On approximate Bayesian methods for large-scale sparse linear inverse problems

Yoann Altmann
School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK

***Abstract*—In this paper, we investigate and compare approximate Bayesian methods for high-dimensional linear inverse problems where sparsity-promoting prior distributions can be used to regularized the inference process. In particular, we investigate fully factorized priors which lead to multimodal and potentially non-smooth posterior distributions such as Bernoulli-Gaussian priors. In addition to the most traditional variational Bayes framework based on mean-field approximation, we compare different implementations of power expectation-propagation (EP) in terms of estimation of the posterior means and marginal variances, using fully factorized approximations. The different methods are compared using low-dimensional examples and we then discuss the potential benefits of power EP for image restoration. These preliminary results tend to confirm that in the case of Gaussian likelihoods, EP generally provides more reliable marginal variances while power EP offers more flexibility for generalised linear inverse problems.**

***Index Terms*—Approximate Bayesian inference, Expectation-Propagation, Linear inverse problems, Variational Bayes, Sparsity.**

I. Introduction

Large scale linear inverse problems are ubiquitous in a variety of imaging problems where images of interest can not only have millions of pixels, but also dozen to hundreds of spectral channels [1]. Similar problems also arise when restoring image sequences (e.g., videos). Statistical inference in such high-dimensional problem often relies on statistical properties of the signal of interest, and in particular its compact representation in particular domains of representation (e.g., Fourier domain, wavelet domains, library of known patterns). As such, leveraging the sparsity or compact representation of parameters of interest has received a significant attention over the last 30 years, in particular when combined with high-dimensional optimization methods [2], [3]. However, high-dimensional inference is often reduced to point estimation (penalized maximum likelihood or maximum a posteriori estimation) and uncertainty quantification remains difficult in general. Within the Bayesian framework, Markov chain Monte Carlo sampling is the most widely used approach to uncertainty quantification but efficient and generic samplers for high-dimensional problems are still required. Variational methods represent a computationally attractive alternative, provided that the approximations capture well the structure of the posterior distribution. While Variational Bayes methods can provide useful approximate posterior means, they often drastically underestimate uncertainties, in particular when mean-field approximations are used. In this paper, we investigate a more general class of variational inference tools called power EP (or $\alpha-$EP) [4], [5] for large scale linear inverse problems where the sparsity-promoting prior can induce a multimodal and potentially non-differentiable posterior distribution.

This work was supported by the Royal Academy of Engineering under the Research Fellowship scheme RF201617/16/31, by the Engineering and Physical Sciences Research Council (EPSRC) Grant number and EP/S000631/1 and the UK MOD University Defence Research Collaboration (UDRC) in Signal Processing.

II. Bayesian model and exact inference

A. Bayesian model

In this work, we address the recovery of $\boldsymbol{x} \in \mathbb{R}^N$ from noisy observations $\boldsymbol{y} \in \mathbb{R}^M$ which result from the transformation of $\boldsymbol{x}$ via a known linear operator $\boldsymbol{A} \in \mathbb{R}^{M\times N}$.

1) Likelihood: While the observation noise could be additive or multiplicative, here we mainly focus on additive Gaussian noise leading to the following likelihood

$$\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{\Sigma}$ is the noise covariance matrix assumed to be diagonal and known. The Gaussian noise assumption is not crucial for some of the methods discussed in this work, but it makes their comparison easier.

