymbol{\Gamma}}_{w,ij})$$

for $j = 1, \ldots, m$. Consequently,

$$\boldsymbol{\Gamma}_{w,j} = \tilde{\boldsymbol{\Gamma}}_{w,j} + \sum_i \tilde{\boldsymbol{\Gamma}}_{w,ij}$$

$$\boldsymbol{m}_{w,j} = \boldsymbol{\Gamma}_{w,j}^{-1}(\tilde{\boldsymbol{\mu}}_{w,j} + \sum_i \tilde{\boldsymbol{\mu}}_{w,ij}).$$

The equations for $\boldsymbol{x}_i$ are similar.

The EP algorithm is used to determine the parameters $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Gamma}}$ of the site terms and, thus, of the full approximation.

2.3 EP Algorithm

The EP algorithm iteratively refines the approximation by minimizing the Kullback–Leibler divergence between a *tilted distribution* and the approximation, which can be shown to correspond to matching their

moments for approximating distributions in the exponential family. The tilted distribution is formed by replacing a single site term of the approximation with the corresponding term in the posterior distribution. The EP algorithm cycles through the site term updates until convergence.

An update of a single site term can be divided into three steps: 1) removal of the site term from the approximation, which results in a *cavity distribution*, 2) computation of the moments of the tilted distribution (cavity times the corresponding factor of the posterior), and 3) update of the site term parameters. The individual steps are described below for the likelihood term updates, with details of the moment matching given in the next section. The overall EP scheme used in this work is a form of the parallel EP algorithm (see, e.g., [13]), where multiple site updates are done in parallel before updating the full approximation, and is given in Algorithm 1. Prior site updates follow the same pattern as the likelihood site updates (sites with the same form in the approximation and the posterior need not be updated, if initialized appropriately).

Algorithm 1: EP algorithm scheme

initialize site parameters;
compute full approximation;
repeat
  for $i \leftarrow 1$ to n and $j \leftarrow 1$ to m **do**
    update ijth likelihood site term:
    1) compute cavity;
    2) compute tilted distribution moments;
    3) update site parameters;
  end
  compute full approximation;
  update prior site terms;
  compute full approximation;
until *convergence*;

1) Cavity distribution The cavity distribution $q^{\setminus ij}$ for the likelihood term $p(y_{ij}|\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i, \theta_j)$ is formed as

$$q^{\setminus ij}(\boldsymbol{w}_j, \boldsymbol{x}_i) \propto q(\boldsymbol{w}_j, \boldsymbol{x}_i) t_{ij}(\boldsymbol{w}_j)^{-1} t_{ij}(\boldsymbol{x}_i)^{-1}.$$

As the full approximation factorizes between $\boldsymbol{w}_j$ and $\boldsymbol{x}_i$, so does the cavity distribution. Focusing on $\boldsymbol{w}_j$, the cavity $q^{\setminus ij}(\boldsymbol{w}_j) = \mathrm{N}(\boldsymbol{w}_j | \boldsymbol{m}_{w,ij}^{\setminus ij}, (\boldsymbol{\Gamma}_{w,ij}^{\setminus ij})^{-1})$ as it is formed as a division of two unnormalized Gaussian densities. The parameters can be identified as

$$\begin{aligned} \boldsymbol{\Gamma}_{w,ij}^{\setminus ij} &= \boldsymbol{\Gamma}_{w,j} - \tilde{\boldsymbol{\Gamma}}_{w,ij} \\ \boldsymbol{m}_{w,ij}^{\setminus ij} &= (\boldsymbol{\Gamma}_{w,ij}^{\setminus ij})^{-1} (\boldsymbol{\Gamma}_{w,j} \boldsymbol{m}_{w,j} - \tilde{\boldsymbol{\mu}}_{w,ij}). \end{aligned}$$

Form and parameters of $q^{\setminus ij}(\boldsymbol{x}_i)$ are found similarly. The joint cavity $q^{\setminus ij}(\boldsymbol{w}_j, \boldsymbol{x}_i)$ is thus a Gaussian distribution.

2) Moments of the tilted distribution The tilted distribution is

$$\hat{p}(\boldsymbol{w}_j, \boldsymbol{x}_i) \propto p(y_{ij}|\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i, \theta_j) q^{\setminus ij}(\boldsymbol{w}_j, \boldsymbol{x}_i).$$

Computation of the moments, that is the means and the covariance matrices

$$\begin{aligned} \hat{\boldsymbol{m}}_{w,ij} &= \mathrm{E}_{\hat{p}}[\boldsymbol{w}_j] \\ \hat{\boldsymbol{m}}_{x,ij} &= \mathrm{E}_{\hat{p}}[\boldsymbol{x}_i] \\ \hat{\boldsymbol{\Gamma}}_{w,ij}^{-1} &= \mathrm{E}_{\hat{p}}[(\boldsymbol{w}_j - \hat{\boldsymbol{m}}_{w,ij})(\boldsymbol{w}_j - \hat{\boldsymbol{m}}_{w,ij})^{\mathrm{T}}] \\ \hat{\boldsymbol{\Gamma}}_{x,ij}^{-1} &= \mathrm{E}_{\hat{p}}[(\boldsymbol{x}_i - \hat{\boldsymbol{m}}_{x,ij})(\boldsymbol{x}_i - \hat{\boldsymbol{m}}_{x,ij})^{\mathrm{T}}] \end{aligned} \tag{2}$$

required for the EP approximation update seems analytically intractable even for a Gaussian likelihood. We show in the next section how the required $2K$-dimensional integrals can be evaluated using only one-dimensional numerical integrals.

