EXPECTATION-PROPAGATION ALGORITHMS FOR LINEAR REGRESSION WITH POISSON NOISE: APPLICATION TO PHOTON-LIMITED SPECTRAL UNMIXING

Yoann Altmann[*]

School of Engineering and Physical Sciences
Heriot-Watt University
Riccarton, Edinburgh, U.K.

Alessandro Perelli[†], *Mike E. Davies*[‡]

School of Engineering
University of Edinburgh
King's Buildings, Edinburgh, U.K.

ABSTRACT

This paper discusses Expectation-Propagation (EP) methods for approximate Bayesian inference in the context of linear regression with Poisson noise. We review two main factor graphs used for generalized linear models and discuss how different EP algorithms can be derived. The estimation performance based on EP approximations is compared to the performance using Monte Carlo sampling from the exact posterior distribution. In particular, we observe that using locally independent or isotropic approximate factors enables more robust and scalable algorithms while providing reliable posterior means and marginal variances.

***Index Terms*—** Expectation-Propagation, Approximate Bayesian inference, linear regression, Poisson noise

1. INTRODUCTION

A popular methodology for approximate Bayesian inference consists of minimizing a Kullback-Leibler (KL) divergence, regarded as an asymmetric discrepancy measure between an exact yet complex distribution and an approximating distribution. Variational Bayes (VB) techniques aim at minimizing the direct KL divergence between the approximate and exact distributions [1]. Alternatively, minimizing the reverse KL divergence (between the approximate and the actual distributions) turns out to be a saddle point problem. While provably convergent double loop algorithms [2] minimizing the underlying Bethe free energy have been proposed, they are generally computationally intensive. Expectation-Propagation (EP) is a family of faster fixed point message passing solvers which minimize the reverse KL divergence locally [3]. Generally, when the EP algorithm converges, it gives a better estimate compared to VB (e.g., does not tend to underestimate posterior variances). Gaussian approximations are classically used within the EP framework [4], leading to more tractable KL divergence minimization problems. To reduce the number of sequential EP updates involving one-dimensional densities, it is possible to use, locally, multivariate Gaussian approximations. In particular, in high dimensional (e.g., imaging) problems, these factors can be constrained to have diagonal covariance matrices. Recently, a class of algorithms called Vector Approximate Message Passing (VAMP) [5] has been derived using EP with isotropic Gaussian messages over vectors, although the noise model can be non-Gaussian. While Poisson likelihoods have recently been considered within EP [6], here we investigate and compare several EP models in terms of estimation performance and uncertainty quantification. More precisely, we compare models involving data augmentation schemes and based on different Gaussian approximations, thus allowing more flexibility than VAMP. These models and associated *a posteriori* estimates are also compared, in terms of estimation performance and uncertainty quantification, to those obtained via Monte Carlo sampling from the true Bayesian model. To the best of our knowledge, this is the first time such a study is conducted for regression with Poisson noise.

The remainder of this paper is organized as follows. Section 2.1 introduces the exact Bayesian model used for linear regression. The two associated factor graphs and associated update rules are presented in Sections 2.2 and 2.3. Section 3 discusses simulations results obtained for a spectral unmixing application and conclusions are finally reported in Section 4.

2. EP FOR REGRESSION WITH POISSON NOISE

2.1. Exact Bayesian model

We consider the estimation of a vector $\boldsymbol{x} = [x_1, \ldots, x_N]^T$ from a set of noisy measurements $\boldsymbol{y} = [y_1, \ldots, y_M]^T$, with $M \geq N$. Precisely, conditioned on the value of $\boldsymbol{x}$, the entries of $\boldsymbol{y}$ are independently distributed according to Poisson distributions, i.e., $f(y_m|\boldsymbol{x}) = \left(\boldsymbol{a}_m^T\boldsymbol{x}\right)^{y_m} \exp\left[-\boldsymbol{a}_m^T\boldsymbol{x}\right]/y_m!$, where the vectors $\{\boldsymbol{a}_m\}_{m=1,\ldots,M}$ are positive and known.

[*] This work was supported by the Royal Academy of Engineering under the Research Fellowship scheme RF201617/16/31 and by the DSTL/EPSRC University Defence Research Collaboration (UDRC) award - Signal processing in the information age, EP/S000631/1.

[†] Since December 2018, A. P. is a H.C. rsted COFUND postdoc fellow at Technical University of Denmark (DTU).