2) Prior model: As mentioned above, we consider fully factorized prior distributions such that

$$f(\boldsymbol{x}|\boldsymbol{\Theta}) = \prod_{n=1}^{N} f_x(x_n|\boldsymbol{\theta}_n), \tag{2}$$

where $f_x(x_n|\boldsymbol{\theta}_n)$ is parameterized by a set of hyperparameters denoted by $\boldsymbol{\theta}_n$, and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_n\}_n$ is known. In the remainder of the paper, $\boldsymbol{\Theta}$ is omitted in the expression of the prior model to simplify notations. When $\boldsymbol{x}$ is expected to only contain a small number of large values, several prior distributions $\{f_x(x_n|\boldsymbol{\theta}_n)\}_n$ can be chosen to encode such prior belief. Classical choices include Laplace distributions [6], Student's t-distributions [7], and spike-and-slab priors [8]. For arguments similar to those motivating the Gaussian noise model, we consider spike-and-slab priors based on i) mixtures of two zero-mean univariate Gaussian distributions (MoG2) and ii) Bernoulli-Gaussian (BG) mixtures. These priors are chosen as they lead to multimodal posterior distributions (in contrast to the product of Laplace priors, when combined with the Gaussian likelihood). Moreover, the BG distributions are non-differentiable with respect to (w.r.t.) $\boldsymbol{x}$, which might limit the range of variational methods applicable for the problem at hand.

B. Posterior distribution and exact inference

Given the likelihood defined in (1) and the prior model in (2), the exact posterior distribution of $\boldsymbol{x}$, i.e. $f(\boldsymbol{x}|\boldsymbol{y})$, can be expressed, up to an (often) intractable constant, as

$$f(\boldsymbol{x}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x}). \tag{3}$$

Estimation of $\boldsymbol{x}$ beyond maximum a posteriori (MAP) estimation is generally challenging in high-dimensional problems ($N \gg 1$) due to the intractable integrals involved to compute (marginal) moments a posteriori. A notable exception is when $f(\boldsymbol{x}|\boldsymbol{y})$ is log-concave, but not necessarily smooth. In such cases, recent MCMC methods can be used to sample approximately from $f(\boldsymbol{x}|\boldsymbol{y})$ [9]–[12]. In more general scenarios, more classical MCMC samplers can be used, including Gibbs sampling to sample sequentially (blocks of) elements of $\boldsymbol{x}$ from their conditional distributions. However, Gibbs sampling and Metropolis-Hastings updates do not scale well with increasing N and alternative methods are required to approximate posterior moments such as means and (marginal) variances efficiently. The next section discusses variational inference methods that can be used to approximate (3) for large scale problems where $N \gg 1$.

III. Scalable variational inference using power expectation-propagation

Variational inference (VI) methods consist of approximating a distribution, $f(\boldsymbol{x}|\boldsymbol{y})$ here, by a so-called approximating distribution $q(\boldsymbol{x})$, such that $f(\boldsymbol{x}|\boldsymbol{y}) \approx q(\boldsymbol{x})$. The distribution $q(\boldsymbol{x})$ is often chosen such that its moments are easy to compute and the different VI methods mainly differ by the similarity measure used to compare $f(\boldsymbol{x}|\boldsymbol{y})$ and $q(\boldsymbol{x})$. For instance, variational Bayesian (VB) methods, which represent the most classical family of VI methods, rely on minimizing the Kullback-Leibler (KL) divergence

$$\mathrm{KL}\left(q(\boldsymbol{x})||f(\boldsymbol{x}|\boldsymbol{y})\right) \tag{4}$$

subject to additional constraints on $q(\boldsymbol{x})$. These constraints can be independence constraints, leading to mean-field VB (MFVB) and/or constraints on the admissible family of approximating distributions, leading to fixed-form VB (FFVB) (see recent tutorial [13] and reference therein). For the Bayesian model in (3) in high-dimensional settings, one of the most suitable VB method is that proposed in [14]. This method considers an extended Bayesian model including binary labels for the spike-and-slab prior and a mean-field approximation is used such that $q(\boldsymbol{x})$ factorizes over the N elements of $\boldsymbol{x}$. While the method was proposed with a BG prior, it can easily be modified for MoG2 priors.

While full factorization of $q(\boldsymbol{x})$ may not be a reasonable assumption for some problems, it is the most common and simple approach to scalability as it does not require handling large covariance or precision matrices. For this reason, in the remainder of this section we investigate VI methods relying on such constraints. Different covariance constraints will be discussed in Section V.