3) Site parameter updates The site parameters of the site terms corresponding to the likelihood term are updated such that the moments of the approximation and the tilted distribution match. This can be seen as a reversion of the cavity computation (now with the approximation q having the moments of the tilted distribution) and leads to the updates:

$$\begin{aligned} \tilde{\boldsymbol{\Gamma}}_{w,ij} &= \hat{\boldsymbol{\Gamma}}_{w,ij} - \boldsymbol{\Gamma}_{w,ij}^{\setminus ij} \\ \tilde{\boldsymbol{\mu}}_{w,ij} &= \hat{\boldsymbol{\Gamma}}_{w,ij} \hat{\boldsymbol{m}}_{w,ij} - \boldsymbol{\Gamma}_{w,ij}^{\setminus ij} \boldsymbol{m}_{w,ij}^{\setminus ij} \end{aligned}$$

and similarly for the $\boldsymbol{x}_i$ part.

We use two stabilizing procedures in the site parameter updates. First, the site precision matrix updates are restricted to produce positive definite matrices (see [10] for discussion of site precision restrictions). This is effected, when needed, by modifying the tilted distribution precision matrix such that the eigenvalues of the new site precision matrix remain positive, and carrying out the site update with this modified precision matrix so that the exact mean of the tilted distribution is preserved in the update. Second, the parameter updates are damped, that is, the parameters are updated to a convex combination of the old and new value [21].

3 MOMENTS OF TILTED DISTRIBUTIONS DEPENDING ON AN INNER PRODUCT OF RANDOM VARIABLES

Computation of the tilted distribution moments in Equation 2 seems intractable already for a Gaussian likelihood $p(y_{ij}|\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i, \theta_j)$, because of the inner product. Our proposal is to use an integral transformation of the Dirac delta function $\delta(\xi) = \frac{1}{2\pi} \int \exp(\imath t \xi) dt$

[22, p. 37–38] to rewrite the integrals. For example, for the normalization of the tilted distribution (where $\boldsymbol{\phi} = (\boldsymbol{w}_j, \boldsymbol{x}_i)$ to shorten the notation):

$$\begin{aligned}
&\int p(y_{ij}|\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i, \theta_j) q^{\setminus ij}(\boldsymbol{\phi}) d\boldsymbol{\phi} \\
&= \iint p(y_{ij}|f, \theta_j)\delta(f - \boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i) df q^{\setminus ij}(\boldsymbol{\phi}) d\boldsymbol{\phi} \\
&= \iint p(y_{ij}|f, \theta_j) \frac{1}{2\pi} \int \exp(\imath t(f - \boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i)) dt df q^{\setminus ij}(\boldsymbol{\phi}) d\boldsymbol{\phi} \\
&= \frac{1}{2\pi} \iint p(y_{ij}|f, \theta_j) \exp(\imath t f) df \\
&\qquad \times \int \exp(-\imath t \boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i)) q^{\setminus ij}(\boldsymbol{\phi}) d\boldsymbol{\phi} dt \\
&= \frac{1}{2\pi} \iint L(t, f) df \int C(t, \boldsymbol{\phi}) d\boldsymbol{\phi} dt,
\end{aligned} \tag{3}$$

where $L(t, f)$ is the integrand over f and $C(t, \boldsymbol{\phi}) = C(t, \boldsymbol{w}_j, \boldsymbol{x}_i)$ is the integrand over $(\boldsymbol{w}_j, \boldsymbol{x}_i)$. We note that the change of integration order is not always valid and deal with such a case below for probit likelihood. Extended derivations of Equation 3 and the following moment integrals are given in the supplementary material.