[‡] M.D. acknowledges support for this work from ERC Advanced grant, C-SENSE, (ERC-ADG-2015-694888) and through the Royal Society Wolfson Research Merit Award.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

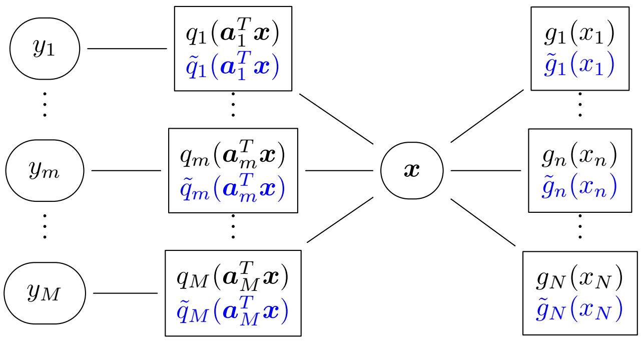
y_1
y_m
y_M
$q_1(\boldsymbol{a}_1^T\boldsymbol{x})$
$\tilde{q}_1(\boldsymbol{a}_1^T\boldsymbol{x})$
$q_m(\boldsymbol{a}_m^T\boldsymbol{x})$
$\tilde{q}_m(\boldsymbol{a}_m^T\boldsymbol{x})$
$q_M(\boldsymbol{a}_M^T\boldsymbol{x})$
$\tilde{q}_M(\boldsymbol{a}_M^T\boldsymbol{x})$
$\boldsymbol{x}$
$g_1(x_1)$
$\tilde{g}_1(x_1)$
$g_n(x_n)$
$\tilde{g}_n(x_n)$
$g_N(x_N)$
$\tilde{g}_N(x_N)$

Fig. 1. First factor graph (FG1) used to perform EP-based regression without auxiliary variables. The circles (resp. rectangular boxes) represent the variable (resp. factor) nodes and the approximate factors are shown in blue.

Assuming mutual independence between the elements of $\boldsymbol{y}$, conditioned on the value of $\boldsymbol{x}$, yields the following joint likelihood $f(\boldsymbol{y}|\boldsymbol{x}) = \prod_{m=1}^{M} f(y_m|\boldsymbol{x}) = \prod_{m=1}^{M} f(y_m|\boldsymbol{a}_m^T\boldsymbol{x})$. Since $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M]^T$ is known, it is omitted in all the conditional distributions. We assume that the prior model $p(\boldsymbol{x})$ is a product of N independent distributions, i.e., $g(\boldsymbol{x}) = \prod_{n=1}^{N} g_n(x_n)$, where $\{g_n(\cdot)\}_n$ are arbitrarily chosen and ensure the positivity of $\boldsymbol{x}$.

Using $f(\boldsymbol{y}|\boldsymbol{x})$ and $g(\boldsymbol{x})$, we can obtain the joint density

$$f(\boldsymbol{y}, \boldsymbol{x}) = f(\boldsymbol{y}|\boldsymbol{x})g(\boldsymbol{x}) = \prod_{m=1}^{M} f(y_m|\boldsymbol{a}_m^T\boldsymbol{x}) \prod_{n=1}^{N} g_n(x_n),$$

and the posterior distribution of $\boldsymbol{x}$ is given by $f(\boldsymbol{x}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{x})g(\boldsymbol{x})/f(\boldsymbol{y})$. Inferring $\boldsymbol{x}$ via maximum a posteriori (MAP) estimation is straightforward if $p(\boldsymbol{x})$ is log-concave since the problem reduces to a convex optimization problem [7, 8, 9]. However, such estimation strategy only provides point estimates and no *a posteriori* confidence measures about $\boldsymbol{x}$. In particular, here we concentrate on estimating the posterior mean and covariance (or marginal variances) of $f(\boldsymbol{x}|\boldsymbol{y})$. Unfortunately, it is in general not possible to compute these quantities analytically. A classical approach to exploit the posterior distribution consists of using simulation methods. In particular, constrained Hamiltonian Monte Carlo methods [10] have been investigated to solve regression problems in the presence of Poisson noise [11] (see also [12] for comparison of samplers including a bouncy particle sampler [13]). However, such methods still suffer from a high computational cost and approximate Bayesian methods (and in particular EP methods) stand as promising alternatives.