A. General power EP method

In a similar fashion to FFVB, in EP methods, the practitioner can choose the set of admissible distributions for $q(\boldsymbol{x})$, generally in the exponential family, and especially Gaussian distributions for high dimensional problems, as is done here. EP and its extensions, such as $\alpha-$EP [5], rely on factorizations of $q(\boldsymbol{x})$ which mimic that of $f(\boldsymbol{x}|\boldsymbol{y})$. More precisely, we factorize $q(\boldsymbol{x}) = q_1(\boldsymbol{x})q_0(\boldsymbol{x})$ using 2 unnormalized multivariate Gaussian densities: $q_1(\boldsymbol{x})$, which will approximate $f(\boldsymbol{y}|\boldsymbol{x})$ and $q_0(\boldsymbol{x})$, which approximates $f(\boldsymbol{x})$. The mean and covariance matrix of $q_i(\boldsymbol{x})$ are denoted by $(\boldsymbol{m}_i, \boldsymbol{D}_i)$ ($i \in \{0; 1\}$) and the mean and covariance of $q(\boldsymbol{x})$, denoted by $(\boldsymbol{m}, \boldsymbol{D})$ satisfy

$$\begin{cases} \boldsymbol{D}^{-1} & = \boldsymbol{D}_1^{-1} + \boldsymbol{D}_0^{-1} \\ \boldsymbol{D}^{-1}\boldsymbol{m} & = \boldsymbol{D}_1^{-1}\boldsymbol{m}_1 + \boldsymbol{D}_0^{-1}\boldsymbol{m}_0 \end{cases} \tag{5}$$

To ensure $\boldsymbol{D}$ is diagonal and positive definite, so are $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$. We now briefly recall the working principle of $\alpha-$EP, which updates sequentially the variational parameters of $q_0(\boldsymbol{x})$ and $q_1(\boldsymbol{x})$ until convergence.

1) Update of $q_1(\boldsymbol{x})$: We first define a so-called cavity distribution $q_{\backslash 1}(\boldsymbol{x}) = q(\boldsymbol{x})/q_1(\boldsymbol{x}) \propto q_0(\boldsymbol{x})$. To update $q_1(\boldsymbol{x})$, we aim to solve

$$q^{new}(\boldsymbol{x}) = \underset{q\in\mathcal{F}}{\arg\min} D_{\alpha_1}(f(\boldsymbol{y}|\boldsymbol{x})q_{\backslash 1}(\boldsymbol{x})||q(\boldsymbol{x})), \tag{6}$$

where $\mathcal{F}$ denotes the set of multivariate Gausian densities with positive definite diagonal covariance matrices and D_{α_1} denotes the $\alpha-$divergence given by

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{(1+\alpha)/2}q(x)^{(1-\alpha)/2}\mathrm{d}x\right), \tag{7}$$

for $\alpha \notin \{-1; 1\}$. When $\alpha = 1$, it reduces to $\mathrm{KL}(p||q)$, and when $\alpha = -1$, it is $\mathrm{KL}(q||p)$. For $\alpha \neq -1$ and following [5], solving (6) reduces to

$$q^{new}(\boldsymbol{x}) = \mathrm{proj}\left[\left(f(\boldsymbol{y}|\boldsymbol{x})q_{\backslash 1}(\boldsymbol{x})\right)^{1/n_1} q(\boldsymbol{x})^{1-1/n_1}\right], \tag{8}$$