On studying the $2K$-dimensional integral $\int C(t, \boldsymbol{w}_j, \boldsymbol{x}_i) d(\boldsymbol{w}_j, \boldsymbol{x}_i)$ the integrand can be seen to be of unnormalized Gaussian form, with complex-valued mean and covariance. In particular, the mean and covariance for the concatenated variable $[\boldsymbol{w}_j^{\mathrm{T}} \ \boldsymbol{x}_i^{\mathrm{T}}]^{\mathrm{T}}$ are

$$\begin{bmatrix} \bar{\boldsymbol{m}}_w \\ \bar{\boldsymbol{m}}_x \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Gamma}_{x,ij}^{\setminus ij} \boldsymbol{\Gamma}_{w,ij}^{\setminus ij} \boldsymbol{\Sigma}^{\mathrm{T}} (\boldsymbol{m}_{w,ij}^{\setminus ij} - \imath t (\boldsymbol{\Gamma}_{w,ij}^{\setminus ij})^{-1} \boldsymbol{m}_{x,ij}^{\setminus ij}) \\ \boldsymbol{\Gamma}_{w,ij}^{\setminus ij} \boldsymbol{\Gamma}_{x,ij}^{\setminus ij} \boldsymbol{\Sigma} (\boldsymbol{m}_{x,ij}^{\setminus ij} - \imath t (\boldsymbol{\Gamma}_{x,ij}^{\setminus ij})^{-1} \boldsymbol{m}_{w,ij}^{\setminus ij}) \end{bmatrix}, \tag{4}$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Gamma}_{w,ij}^{\setminus ij} \boldsymbol{\Gamma}_{x,ij}^{\setminus ij} + t^2 \boldsymbol{I})^{-1}$, and

$$\bar{\boldsymbol{V}} = \begin{bmatrix} \boldsymbol{\Gamma}_{w,ij}^{\setminus ij} & \imath t \boldsymbol{I} \\ \imath t \boldsymbol{I} & \boldsymbol{\Gamma}_{x,ij}^{\setminus ij} \end{bmatrix}^{-1}. \tag{5}$$

While not perhaps immediately clear, such Gaussian integral can be computed analytically when the real part of $\bar{\boldsymbol{V}}$ is positive definite and behaves similarly to the common real-valued version [23, p. 10]. With this result, the normalization constant of the Gaussian form $C(t, \boldsymbol{w}_j, \boldsymbol{x}_i)$ can be seen to be

$$D(t) = (|\boldsymbol{\Gamma}_{w,ij}^{\setminus ij}||\boldsymbol{\Gamma}_{x,ij}^{\setminus ij}||\bar{\boldsymbol{V}}|)^{\frac{1}{2}} \exp(-\frac{1}{2} d(t)),$$

where

$$\begin{aligned}
d(t) = t^2 &((\boldsymbol{m}_{w,ij}^{\setminus ij})^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\Gamma}_{w,ij}^{\setminus ij} \boldsymbol{m}_{w,ij}^{\setminus ij} \\
&+ (\boldsymbol{m}_{x,ij}^{\setminus ij})^{\mathrm{T}} \boldsymbol{\Sigma}^{\mathrm{T}} \boldsymbol{\Gamma}_{x,ij}^{\setminus ij} \boldsymbol{m}_{x,ij}^{\setminus ij}) \\
&+ 2\imath t (\boldsymbol{m}_{w,ij}^{\setminus ij})^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{\Gamma}_{w,ij}^{\setminus ij} \boldsymbol{\Gamma}_{x,ij}^{\setminus ij} \boldsymbol{m}_{x,ij}^{\setminus ij},
\end{aligned}$$

and the mean and covariance are given in Equations 4 and 5. Note that $\bar{\boldsymbol{m}}_w$, $\bar{\boldsymbol{m}}_x$, $\bar{\boldsymbol{V}}$ and $\boldsymbol{\Sigma}$ are functions of t, although we don't explicitly write the dependency.

Assuming $L(t, f)$ can be integrated over f either analytically or numerically, we are left with integration over t in Equation 3, which seems analytically intractable. However, since the problem has been reduced from $2K$-dimensional integration to only one-dimensional, it can now be efficiently implemented numerically. Computing the normalization constant $\hat{Z} = \frac{1}{2\pi} \int L(t) D(t) dt$ of the tilted distribution requires evaluation of one one-dimensional integral. The mean $\hat{\boldsymbol{m}}_{w,ij}$ is evaluated as $\frac{1}{2\pi} \frac{1}{\hat{Z}} \int \bar{\boldsymbol{m}}_w L(t) D(t) dt$ and requires K one-dimensional integrals. $\hat{\boldsymbol{m}}_{x,ij}$ is computed similarly. The covariance $\hat{\boldsymbol{\Gamma}}_{w,ij}^{-1}$ can be evaluated as $\frac{1}{2\pi} \frac{1}{\hat{Z}} \int (\bar{\boldsymbol{m}}_w \bar{\boldsymbol{m}}_w^{\mathrm{T}} + \boldsymbol{\Gamma}_{x,ij}^{\setminus ij} \boldsymbol{\Sigma}) L(t) D(t) dt - \hat{\boldsymbol{m}}_{w,ij} \hat{\boldsymbol{m}}_{w,ij}^{\mathrm{T}}$ and requires $\frac{K(K+1)}{2}$ one-dimensional integrals. $\hat{\boldsymbol{\Gamma}}_{x,ij}^{-1}$ is computed similarly.