2.2. Existing EP approximation

The EP method, used for regression with Gaussian noise [4] and generalised linear models [14], approximates each exact factor $f(y_m|\boldsymbol{a}_m^T\boldsymbol{x}) = q_m(\boldsymbol{a}_m^T\boldsymbol{x})$ (resp. $g_n(\mathbf{x}_n)$) with a simpler factor $\tilde{q}_m(\boldsymbol{a}_m^T\boldsymbol{x})$ (resp. $\tilde{g}_n(x_n)$) so that

$$f(\boldsymbol{y}, \boldsymbol{x}) \approx \prod_{m=1}^{M} \tilde{q}_m(\boldsymbol{a}_m^T\boldsymbol{x}) \prod_{n=1}^{N} \tilde{g}_n(x_n) = Q(\boldsymbol{x}), \tag{1}$$

where all the approximate factors belong to the same family of distributions (Gaussian distributions here) and so does $Q(\boldsymbol{x})$, but they do not need to be normalized densities. In contrast to [15], here it does not seem possible (without auxiliary variables) to gather all the likelihood factors depicted in Fig. 1 (referred to as FG1) into a single factor, because of the positivity constraints imposed by the Poisson likelihood. To optimize $Q(\boldsymbol{x})$ so that Eq. (1) is satisfied, EP sequentially refines the factors $\left\{\tilde{q}_m(\boldsymbol{a}_m^T\boldsymbol{x})\right\}_m$ and $\{\tilde{g}_n(x_n)\}_n$ by minimizing the following Kullback-Leibler (KL) divergences

$$\begin{cases} \min\limits_{\tilde{q}_m} KL\left(q_m(\boldsymbol{a}_m^T\boldsymbol{x})Q^{\backslash m}(\boldsymbol{x})||\tilde{q}_m(\boldsymbol{a}_m^T\boldsymbol{x})Q^{\backslash m}(\boldsymbol{x})\right), \\ \min\limits_{\tilde{g}_n} KL\left(g_n(x_n)Q^{\backslash n}(\boldsymbol{x})||\tilde{g}_n(x_n)Q^{\backslash n}(\boldsymbol{x})\right), \end{cases} \tag{2}$$

where the cavity distributions $Q^{\backslash m}(\boldsymbol{x}) = Q(\boldsymbol{x})/\tilde{q}_m(\boldsymbol{a}_m^T\boldsymbol{x})$ and $Q^{\backslash n}(\boldsymbol{x}) = Q(\boldsymbol{x})/\tilde{g}_n(x_n)$ are Gaussian. Solving Eq. (2) thus reduces to matching the mean and covariance of $Q(\boldsymbol{x})$ and of the so-called tilted distributions $q_m(\boldsymbol{a}_m^T\boldsymbol{x})Q^{\backslash m}(\boldsymbol{x})$ (resp. $\tilde{g}_n(x_n)Q^{\backslash n}(\boldsymbol{x})$). In [6], the authors showed that the problem in the first row of Eq. (2) can be solved analytically by computing sequentially one-dimensional integrals. If $g_n(\cdot)$ is a truncated Gaussian or an exponential distribution, the second row of Eq. (2) can be solved by computing the mean and variance of a one-dimensional truncated Gaussian distribution. For more complex priors, Gaussian quadratures or Laplace approximations [16, 17] can be used. In a similar fashion to [15], we used a damping strategy here to reduce convergence issues (see Section 3). The resulting EP model is denoted EP-F (full covariance) in the remainder of the paper.

	EP-F	EP-DF	EP-DD	EP-ID	EP-II
$\boldsymbol{\Sigma}_1$	–	diag.	diag.	iso.	iso.
$\boldsymbol{S}_1$	–	full	diag.	diag.	iso.
$\boldsymbol{S}$	full	full	diag.	diag.	diag.

Table 1. Properties of $(\boldsymbol{\Sigma}_1, \boldsymbol{S}_1, \boldsymbol{S})$ for the different EP approximations. Using FG2, $\boldsymbol{\Sigma}_0$ and $\boldsymbol{S}_0$ are assumed diagonal.