with $1/n_1 = (1+\alpha_1)/2$ and $\mathrm{proj}\left[\cdot\right]$ is the KL-projection onto $\mathcal{F}$. Since the Gaussian distributions in $\mathcal{F}$ have diagonal covariance matrices, computing $\mathrm{proj}\left[h(\boldsymbol{x})\right]$ means computing the mean and covariance matrix of $h(\boldsymbol{x})$, and returning a Gaussian with the same mean and a diagonal covariance matrix whose diagonal elements match those of the covariance of the density $h(\boldsymbol{x})$. Note that if $\alpha_1 = -1$, Eq. (6) reduces to applying MFVB to the tilted distribution $\tilde{p}_1(\boldsymbol{x}) \propto f(\boldsymbol{y}|\boldsymbol{x})q_{\backslash 1}(\boldsymbol{x})$. In this case, $q^{new}(\boldsymbol{x})$ has the same mean as $\tilde{p}_1(\boldsymbol{x})$ and the diagonal of its covariance matrix is the inverse of the main diagonal of the precision matrix of $\tilde{p}_1(\boldsymbol{x})$. Computing the projection in (8) is the main difficult task in EP and power EP. However, since the likelihood and $q(\boldsymbol{x})$ are Gaussian here, this can be done effectively via Monte Carlo sampling (see next paragraph). Once $q^{new}(\boldsymbol{x})$ is computed, the updated $q_1(\boldsymbol{x})$ is obtained using $q_1(\boldsymbol{x}) \propto q^{new}(\boldsymbol{x})/q_0(\boldsymbol{x})$ (if no damping is used).

2) Update of $q_0(\boldsymbol{x})$: $q_0(\boldsymbol{x})$ is updated by solving

$$q^{new}(\boldsymbol{x}) = \arg\min_{q\in\mathcal{F}} D_{\alpha_0}(f(\boldsymbol{x})q_{\backslash 0}(\boldsymbol{x})||q(\boldsymbol{x})), \tag{9}$$

with $q_{\backslash 0}(\boldsymbol{x}) = q(\boldsymbol{x})/q_0(\boldsymbol{x}) \propto q_1(\boldsymbol{x})$, and by setting $q_0(\boldsymbol{x}) \propto q^{new}(\boldsymbol{x})/q_1(\boldsymbol{x})$. Damping the updates of $q_1(\boldsymbol{x})$ and $q_0(\boldsymbol{x})$, is possible with EP and $\alpha-$EP (e.g., see [5], [15]). This strategy generally slows down the convergence but makes the updates more stable.

B. Selection of the divergence parameters

The $\alpha-$EP framework allows the practionner to choose the parameters (α_0, α_1), which can in principle be different. Different options can be investigated but a trade-off must be found between quality of the final estimates, computational cost of each update and overall stability of the algorithm (whose convergence is not guaranteed). As discussed in [16], small (i.e., negative) values of α lead to exclusive divergences which tend to capture well one of the modes of the distributions while large values of α lead to more inclusive divergences such that $q(\boldsymbol{x})$ tends to also cover the tails of $p(\boldsymbol{x})$ when minimizing (7) w.r.t. $q(\boldsymbol{x})$. In the context of this paper, were $q_0(\boldsymbol{x})$, $q_1(\boldsymbol{x})$ and therefore $q(\boldsymbol{x})$ have diagonal covariance matrices, we can explore efficiently various values of (α_0, α_1). Eq. (9) can be minimized analytically and easily for $\alpha_0 = 1$ since the prior is fully factorized and composed of MoG2 or BG distributions (see [15]). For other values of α_0, solving Eq. (9) is harder and not as scalable, thus we fix $\alpha_0 = 1$.