In total, one iteration of the EP algorithm requires evaluating $O(nmK^2)$ one-dimensional numerical integrals. To decrease the computational burden, one can consider restricting the site precision matrices $\tilde{\boldsymbol{\Gamma}}_{w,ij}$ and $\tilde{\boldsymbol{\Gamma}}_{x,ij}$ to diagonal. Only the diagonal elements of the tilted distribution covariance matrices would then be needed for the site updates, which would require in total only $2K$ one-dimensional integrals for the covariances in one likelihood site update and $O(nmK)$ integrals for one EP iteration. However, we did not test this in the current work.

3.1 Gaussian Likelihood

For the Gaussian likelihood $p(y_{ij}|f) = \mathrm{N}(y_{ij}|f, \theta_j)$, $L(t, f) = \mathrm{N}(y_{ij}|f, \theta_j) \exp(\imath t f)$ and integration over f gives $L(t) \propto \exp(y_{ij} \imath t - \frac{1}{2} \theta_j t^2)$.

3.2 Probit Likelihood

For the probit likelihood $p(y_{ij}|f) = \Phi(y_{ij} f)$, where Φ is the cumulative distribution function of the standard normal distribution and $y_{ij} \in \{-1, +1\}$, the integration of $L(t, f) = \Phi(y_{ij} f) \exp(\imath t f)$ over f diverges. Our proposed solution is to add the term $\exp(s(\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i))$ to the tilted distribution and take the part $\exp(-sf)$ to the $L(t, f)$ integral, leaving $\exp(s\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i)$ to $C(t, \boldsymbol{w}_j, \boldsymbol{x}_i)$ when using the Dirac delta trick. Choosing $s > 0$ for $y_{ij} = +1$ and $s < 0$ for $y_{ij} = -1$ makes the integral of $L(t, f)$ over f convergent, giving $L(t) \propto \frac{y_{ij}}{s - \imath t} \exp(-\frac{1}{2}(t^2 + 2st\imath))$. Additionally, s needs to be selected such that the real part of $\bar{\boldsymbol{V}}$ is positive definite for the integration of $C(t, \boldsymbol{w}_j, \boldsymbol{x}_i)$ over $(\boldsymbol{w}_j, \boldsymbol{x}_i)$.

3.3 Other Likelihoods

Generally, a likelihood function $p(y_{ij}|f,\theta)$ may require numerical integration of the corresponding $L(t,f)$ over f. Furthermore, θ can also be assumed unknown and given a prior distribution. The EP algorithm can then be extended to approximate the posterior of θ (see [24]). In this case $L = L(t,f,\theta)$ depends also on θ and the moments of the tilted distribution with regard to θ need to be computed. In general, this requires numerical integration, but in some cases a convenient representation of $L_l(t) = \int \theta^l L(t,f,\theta) df d\theta$, for $l = 0,1,2$, may be available.

3.4 Implementation Details

We note some implementation details here. Firstly, as the determinant $|\bar{\boldsymbol{V}}| = |\boldsymbol{\Sigma}^{-1}|^{-1}$ is a function of t, we need to be able to compute it efficiently for various values of t. The eigendecomposition of $\boldsymbol{\Gamma}_{w,ij}^{\backslash ij}\boldsymbol{\Gamma}_{x,ij}^{\backslash ij}$ (the eigenvalues and -vectors of which can be shown to be real-valued and the eigenvalues positive, when $\boldsymbol{\Gamma}_{w,ij}^{\backslash ij}$ and $\boldsymbol{\Gamma}_{x,ij}^{\backslash ij}$ are positive definitive) can be used for this. The value of the determinant at t is the inverse of the product of the eigenvalues λ_l, which have been shifted by t^2: $\prod_l (\lambda_l + t^2)^{-1}$. The eigendecomposition can also be used in other computations involving $\boldsymbol{\Sigma}$.

Secondly, the real part of the integrand in the tilted distribution moments can be seen to be an even function of t. Thus, the numerical integration can be performed from 0 to ∞ instead of the full real axis. Similarly, the imaginary part can be seen to be an odd function of t and will always vanish.

We implement the numerical integration over t using Simpson's composite rule. Some care is needed in the selection of the number of evaluation points as the integrand can be oscillatory. Writing the integrand for the computation of the normalization constant $\hat{Z}$ as $\exp(a(t) + \imath b(t))$, one can see using the Euler's formula that $a(t)$ defines the decay of the integrand as t increases and $b(t)$ defines the oscillatory behavior. We use second order Taylor expansions of $a(t)$ and $b(t)$ around $t = 0$ to efficiently determine a suitable end point of the integration and the number of required evaluation points.

4 EXPERIMENTS

In this section, we study the accuracy of the EP approximation in sparse principal component analysis (SPCA) models with Gaussian and probit likelihoods and compare to alternative inference methods.