2.3. EP using data augmentation

The scheme discussed in Section 2.2 induces $N + M$ sequential updates of the approximate factors at each iteration, which may lead to slow convergence if M is large. Here we discuss an alternative factor graph (FG2), depicted in Fig. 2 which is based on a data augmentation scheme. This scheme allows some updates to be performed independently and thus improve the convergence speed of the resulting EP-based methods. Let $\boldsymbol{u} = [u_1, \ldots, u_M]^T$ be a vector of auxiliary variables, the model in Eq. (1) can be extended as $f(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{x}) =$

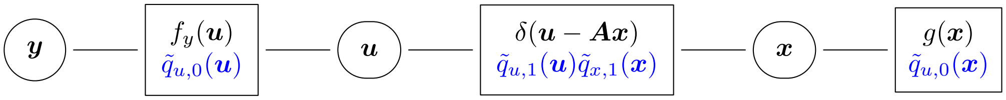
$\boldsymbol{y}$
$f_y(\boldsymbol{u})$
$\tilde{q}_{u,0}(\boldsymbol{u})$
$\boldsymbol{u}$
$\delta(\boldsymbol{u}-\boldsymbol{A}\boldsymbol{x})$
$\tilde{q}_{u,1}(\boldsymbol{u})\tilde{q}_{x,1}(\boldsymbol{x})$
$\boldsymbol{x}$
$g(\boldsymbol{x})$
$\tilde{q}_{u,0}(\boldsymbol{x})$

Fig. 2. Second factor graph (FG2) used to perform EP-based estimation using auxiliary variables.

$f(\boldsymbol{y}|\boldsymbol{u})f(\boldsymbol{u}|\boldsymbol{x})g(\boldsymbol{x})$, with $f(\boldsymbol{u}|\boldsymbol{x}) = \prod_{m=1}^{M}\delta(u_m - \boldsymbol{a}_m^T\boldsymbol{x}) = \delta(\boldsymbol{u}-\boldsymbol{A}\boldsymbol{x})$, where $\delta(\cdot)$ denotes the Dirac delta function. Note that Eq. (1) is recovered using $\int f(\boldsymbol{y},\boldsymbol{u},\boldsymbol{x})d\boldsymbol{u}$. As in Section 2.2, the exact joint density can be approximated by a product of unnormalized multivariate Gaussian distributions (see Fig. 2), i.e., $\tilde{q}_{u,i}(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ and $\tilde{q}_{x,i}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{m}_i,\boldsymbol{S}_i)$, $\forall i \in \{0;1\}$. In particular, the factor $\delta(\boldsymbol{u}-\boldsymbol{A}\boldsymbol{x})$ is approximated such that $\boldsymbol{u}$ and $\boldsymbol{x}$ are a posteriori independent with

$$Q(\boldsymbol{u},\boldsymbol{x}) = Q(\boldsymbol{u})Q(\boldsymbol{x}) = \tilde{q}_{u,0}(\boldsymbol{u})\tilde{q}_{u,1}(\boldsymbol{u})\tilde{q}_{x,0}(\boldsymbol{x})\tilde{q}_{x,1}(\boldsymbol{x}).$$

Moreover, the vectors $\boldsymbol{u}$ and $\boldsymbol{x}$ are a posteriori Gaussian with mean and covariance $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $(\boldsymbol{m},\boldsymbol{S})$, respectively. Since $f(\boldsymbol{y}|\boldsymbol{u}) = f_y(\boldsymbol{u})$ and $g(\boldsymbol{x})$ can be factorized as a product of M and N independent terms, respectively, we assume that $\boldsymbol{\Sigma}_0$ and $\boldsymbol{S}_0$ are diagonal. As in Section 2.2, the EP algorithms will update the different factors sequentially, but depending on the structure of $(\boldsymbol{\Sigma}_1,\boldsymbol{S}_1)$, some updates can be performed independently and in a parallel manner. To allow parallel updates, we impose that $\boldsymbol{\Sigma}_1$ is either diagonal or isotropic, leading to $\boldsymbol{\Sigma}$ being diagonal. This approach is particularly interesting here since $M \geq N$ as it avoids inverting the potentially large and ill-conditioned matrix $\boldsymbol{\Sigma}_1$. The matrix $\boldsymbol{S}_1$ can also be either full, diagonal or isotropic (i.e., proportional to the identity matrix). The different EP-models are summarized in Table 1, where ”F”, ”D” and ”I” in the EP acronyms stand for ”full”, ”diagonal” and ”isotropic” covariance matrices. We now discuss how the different updates can be performed.