For α_1, we have much more flexibility since, for $\alpha_1 \neq -1$,

$$\tilde{f}_1(\boldsymbol{x}) \propto \left(f(\boldsymbol{y}|\boldsymbol{x})q_{\backslash 1}(\boldsymbol{x})\right)^{1/n_1} q(\boldsymbol{x})^{1-1/n_1} \tag{10}$$

is Gaussian, with mean and covariance matrix $(\boldsymbol{\mu}, \boldsymbol{S})$ such that

$$\begin{cases} \boldsymbol{S}^{-1} &= \dfrac{\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A}}{n_1} + \dfrac{\boldsymbol{D}_0^{-1}}{n_1} + \dfrac{(n_1-1)\boldsymbol{D}^{-1}}{n_1} \\ \boldsymbol{S}^{-1}\boldsymbol{\mu} &= \dfrac{\boldsymbol{\Sigma}^{-1}\boldsymbol{A}^T\boldsymbol{y}}{n_1} + \dfrac{\boldsymbol{D}_0^{-1}\boldsymbol{m}_0 + (n_1-1)\boldsymbol{D}^{-1}\boldsymbol{m}}{n_1}. \end{cases} \tag{11}$$

Computing $\boldsymbol{\mu}$ can be done without inverting $\boldsymbol{S}^{-1}$, e.g. using preconditioned gradient descent since left-multiplying by $\boldsymbol{S}^{-1}$ is generally fast. As mentioned above, only $\text{diag}(\boldsymbol{S})$ (not the full matrix covariance matrix $\boldsymbol{S}$) is required to compute (8). Given the structure of $\boldsymbol{S}^{-1}$, it is easy to approximate $\text{diag}(\boldsymbol{S})$ via Monte Carlo sampling [17], [18], even in high dimensions, and this is the approach adopted here. Although α_1 can in principle take any value, $\alpha_1 > 1$ does not lead to stable updates and we restrict our analysis to $\alpha_1 \in [-1, 1]$. In the remainder of the paper, we use the notation $\alpha-\text{EP}_{\alpha_1}$ to highlight which α_1 value is used. Note that $\alpha-\text{EP}_1$ correspond to the EP algorithm described in [15], and that solving (6) involves the same cost $\forall\alpha_1 > -1$ (which is usually higher than for $\alpha_1 = -1$ (VB)).

IV. Experiments

A. 2D Gaussian likelihood

Prior to investigating large problems, we first visualise the results of the different methods discussed in this paper using a simple example with $N = M = 2$. We set $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ such that $(\boldsymbol{A}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{A})^{-1} = [0.2, 0.255; 0.255; 0.4]$, the ground truth $\boldsymbol{x}_0$ is set to $\boldsymbol{x}_0 = [1.2, 0]^T$. We then consider two prior models

1) a MoG2 prior, the same for both components of $\boldsymbol{x}$, with $(m_1, s_1^2) = (0, 0.01)$ and $(m_2, s_2^2) = (0, 10)$ and the prior weight of the first Gaussian distribution is set to 0.70 to promote small values.
2) a BG prior, the same for both components of $\boldsymbol{x}$, with $(m_2, s_2^2) = (0, 10)$ and the prior weight of the Dirac delta function is set to 0.70.

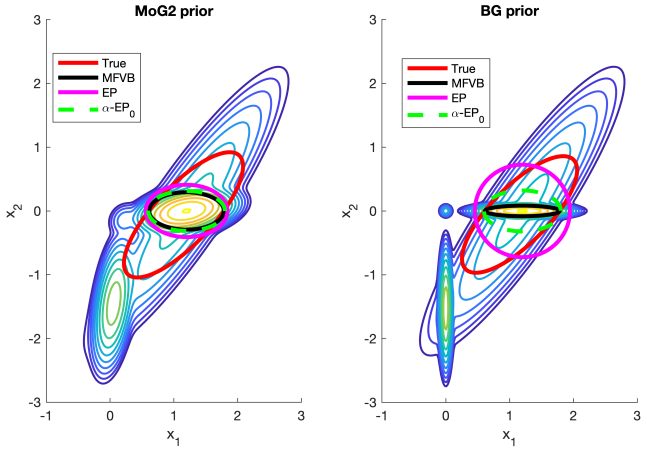

Fig. 1. Examples of variational approximation of 2D posterior using MVFB and $\alpha-$EP. The ellipses show the approximate regions of high posterior density. The red ellipses are obtained from Gaussian densities which have the same means and covariance matrices as the true posteriors. In these examples, all methods correctly identify the primary modes of the posterior distributions and provide similar posterior mean estimates.