4.1 Sparse PCA – Gaussian Likelihood

The SPCA model with Gaussian likelihood is

$$\begin{aligned} p(y_{ij}|\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i) &= \mathrm{N}(y_{ij}|\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i, 1) \\ p(x_{ik}) &= \mathrm{N}(x_{ik}|0,1) \\ p(w_{jk}|\tau^2,\gamma_{jk}) &= \mathrm{N}(w_{jk}|0,\tau^2)^{\gamma_{jk}}\delta(w_{jk})^{1-\gamma_{jk}} \\ p(\gamma_{jk}|\omega) &= \mathrm{Bernoulli}(\gamma_{jk}|\omega), \end{aligned} \tag{6}$$

where $i = 1,\ldots,n$, $j = 1,\ldots,m$, $k = 1,\ldots,K$ for n samples, m observed variables, and K latent variables. The binary parameter γ_{jk} indicates whether w_{jk} is allowed to be non-zero, with the prior probability ω governing the sparsity. For simplicity, the residual variance and the prior parameters τ^2 and ω are here assumed given.

Following Sharp and Rattray [25] we generate 50 replicate datasets from the SPCA model in each of the four configurations: $n = 200$, $K = 1$, $\omega = 0.1$, $m = 800, 1000, 1333, 2000$ with $\tau^2 = 0.125, 0.1, 0.075, 0.05$ respectively. Data generation was done using the Matlab code of Sharp and Rattray [25][1].

We compare four inference methods for the model: 1) EP as proposed here, 2) VB-EP hybrid [19], where the likelihood terms are updated using a mean-field variational Bayes algorithm and the prior using EP, 3) dense message passing (DMP) [25] and 4) collapsed Gibbs sampling [19]. EP updates for the sparse prior of $\boldsymbol{w}_j$ can be found in [15, 19]. Matlab code of Sharp and Rattray [25] was used for DMP[1]. Gibbs sampling was run for 10000 iterations, of which 1000 were discarded as burn-in. It was initialized using the VB-EP hybrid. PCA was used to initialize the other methods. Each computation method was run for seven different settings of the prior parameter ω (while τ^2 was set to its true value).

Assuming Gibbs sampling provides the best characterization of the posterior distribution, the deterministic approximations are compared against it using three statistics: mean squared error (w.r.t. the Gibbs result) in the posterior mean of $\boldsymbol{w}$ and $\boldsymbol{x}$, denoted MSE($\boldsymbol{w}$) and MSE($\boldsymbol{x}$), respectively, and mean absolute error in the posterior probabilities $p(\boldsymbol{\gamma} = 1)$, MAE($p(\boldsymbol{\gamma} = 1)$). We also compare the area under the ROC-curve (AUC) for the identification of true non-zero w_j and the cosine angle ρ between the true, data generating $\boldsymbol{w}$ and its posterior mean estimate. The latter is suggested by Rattray et al. [19], Sharp and Rattray [25] and they also provide formulas and code[1] for computing the theoretical optimal performance with regard to it. Robust statistics (median, quartiles) are used to summarize the results over the 50 replicate datasets to diminish the effect of possible occasional poor convergence.

[1] Available at `http://www.cs.man.ac.uk/~sharpk/`.

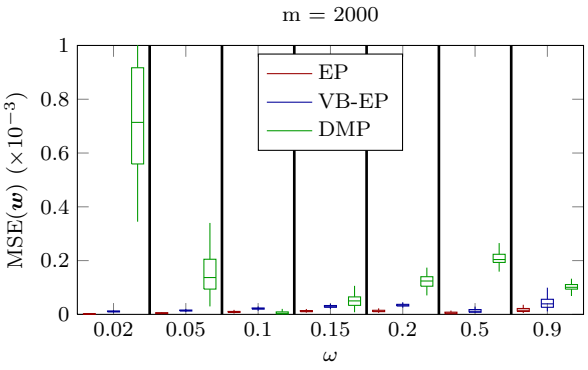

Figure 1: Boxplot of MSE($\boldsymbol{w}$) over the 50 replicate datasets with $m = 2000$ for Gaussian SPCA.

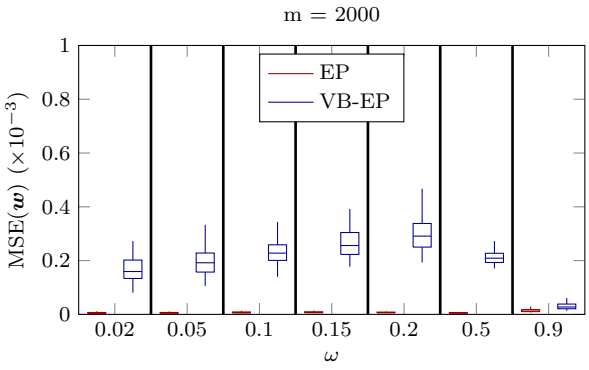

Figure 2: Boxplot of MSE($\boldsymbol{w}$) over the 50 replicate datasets with $m = 2000$ for probit SPCA.