Update of $\tilde{q}_{u,0}$: the update of $\tilde{q}_{u,0}$ reduces to minimizing

$$KL\left(f_y(\boldsymbol{u})\tilde{q}_{u,1}(\boldsymbol{u})||\tilde{q}_{u,0}(\boldsymbol{u})\tilde{q}_{u,1}(\boldsymbol{u})\right), \tag{3}$$

w.r.t. $\tilde{q}_{u,0}$. Since the distribution $\hat{q}_{u,1}(\boldsymbol{u}) = f_y(\boldsymbol{u})\tilde{q}_{u,1}(\boldsymbol{u})$ factorizes over each u_n, its mean and diagonal covariance can be estimated component-wise using the method proposed in [6]. The positivity of diag($\boldsymbol{\Sigma}_0$), the diagonal of $\boldsymbol{\Sigma}_0$, is ensured by including the additional constraints diag($\boldsymbol{\Sigma}^{-1}$) $\geq$ diag($\boldsymbol{\Sigma}_1^{-1}$) (element-wise) (see [15]). In this case, $(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ can be found analytically.

Update of $\tilde{q}_{x,0}$: the update of $\tilde{q}_{x,0}$ reduces to minimizing

$$KL\left(g(\boldsymbol{x})\tilde{q}_{x,1}(\boldsymbol{x})||\tilde{q}_{x,0}(\boldsymbol{x})\tilde{q}_{x,1}(\boldsymbol{x})\right). \tag{4}$$

If $\boldsymbol{S}_1$ is full, the moments of the tilted distribution $\hat{q}_{x,0}(\boldsymbol{x}) = g(\boldsymbol{x})\tilde{q}_{x,1}(\boldsymbol{x})$ cannot be computed analytically due to the positivity constraints imposed on $\boldsymbol{x}$ in $g(\boldsymbol{x})$. However, the elements of $\boldsymbol{m}_0$ and the diagonal of $\boldsymbol{S}_0$ can be updated sequentially (N updates), as in Section 2.2. If $\boldsymbol{S}_1$ is diagonal, so is $\hat{\boldsymbol{S}}_0$, the covariance of the tilted distribution $\hat{q}_{x,0}(\boldsymbol{x})$. Thus, $(\hat{\boldsymbol{S}}_0,\boldsymbol{m}_0)$ can be computed analytically (using a single update). The mean $\boldsymbol{m}_0$ is updated so that $\boldsymbol{m}$ matches the mean of $\hat{q}_{x,0}(\boldsymbol{x})$. The diagonal of $\boldsymbol{S}_0$ is updated by imposing positivity constraints during the minimization of Eq. (4), in a similar fashion to the update of $\tilde{q}_{u,0}$.

Update of $(\tilde{q}_{u,1},\tilde{q}_{x,1})$: the update of $(\tilde{q}_{u,1},\tilde{q}_{x,1})$ consists of minimizing the KL divergence

$$KL\left(\delta(\boldsymbol{u}-\boldsymbol{A}\boldsymbol{x})\tilde{q}_{u,0}(\boldsymbol{u})\tilde{q}_{x,0}(\boldsymbol{x})||Q(\boldsymbol{u},\boldsymbol{x})\right) \tag{5}$$

w.r.t. $(\tilde{q}_{u,1},\tilde{q}_{x,1})$. Computing the full covariance matrix associated with the tilted distribution $\hat{q}_{u,x}(\boldsymbol{u},\boldsymbol{x}) = \delta(\boldsymbol{u}-\boldsymbol{A}\boldsymbol{x})\tilde{q}_{u,0}(\boldsymbol{u})\tilde{q}_{x,0}(\boldsymbol{x}))$ is not possible but since $Q(\boldsymbol{u},\boldsymbol{x}) = Q(\boldsymbol{u})Q(\boldsymbol{x})$ with $Q(\boldsymbol{u}) = \prod_{m=1}^{M}Q_m(u_m)$, it is sufficient to compute the moments of the marginals of $\hat{q}_{u,x}(\boldsymbol{u},\boldsymbol{x})$ w.r.t. $\boldsymbol{x}$ and $\{u_m\}_m$. It can be shown [14] that the marginal $\int \hat{q}_{u,x}(\boldsymbol{u},\boldsymbol{x})d\boldsymbol{u} = \tilde{q}_{u,0}(\boldsymbol{A}\boldsymbol{x})\tilde{q}_{x,0}(\boldsymbol{x})$ is proportional to a multivariate Gaussian distribution whose mean and covariance $(\hat{\boldsymbol{m}}_1,\hat{\boldsymbol{S}}_1)$ can be computed analytically. If $\boldsymbol{S}_1$ is full, it is updated using $\boldsymbol{S}_1^{-1} = \hat{\boldsymbol{S}}_1^{-1} - \boldsymbol{S}_0^{-1}$ and $\boldsymbol{m}_1 = \boldsymbol{S}_1(\hat{\boldsymbol{S}}_1^{-1}\hat{\boldsymbol{m}}_1 - \boldsymbol{S}_0^{-1}\boldsymbol{m}_0)$. If $\boldsymbol{S}_1$ is diagonal or proportional to the identity matrix, $(\boldsymbol{m}_1,\boldsymbol{S}_1)$ is obtained by minimizing (w.r.t. $\tilde{q}_{x,1}$) the KL divergence $KL\left(\int \hat{q}_{u,x}(\boldsymbol{u},\boldsymbol{x})d\boldsymbol{u}||Q(\boldsymbol{x})\right)$ between two Gaussian densities, subject to the appropriate constraints (positivity and/or equality of the diagonal elements of $\boldsymbol{S}_1$). Estimating the full covariance of $\int \hat{q}_{u,x}(\boldsymbol{u},\boldsymbol{x})d\boldsymbol{x}$ is not necessary here since $\boldsymbol{\Sigma}_1$ is assumed to be diagonal. Thus, it is sufficient to compute the marginal means and variances of the marginals $\int \hat{q}_{u,x}(\boldsymbol{u},\boldsymbol{x})d\boldsymbol{u}_{\backslash m}d\boldsymbol{x}, \forall m$, where $\boldsymbol{u}_{\backslash m}$ consists of the elements of $\boldsymbol{u}$ whose mth element has been removed. These marginals are proportional to univariate Gaussian distributions and the update of $\tilde{q}_{u,1}$ also reduces to minimizing a KL divergence between two Gaussian distributions.