For these two priors, exact computation of the posterior is tractable, and it is possible to apply the MFVB method from [14] for comparison. Figure 1 depicts the regions of high density approximated by the different methods for the two scenarios. The level sets represent the actual posterior distribution, although on the right-hand side, a smoothed representation of the exact posterior is displayed (the true posterior is a mixture of degenerated densities). This figure illustrates how MFVB and more generally the use of exclusive divergences tend to underestimate posterior variances, even in small dimensions. The $\alpha-\text{EP}_{-1}$ results are not displayed here as they are very similar to those of MFVB, and by increasing α_1, the estimated marginal variances increase and become closer to the actual marginal variances of the true posterior.

B. Estimation performance

We now investigate the estimation of a sparse vector of length $N = 100$ using $M \in \{50; 100; 500\}$ noisy observations. The matrix $\boldsymbol{A}$ is constructed by multiplying an $M \times N$ and an $N \times N$ random matrix with i.i.d. elements drawn for a standard Gaussian distribution. Moreover, $\boldsymbol{\Sigma} = 0.01\boldsymbol{I}_M$ and the ground truth $\boldsymbol{x}_0$ is drawn from a product of independent and identical BG priors with $p(x_i = 0) = 0.73$ and where the slab is Gaussian with mean zero and variance equal to 10. The

Algorithm		M=5N	M=N	M=N/2
MFVB	BG	0.0060(0.016)	0.12(0.22)	0.74(0.0008)
	MoG2	0.0001(4.10^{-5})	0.0005(0.0001)	1.81(0.60)
$\alpha-\mathbf{EP}_{-1}$	BG	0.0001(4.10^{-5})	0.0005(0.0001)	1.34(0.36)
	MoG2	0.0001(4.10^{-5})	0.0005(0.0001)	1.34(0.36)
$\alpha-\mathbf{EP}_{-0.5}$	BG	0.0085(0.004)	0.0098(0.024)	0.60(0.0031)
	MoG2	0.0085(0.004)	0.0098(0.024)	0.60(0.0031)
$\alpha-\mathbf{EP}_{0}$	BG	0.0014(0.0018)	0.0010(0.0026)	0.52(0.012)
	MoG2	0.0014(0.0018)	0.0010(0.0026)	0.52(0.012)
$\alpha-\mathbf{EP}_{0.5}$	BG	0.0001(0.0001)	0.0005(0.0001)	0.49(0.015)
	MoG2	0.0001(0.0001)	0.0005(0.0001)	0.49(0.015)
EP	BG	0.0060(0.016)	0.12(0.22)	0.74(0.0008)
	MoG2	0.0001(4.10^{-5})	0.0005(0.0001)	1.81(0.60)

TABLE I
RMSE OF THE APPROXIMATE POSTERIOR MEAN (THE NUMBERS IN BRACKETS INDICATE THE STANDARD DEVIATIONS OVER 30 NOISE REALIZATIONS).

experiments are repeated with 30 noise realizations. For all the $\alpha-$EP results, the diagonals of $\boldsymbol{S}$ in (11) are estimated from 1000 independent Monte Carlo samples. Table IV-B reports the root mean squared errors between the ground truth $\boldsymbol{x}_0$ and the approximate MMSE estimates obtained by different approximate methods using as prior i) the BG prior used to generate $\boldsymbol{x}_0$ and ii) a MoG2 prior with the variance of the spike equal to 0.01 (and the other parameters as for the true prior). Note that the results of $\alpha-\text{EP}_1$ are the same as those of EP in our experiments and are thus not duplicated. When the problem is not severely ill-posed (e.g., when $M = 5N$), MFVB and EP tend to provide RMSEs lower than $\alpha-$EP, which also present larger standard deviations. When the likelihood becomes less informative due to $M < N$ $\alpha-$EP provides slightly smaller RMSEs than MFVB and EP. Moreover, $\alpha-$EP tends to provide similar estimates if based on the BG or MoG2 prior. We believe the variations of $\alpha-$EP are partly due to convergence issues as the algorithm requires damping to be stable. While EP also often needs to be damped, it seems to be less sensitive than $\alpha-$EP in the examples considered here. In addition to the RMSE of the approximate posterior mean, we also visualize the marginal variances estimated by the different methods. Fig. 2 shows examples of estimated variances for $M = 500$. As in Section IV-A, we observe that EP provides the largest marginal variances, expected to be closer to the actual marginal variances.