Table 1: Performance statistics over the 50 replicate datasets for Gaussian SPCA for $m = 2000$ and $\omega = 0.1$. IQR is the inter-quartile range. DMP gives only posterior mean estimate for $\boldsymbol{w}$ and statistics depending on other parameters are not available for it.

	median	IQR
MSE($\boldsymbol{w}$) ($\times 10^{-4}$)		
EP	0.09	0.07 – 0.11
VB-EP	0.22	0.19 – 0.24
DMP	0.04	0.02 – 0.09
MSE($\boldsymbol{x}$) ($\times 10^{-2}$)		
EP	0.66	0.51 – 0.80
VB-EP	1.21	1.04 – 1.42
MAE($p(\boldsymbol{\gamma} = 1)$) ($\times 10^{-2}$)		
EP	0.40	0.34 – 0.47
VB-EP	0.95	0.86 – 1.10
AUC		
Gibbs	0.80	0.79 – 0.81
EP	0.80	0.79 – 0.81
VB-EP	0.80	0.79 – 0.81
ρ		
Gibbs	0.87	0.86 – 0.89
EP	0.87	0.86 – 0.89
VB-EP	0.87	0.86 – 0.89
DMP	0.87	0.86 – 0.89

Figure 1 shows the MSE($\boldsymbol{w}$) for $m = 2000$ as a function of the prior parameter ω and Table 1 lists the performance statistics for $m = 2000$ and $\omega = 0.1$. The errors in mean $\boldsymbol{w}$ are practically small, except for DMP when ω is not set to the data-generating value. This is because DMP strongly constrains the length of $\boldsymbol{w}$ in the algorithm according to the prior. The differences between EP and VB-EP are small, but EP agrees better with Gibbs sampling in this case. The same is true for the MSE($\boldsymbol{x}$) and MAE($p(\boldsymbol{\gamma}) = 1$) values presented in Table 1. There are no differences between the methods with regard to AUC or ρ. The theoretically optimal performance with regard to ρ is 0.88, which confirms that all of the methods perform well in this respect. Similar conclusions hold for all data-generating values of m. Boxplots of the statistics in each simulation setting are shown in the supplementary material.

4.2 Sparse PCA – Probit Likelihood

The above analysis was replicated using the same model (Equation 6), but with the probit likelihood:

$$p(y_{ij}|\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i) = \Phi(y_{ij}\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{x}_i),$$

where $y_{ij} \in \{-1, +1\}$. The same datasets were also used, but the observations were made binary-valued by a cutoff at zero. Three methods are compared: 1) EP, 2) VB-EP hybrid and 3) collapsed Gibbs sampling. The latter two were adapted for probit likelihood using the auxiliary variable representation of probit [26].

Figure 2 summarizes the MSE($\boldsymbol{w}$) for $m = 2000$ as a function of the prior parameter ω and Table 2 lists the performance statistics for $m = 2000$ and $\omega = 0.1$. The differences between the Gibbs sampling result and the deterministic approximations are now mostly larger than in the Gaussian likelihood case. EP agrees better with Gibbs with regard to MSE($\boldsymbol{w}$), MSE($\boldsymbol{x}$) and MAE($p(\boldsymbol{\gamma} = 1)$). AUC and ρ show little difference between the methods. Similar conclusions hold generally for all the simulation settings (boxplots of the performance statistics are presented in the supplementary material).

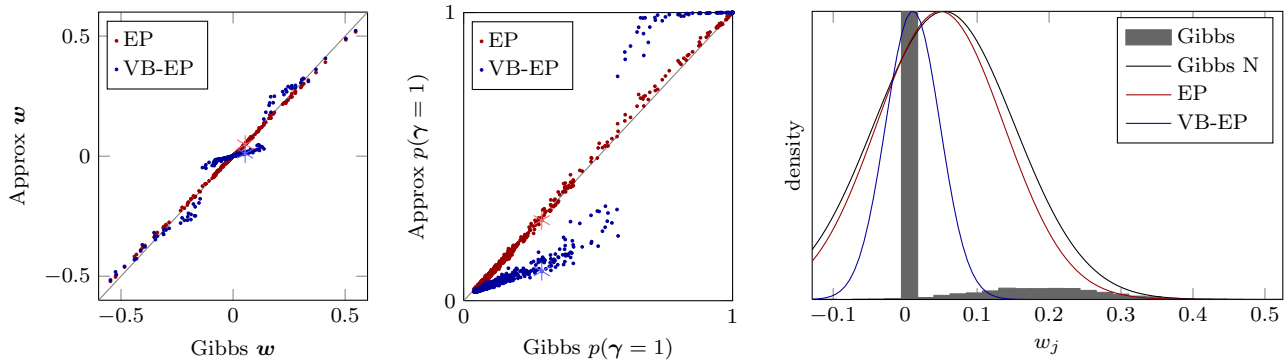

Figure 3: Left panel: Posterior mean of $\boldsymbol{w}$ for the Gibbs sampling (x-axis) and EP and VB-EP (y-axis) for a dataset with $m = 2000$ and the prior parameter ω set at 0.1. Middle panel: similar to left panel but showing posterior probabilities $p(\boldsymbol{\gamma} = 1)$. Right panel: Marginal posterior histogram and densities of w_j for the variable marked with star in the other panels. Histogram is based on the Gibbs samples, Gibbs N is a normal distribution with mean and variance taken from the Gibbs samples. Densities are scaled to have the same maximum value.