3. RESULTS

We compare the estimation performance of the EP algorithms for unmixing spectral measurements from the single-photon multispectral Lidar system used in [18]. The matrix $\boldsymbol{A}$, whose columns are depicted in [18], consists of $N = 15$ spectral signatures of materials observed at $M = 33$ different spectral bands. Most of the signatures are highly correlated and the condition number of $\boldsymbol{A}$ is 1376. The quality of the data is tuned by scaling $\boldsymbol{A}$ using $\alpha\boldsymbol{A}$ where $\alpha \in \{5, 10, 20, 50, 100, 500, 1000, 5000\}$. For each value of α, the results are averaged over 2000 realizations of $\boldsymbol{x}$ drawn from $g(\boldsymbol{x})$ (product of independent exponential distributions with mean equal to 1). The resulting average photon counts $\mathrm{E}\left[y_m\right]$ range from 10 ($\alpha = 5$) to 10^4 ($\alpha = 5000$). Note that similar results have been obtained with truncated Gaussian priors. All the EP algorithms have been damped as in [15] using $\epsilon = 0.7$ set to reduce convergence issues. They are compared to an MCMC algorithm similar to that in [19], used to sample the exact posterior of $\boldsymbol{x}$, using $N_{MC} = 15000$ Monte Carlo iterations (including $N_{bi} = 5000$ burn-in iterations), and the estimated parameters (posterior means and covariances) are considered as exact estimates.

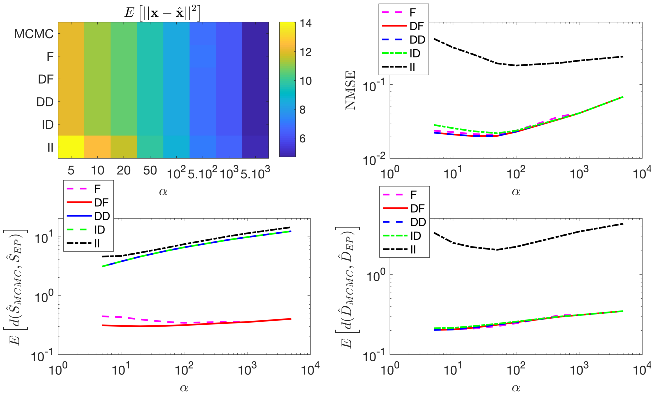

Fig. 3. Top MSEs (left) and NMSEs (right) as a function of α. Bottom: log-Euclidean distances between the actual posterior covariance and the covariances estimated via EP (left) and between these matrices whose off-diagonal terms are set to zero (right). The acronyms are those used in Table 1.