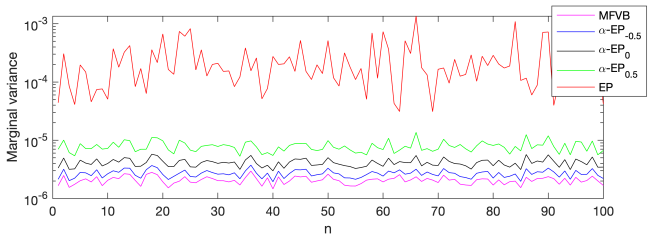

Fig. 2. Example of approximate marginal variances estimated by MFVB, $\alpha-$EP and EP with the MoG2 prior and $M = 500$.

C. Image deconvolution

Finally, to illustrate how $\alpha-$EP can be used for very large inverse problems, we investigate an image deconvolution problem using a 512×512 pixels image (cameraman). In this example $\boldsymbol{x}$ represents the wavelet coefficients of the unknown image, i.e., $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{W}^{-1}$, where $\boldsymbol{B}$ represent a 2D convolution and $\boldsymbol{W}$ is an invertible 2D wavelet transform (Daubechies 5 with 3 levels here). The coefficients of the coarse approximation are assigned the same weakly informative Gaussian prior and the detail coefficients are assigned the same MoG2 prior with the two means set to zeros, the variances set as $(1; 5.10^3)$ and the weight of the spike is set as 0.70. These prior parameters have not been optimized and a full comparison with alternative approaches is left to future work due to space constraints. The results discussed here are primarily intended to illustrate the potential of the methods for high-dimensional inference. Using $\hat{\boldsymbol{x}} = \boldsymbol{m}$ and $\boldsymbol{D}$ estimated by $\alpha-$EP, it is possible to approximate the unknown image using $\boldsymbol{W}^{-1}\boldsymbol{m}$ and its covariance matrix using $\boldsymbol{W}^{-1}\boldsymbol{D}\boldsymbol{W}^{-T}$. Again, to avoid the manipulation of large matrices, we only estimate its diagonal for visualization purposes, using Monte Carlo sampling [17]. An example of deconvolution is depicted in Fig. 3, where the observed image is blurred using a 9×9 pixels blur and blurred signal-to-noise ratios of 30dB and 20dB. We can observe that the marginal variances generally decrease as the BSNR increases (the data become less informative) and the uncertainties are larger in textured regions and around edges.

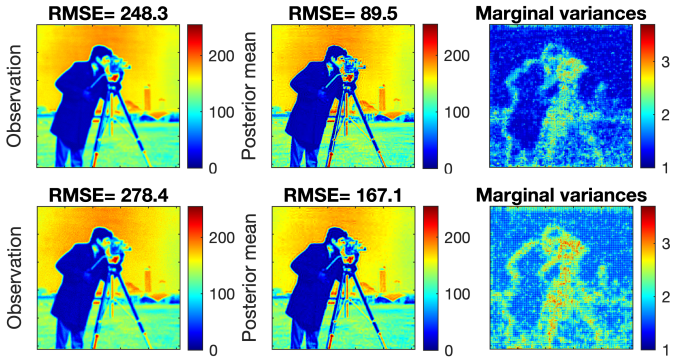

Fig. 3. Example of deconvolution results using EP and a MoG2 prior for BSNR=30dB (top) and BSNR=20dB (bottom).