Table 2: Performance statistics over the 50 replicate datasets for probit SPCA for $m = 2000$ and $\omega = 0.1$.

	median	IQR
MSE($\boldsymbol{w}$) ($\times 10^{-4}$)		
EP	0.07	0.05 – 0.09
VB-EP	2.28	2.01 – 2.59
MSE($\boldsymbol{x}$) ($\times 10^{-2}$)		
EP	0.91	0.79 – 1.11
VB-EP	2.00	1.39 – 2.63
MAE($p(\boldsymbol{\gamma} = 1)$) ($\times 10^{-2}$)		
EP	0.55	0.48 – 0.60
VB-EP	3.49	3.36 – 3.66
AUC		
Gibbs	0.75	0.73 – 0.77
EP	0.75	0.73 – 0.77
VB-EP	0.75	0.73 – 0.77
ρ		
Gibbs	0.77	0.73 – 0.80
EP	0.77	0.73 – 0.80
VB-EP	0.74	0.68 – 0.78

The difference between the Gibbs sampling and the deterministic approximations can now be seen to be also of more practical interest. A common case is shown in Figure 3 for one of the datasets with $m = 2000$. EP gives accurate posterior mean estimates of $\boldsymbol{w}$. However, VB-EP markedly underestimates w_j sizes for small-sized coefficients. EP is also accurate for the posterior probabilities $p(\boldsymbol{\gamma} = 1)$, while VB-EP clearly pushes the estimates towards extremes. The rightmost panel in Figure 3 shows the marginal posterior approximations of w_j for a variable with $p(\gamma_j = 1) = 0.29$ given by Gibbs sampling (0.28 and 0.10 by EP and VB-EP, respectively). The VB-EP posterior approximation is drawn close to the mode at zero, while EP covers better the whole posterior mass. This kind of difference in the behavior of EP and VB has been observed previously (see, e.g., [17, 18]).

We also tested how the inference methods perform, when the latent dimensionality K is misspecified in the model. We set $K = 5$ and $\omega = 0.02$ (i.e., the true sparsity with the misspecified latent dimensionality) and computed the posterior approximations for the same dataset as analyzed above. Table 3 shows the number of coefficients with posterior probability $p(\gamma_{jk} = 1) > 0.05$ (an arbitrary threshold[2]) for each inferred latent dimension k. All of the methods correctly find that only one latent dimension is active and shut off the other four dimensions.

We also reproduced the middle panel of Figure 3 using the first, active latent dimension of the approximations for the misspecified model. This is shown in Figure 4. The result is very similar to the one with the correctly specified model.

[2]The reason that the threshold produces a lower number in the first latent dimension for VB-EP than for Gibbs sampling and EP is that VB-EP underestimates posterior probabilities at the low end. However, VB-EP is not worse than EP in separating the true data-generating coefficients as evidenced by the presented AUC comparisons.

Table 3: Comparison between the inference methods of the number of coefficients with posterior probability $p(\gamma_{jk} = 1) > 0.05$ for each of the latent dimensions k for a dataset with $m = 2000$ and true latent dimensionality of 1 and true sparsity $\omega = 0.1$ (i.e., 200 non-zero data-generating coefficients w_{jk}). The parameters of the fitted model were $K = 5$ and $\omega = 0.02$.

k	Gibbs	EP	VB-EP
1	143	141	62
2	3	3	0
3	1	0	1
4	0	0	0
5	2	0	0

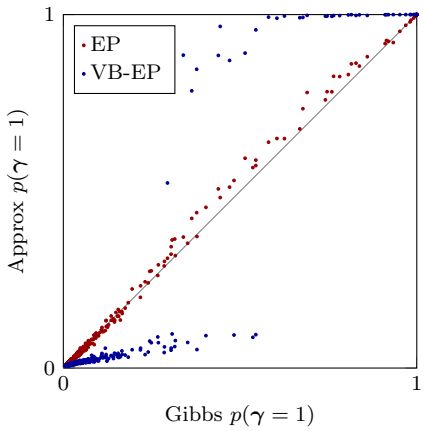

Figure 4: Posterior probabilities $p(\gamma = 1)$ for the first latent dimension for Gibbs sampling (x-axis) and EP and EP-VB (y-axis) for a dataset with $m = 2000$ and true latent dimensionality of 1 and true sparsity $\omega = 0.1$ (i.e., 200 non-zero data-generating coefficients w_{jk}). The parameters of the fitted model were $K = 5$ and $\omega = 0.02$.