Fig. 3 compares the estimation performance of the different EP methods with the results obtained via MCMC. Precisely, the top left subplot depicts the mean square errors (MSEs) $E\left[||\mathbf{x}-\hat{\mathbf{x}}||^2\right]$, where $\hat{\mathbf{x}}$ is the posterior mean computed with MCMC or EP approximation. The top right subplot compares the mean normalised mean squared errors between the posterior means computed by MCMC and those approximated by EP, i.e., $NMSE = \dfrac{||\hat{\mathbf{x}}_{EP} - \hat{\mathbf{x}}_{MCMC}||^2}{||\mathbf{x} - \hat{\mathbf{x}}_{MCMC}||^2}$. The errors are normalised by $||\mathbf{x} - \hat{\mathbf{x}}_{MCMC}||^2$ to highlight the performance degradation w.r.t. the MCMC algorithm. These two subplots show that EP-F, EP-DF, EP-DD and EP-ID generally lead to similar performance in estimating the posterior mean with NMSEs remaining below 7%, in contrast to EP-II which leads to significantly worse results. Moreover, it is worth noting that EP-DF performs slightly better than EP-F, mostly due to the slower convergence of EP-F, which is also more prone to oscillations not attenuated by damping ($\leq 0.1\%$ of all the signals processed in this study, for all methods). The posterior covariance matrix estimation is assessed using the log-Euclidean distance [20], defined as $d(\boldsymbol{X}_1, \boldsymbol{X}_2) = ||\log(\boldsymbol{X}_1) - \log(\boldsymbol{X}_2)||_F$, for positive semi-definite matrices $(\boldsymbol{X}_1, \boldsymbol{X}_2)$, where $\log(\cdot)$ is the matrix logarithm. The bottom left subplot of Fig 3 compares the distances between the posterior covariance matrices obtained via EP ($\hat{\boldsymbol{S}}_{EP}$) to those obtained by MCMC ($\hat{\boldsymbol{S}}_{MCMC}$). The bottom right subplot compares the performance regarding marginal posterior variance estimation, using $\hat{\boldsymbol{D}}_{EP}$ and $\hat{\boldsymbol{D}}_{MCMC}$, the diagonal matrices whose diagonal elements match those of $\hat{\boldsymbol{S}}_{EP}$ and $\hat{\boldsymbol{S}}_{MCMC}$, respectively. While EP-F and EP-DF are, by construction, the only methods able to capture the posterior correlations (bottom left subplot of Fig. 3), it is interesting to note that EP-DD and EP-ID can estimate satisfactorily the marginal variances. Conversely, EP-II yields much larger distances. It is worth mentioning that, again, EP-DF performs slightly better than EP-F, probably due to the convergence issues mentioned above.

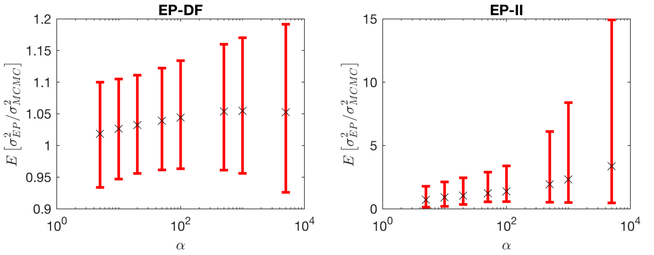

Fig. 4. Analysis of the marginal variances obtained with EP-DF and EP-II, compared to MCMC. The error bars correspond to the 5th-95th percentile intervals

Finally, Fig. 4 compares the mean ratios between the (actual) N marginal variances obtained by MCMC $\hat{\sigma}^2_{MCMC}$ and those approximated by EP $\hat{\sigma}^2_{EP}$, using EP-DF and EP-II. The results obtained with EP-F, EP-DD and EP-ID are similar to those obtained with EP-DF and are not presented here. This figure shows that EP-DF only slightly overestimates the marginal variances (by less than 6% on average) with a small variability of the marginal variance estimates (the 5th-95th percentile interval only varies by 20% in extreme). In contrast, EP-II tends to overestimate more significantly the marginal variances (by $\approx$ 4 times for large α) and the variability in these estimates is large and thus they do not allow for reliable uncertainty quantification.

4. CONCLUSION

We compared several EP models for regression with Poisson observation noise. Using an extended factor graph, constraints can be imposed to factor nodes, which allows simpler and parallel updates. The EP-based estimates obtained are similar to those obtained via Monte Carlo sampling, but at a much lower cost (computational comparisons omitted here due to space constraints). While graph constraints will impact the quality of the approximations, the EP approximations can be used to develop more scalable inference processes, either for inference in higher dimensions or for hierarchical graphs [15, 14]. Interestingly, additional structural constraints seem to also reduce convergence issues in practice. Future work include a deeper analysis of these preliminary observations and applications to more complex priors (e.g., using spatial correlations as in [21]) and noise models (e.g., binomial).