V. CONCLUSION

In this paper, we compared several variational inference methods for small and then large scale sparse linear inverse problems with Gaussian likelihood. We illustrated how $\alpha-$EP can stand as a trade off between VB and EP methods, in terms of the marginal variance estimation. In the examples considered, it seems EP was the most attractive method but these initial results should be nuanced. All the methods considered here rely on fully factorized approximations, which may not be a reasonable constraint for some inverse problems. Adopting different, problem-specific, covariance constraints can favor the use of specific $\alpha-$divergences. For instance, low-rankness can be incorporated easily using FFVB [13]. On the other hand, EP does not require the posterior distribution to be differentiable

REFERENCES

[1] J. Peng, W. Sun, H.-C. Li, W. Li, X. Meng, C. Ge, and Q. Du, "Low-rank and sparse representation for hyperspectral image processing: A review," *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–35, 2021.

[2] J.-L. Starck and M. J. Fadili, "An overview of inverse problem regularization using sparsity," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 1453–1456.

[3] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009. [Online]. Available: https://doi.org/10.1137/080716542

[4] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 362–369.

[5] T. Minka, "Power EP," Microsoft Research Ltd., Tech. Rep. MSR-TR-2004-149, January 2004.

[6] M. W. Seeger, "Bayesian inference and optimal design for the sparse linear model," *Journal of Machine Learning Research*, vol. 9, no. 26, pp. 759–813, 2008.

[7] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, p. 211–244, sep 2001. [Online]. Available: https://doi.org/10.1162/15324430152748236

[8] E. I. George and R. E. McCulloch, "Approaches for Bayesian variable selection," *Statistica Sinica*, vol. 7, no. 2, pp. 339–373, 1997. [Online]. Available: http://www.jstor.org/stable/24306083

[9] A. Durmus, E. Moulines, and M. Pereyra, "Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 473–506, 2018.

[10] M. Vono, N. Dobigeon, and P. Chainais, "Split-and-augmented Gibbs sampler - Application to large-scale inference problems," *IEEE Transactions on Signal Processing*, vol. 67, no. 6, pp. 1648–1661, 2019.

[11] A. F. Vidal, V. De Bortoli, M. Pereyra, and A. Durmus, "Maximum likelihood estimation of regularization parameters in high-dimensional inverse problems: An empirical bayesian approach part i: Methodology and experiments," *SIAM J. Imag. Sci.*, vol. 13, no. 4, pp. 1945–1989, 2020.

[12] L. Vargas, M. Pereyra, and K. C. Zygalakis, "Accelerating proximal markov chain monte carlo by using an explicit stabilised method," *SIAM Journal on Imaging Sciences*, vol. 13, no. 2, pp. 905–935, 2019.

[13] M.-N. Tran, T.-N. Nguyen, and V.-H. Dao, "A practical tutorial on variational Bayes," 2021.

[14] P. Carbonetto and M. Stephens, "Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies," *Bayesian Analysis*, vol. 7, no. 1, pp. 73 – 108, 2012. [Online]. Available: https://doi.org/10.1214/12-BA703

[15] J. M. 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, 2015.

[16] T. Minka, "Divergence measures and message passing," Microsoft Research Ltd., Tech. Rep. MSR-TR-2005-173, December 2005.

[17] G. Papandreou and A. L. Yuille, "Gaussian sampling by local perturbations," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'10. Red Hook, NY, USA: Curran Associates Inc., 2010, p. 1858–1866.

[18] P. Sidén, F. Lindgren, D. Bolin, and M. Villani, "Efficient covariance approximations for large sparse precision matrices," *Journal of Computational and Graphical Statistics*, vol. 27, no. 4, pp. 898–909, 2018.