5 CONCLUSIONS

In this work, we have shown how expectation propagation can be applied to Bayesian models, where the likelihood depends on an inner product of two multivariate random variables. This expands the applicability of EP to a wide variety of important linear models in statistical machine learning. The presented experiments show that the EP posterior approximation can be markedly better than a variational Bayes approximation in some cases. A trade-off is that the EP algorithm is slower and requires numerical integration. However, the algorithm is easily parallelizable.

The sparse principal component analysis is only one example of a model, where the proposed EP algorithm is applicable. An important future research direction is a more extensive and detailed characterization of the properties of the approximation in different types of prior models for the parameters $\boldsymbol{w}$ and $\boldsymbol{x}$ and as a building block in more complex probability models. Notably, the presented EP algorithm is applicable with any prior model as long as the posterior distributions of $\boldsymbol{w}$ and $\boldsymbol{x}$ can be approximated as Gaussian. The integrals over the auxiliary variable t and the EP update scheme should also be more carefully analyzed to optimize the speed and stability of the algorithm.

Matlab code implementing the SPCA model using EP, VB-EP hybrid and Gibbs sampling is available at `http://becs.aalto.fi/en/research/bayes/epwx/`.

Acknowledgements

This work was supported by the Finnish Doctoral Programme in Computation Sciences FICS (TP) and the Academy of Finland (grant 218248 to AV). We acknowledge the computational resources provided by Aalto Science-IT project. We thank Arno Solin, Ville Tolvanen and the anonymous reviewers for comments on the manuscript.

References

[1] Christopher M Bishop. Bayesian PCA. In *Advances in Neural Information Processing Systems*, volume 11, pages 382–388, 1999.

[2] Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013.

[3] Sepp Hochreiter, Ulrich Bodenhofer, Martin Heusel, Andreas Mayr, Andreas Mitterecker, Adetayo Kasim, Tatsiana Khamiakova, Suzy Van Sanden, Dan Lin, Willem Talloen, et al. FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527, 2010.

[4] Matthias Seeger and Guillaume Bouchard. Fast variational Bayesian inference for non-conjugate matrix factorization models. In *JMLR Workshop and Conference Proceedings: AISTATS 2012*, volume 22, pages 1012–1018, 2012.

[5] Piyush Rai and Hal Daumé III. Infinite predictor subspace models for multitask learning. In *JMLR Workshop and Conference Proceedings: AISTATS 2010*, volume 9, pages 613–620, 2010.

[6] Paul Gustafson. *Measurement Error and Misclassification in Statistical Epidemiology: Impacts and Bayesian Adjustments*. Interdisciplinary Statistics Series. Chapman & Hall/CRC, 2004.

[7] Christopher M Bishop. Variational principal components. In *Proceedings of the Ninth International Conference on Artificial Neural Networks, ICANN'99*, volume 1, pages 509–514, 1999.

[8] Zoubin Ghahramani and Matthew J Beal. Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems*, volume 12, pages 449–455, 2000.

[9] Hagai Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

[10] Thomas Minka. *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[11] Thomas P Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

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

[13] Marcel van Gerven, Botond Cseke, Robert Oostenveld, and Tom Heskes. Bayesian source localization with the multivariate Laplace prior. In *Advances in Neural Information Processing Systems*, volume 22, pages 1901–1909, 2009.

[14] Marcel van Gerven, Botond Cseke, Floris de Lange, and Tom Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50:150–161, 2010.

[15] Daniel Hernández-Lobato, José M Hernández-Lobato, and A. Suárez. Expectation propagation for microarray data classification. *Pattern Recognition Letters*, 31(12):1618–1626, 2010.

[16] Daniel Hernández-Lobato, José M Hernández-Lobato, and Pierre Dupont. Generalized spike-and-slab priors for Bayesian group feature selection using expectation propagation. *Journal of Machine Learning Research*, 14:1891–1945, 2013.

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

[18] Jaakko Riihimäki, Pasi Jylänki, and Aki Vehtari. Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *Journal of Machine Learning Research*, 14:75–109, 2013.

[19] Magnus Rattray, Oliver Stegle, Kevin Sharp, and John Winn. Inference algorithms and learning theory for Bayesian sparse factor analysis. *Journal of Physics: Conference Series*, 197(1), 2009.

[20] Edward Challis and David Barber. Affine independent variational inference. In *Advances in Neural Information Processing Systems*, volume 25, pages 2195–2203. 2012.

[21] Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

[22] Frank WJ Olver, Daniel W Lozier, Ronald F Boisvert, and Charles W Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[23] Yu A Neretin. *Lectures on Gaussian Integral Operators and Classical Groups*. European Mathematical Society, 2011.

[24] Pasi Jylänki, Aapo Nummenmaa, and Aki Vehtari. Expectation propagation for neural networks with sparsity-promoting priors. *arXiv preprint arXiv:1303.6938*, 2013.

[25] Kevin Sharp and Magnus Rattray. Dense message passing for sparse principal component analysis. In *JMLR Workshop and Conference Proceedings: AISTATS 2010*, volume 9, pages 725–732, 2010.

[26] Mark Girolami and Simon Rogers. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006.