5. REFERENCES

[1] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[2] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (cccp)," in *Advances in neural information processing systems*. 2002, pp. 1033–1040, MIT Press.

[3] T. P. Minka, "Expectation propagation for approximate bayesian inference," in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[4] M. W. Seeger, "Bayesian inference and optimal design for the sparse linear model," *J. Mach. Learn. Res.*, vol. 9, pp. 759–813, June 2008.

[5] P. Schniter, S. Rangan, and A. K. Fletcher, "Vector approximate message passing for the generalized linear model," in *50th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2016, pp. 1525–1529.

[6] Y.-J. Ko and M. W. Seeger, "Expectation propagation for rectified linear poisson regression," in *Asian Conference on Machine Learning*, Hong Kong, Nov. 2016, vol. 45 of *Proceedings of Machine Learning Research*, pp. 253–268.

[7] M. Figueiredo and J. Bioucas-Dias, "Restoration of Poissonian images using alternating direction optimization," *IEEE Trans. Image Processing*, vol. 19, no. 12, pp. 3133–3145, 2010.

[8] Y. Altmann, A. Maccarone, A. Halimi, A. McCarthy, G. S. Buller, and S. McLaughlin, "Efficient range estimation and material quantification from multispectral lidar waveforms," in *Proc. Sensor Signal Processing for Defence (SSPD) Conference*, Edinburgh, UK, Sept. 2016.

[9] R. Tobin, Y. Altmann, X. Ren, A. McCarthy, R. A. Lamb, S. McLaughlin, and G. S. Buller, "Comparative study of sampling strategies for sparse photon multispectral lidar imaging: towards mosaic filter arrays," *Journal of Optics*, vol. 19, no. 9, pp. 094006, 2017.

[10] S. Brooks, *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis, 2011.

[11] Y. Altmann, A. Maccarone, A. McCarthy, G. Newstadt, G. S. Buller, S. McLaughlin, and A. Hero, "Robust spectral unmixing of sparse multispectral lidar waveforms using gamma Markov random fields," *IEEE Trans. Comput. Imaging*, vol. 3, no. 4, pp. 658–670, Dec. 2017.

[12] J. Tachella, Y. Altmann, M. Pereyra, and J.-Y. Tourneret, "Bayesian restoration of high-dimensional photon-starved images," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Rome, Italy, Sept. 2018.

[13] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet, "The bouncy particle sampler: A nonreversible rejection-free markov chain monte carlo method," *Journal of the American Statistical Association*, vol. 113, no. 522, pp. 855–867, 2018.

[14] A.S.I. Kim and M. P. Wand, "On expectation propagation for generalised, linear and mixed models," *Australian & New Zealand Journal of Statistics*, vol. 60, no. 1, pp. 75–102, 2018.

[15] J. Hernández-Lobato, D. Hernández-Lobato, and A. Suárez, "Expectation propagation in linear regression models with spike-and-slab priors," *Machine Learning*, vol. 99, no. 3, pp. 437–487, Jun 2015.

[16] M. P. Wand, J. T. Ormerod, S. A. Padoan, and R. Frhwirth, "Mean field variational Bayes for elaborate distributions," *Bayesian Anal.*, vol. 6, no. 4, pp. 847–900, 2011.

[17] A. Perelli, M. A. Lexa, A. Can, and M. E. Davies, "Denoising message passing for X-ray computed tomography reconstruction," *CoRR*, vol. abs/1609.04661, 2016.

[18] Y. Altmann, A. Maccarone, A. McCarthy, S. McLaughlin, and G. S. Buller, "Spectral classification of sparse photon depth images," *Opt. Express*, vol. 26, no. 5, pp. 5514–5530, March 2018.

[19] P. Caramazza, K. Wilson, G. Gariepy, J. Leach, S. McLaughlin, D. Faccio, and Y. Altmann, "Enhancing the recovery of a temporal sequence of images using joint deconvolution," *Scientific Reports*, vol. 8, no. 1, pp. 5257, Jun 2018.

[20] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-Euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic Resonance in Medicine*, vol. 56, no. 2, pp. 411–421, Aug. 2006.

[21] J. Vila, P. Schniter, and J. Meola, "Hyperspectral unmixing via turbo bilinear approximate message passing," *IEEE Trans. Comput. Imaging*, vol. 1, no. 3, pp. 143–158, Sept 2015.