Expectation-Propagation with Low-Rank Constraints for Linear Inverse Problems

Dongrui Shen
Department of Mathematics
City University of Hong Kong
Hong Kong
dongrshen2@cityu.edu.hk

Stephen McLaughlin, Yoann Altmann
School of Engineering and Physical Sciences
Heriot-Watt University
Edinburgh, United Kingdom
{S.McLaughlin, Y.Altmann}@hw.ac.uk

***Abstract*—In this work, we address the problem of scalable approximate inference and covariance estimation for linear inverse problems using Expectation-Propagation (EP). Traditional EP methods rely on Gaussian approximations with either diagonal or full covariance structures. Full covariance matrices can capture correlation but do not scale as the dimensions of the problem increases, while diagonal matrices scale better but omit potentially important correlations. For the first time to the best of our knowledge, we propose to investigate low-rank decompositions within EP for linear regression. The potential benefits of such covariance structures are illustrated thorough simulations results obtained on sparse linear regression problems and a more challenging spectral unmixing problem where the sparse mixing coefficients are, in addition, subject to positivity constraints.**

***Index Terms*—Variational inference, sparse regression, Expectation-Propagation, low-rank representation, spectral unmixing.**

I. Introduction

Monte Carlo sampling has been the gold standard for Bayesian inference when posterior moments cannot be computed analytically. Significant improvements have been achieved in terms of scalability of Markov chain Monte Carlo (MCMC) methods, using Langevin diffusion [1] and Hamiltonian Monte Carlo methods [2]. Still, sampling high-dimensional distributions exhibiting multiple modes and strong correlations remains challenging.

Variational inference (VI) stands as an alternative to MCMC for high-dimensional inference. It consists of approximating the distribution of interest by a simpler distribution, whose moments are easier to compute. Variational Bayes (VB) is the most classical family of VI methods, due to the simplicity of its implementation. Another family of methods is referred to as Expectation-Propagation (EP) [3], which primarily differ from VB by the similarity measure (divergence) used to compare the true distribution of interest to its approximation. Early VB algorithm have relied on mean-field approximations to make the VB updates tractable, but such assumptions neglect potentially important correlations in the approximating distribution. More recently, another family of VB methods, referred to as fixed-form VB (FFVB) [4], where the practitioner can select the approximating distribution from a parametric family. An example of application of FFVB with Gaussian approximations and structured covariances has been recently discussed in [5]. This method is referred to as variational approximation with factor covariance (VAFC).

EP is an alternative to VB and uses a reverse Kullback-Leibler (KL) divergence. EP and VB can be interpreted within the more general framework of Power EP [6]. Although EP is in general more difficult to implement than VB, it suffers less from underestimation of the distribution variances or covariances. Hence it is a promising tool for more reliable uncertainty quantification. For high-dimensional problems, EP often relies on using multivariate Gaussian distributions as approximating distributions but full, unstructured covariance matrices can be hard to manipulate and store for large problems. Using diagonal matrices, referred to as diagonal EP, partially overcomes this issue [7], [8] but the resulting approximation still relies on a mean-field like assumption, which may not be appropriate.

In this preliminary work, we investigate how to incorporate and leverage low-rank structures within EP to partially capture correlation while keeping the EP update simple. Ultimately, such methods would be useful for large scale inverse problems, but in this work we concentrate on medium scale sparse linear inverse problems, for which the exact posterior distribution is still tractable. In the last set of experiments, we apply the proposed EP algorithm to a spectral unmixing problem, and discuss current limitations of the approach.

The rest of the paper is organised as follows: In section II, we briefly introduce the Bayesian model used for linear regression. Section III describes the EP algorithm with diagonal plus low-rank constraints. We demonstrate the performance of the proposed EP algorithm with several simulations in Section IV. Finally, the conclusions are summarised in section V.

Notation: $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the probability density function of a (multivariate) Gaussian random variable $\boldsymbol{x}$ with mean μ and covariance matrix $\boldsymbol{\Sigma}$. The KL divergence $KL(P(\boldsymbol{x})\|Q(\boldsymbol{x})) = \int P(\boldsymbol{x}) \log\left(P(\boldsymbol{x})/Q(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}$ is used to measure the discrepancy between the densities $P(\boldsymbol{x})$ and $Q(\boldsymbol{x})$, and $\mathrm{tr}(\cdot)$ and $|\cdot|$ denote the trace and the determinant of a matrix, respectively.

This research was supported by the Royal Academy of Engineering under the Research Fellowship scheme RF201617/16/31 and by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/T00097X/1.

II. Exact Bayesian model

We consider a classical linear forward model where the observed data $\boldsymbol{y} \in \mathbb{R}^L$ result from the noisy transformation of the signal of interest $\boldsymbol{x} \in \mathbb{R}^R$ by a linear operator represented by the matrix $\boldsymbol{A} \in \mathbb{R}^{L \times R}$. More precisely, the observation noise is additive, zero-mean, Gaussian distributed with known covariance matrix $\boldsymbol{\Sigma}_0$, leading to

$$f(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Sigma}_0). \tag{1}$$

We address the problem of recovering $\boldsymbol{x}$ under the assumption that $\boldsymbol{x}$ is sparse. Note that if $\boldsymbol{x}$ was sparse in a specific basis, the equation above could be reformulated easily to estimate the sparse representation of the signal of interest. In the spectral unmixing context, the R columns of $\boldsymbol{A}$ represent spectral signatures of known materials referred to as endmembers, and the elements in $\boldsymbol{x}$ are the corresponding material abundances. The supervised spectral unmixing problem is typically an ill-posed problem, even if $R < L$, as the columns of $\boldsymbol{A}$ can be highly correlated. While enforcing the positivity of the elements of $\boldsymbol{x}$ regularises the spectral unmixing problem it has been shown that promoting sparse mixtures [9] can drastically improve the results. Here, we consider a sparsity-promoting prior model that factorize over the elements of $\boldsymbol{x}$, i.e.,

$$f(\boldsymbol{x}|\boldsymbol{\Theta}) = \prod_{r=1}^{R} f(x_r|\boldsymbol{\theta}_r) \tag{2}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_r\}_r$ is the set of hyper-parameters controlling the univariate priors $f(\cdot|\boldsymbol{\theta}_r)$. Classical choices include Laplace distributions, Student's t distributions, and (non-log-concave) spike-and-slab priors. Here, we consider a spike-and-slab prior that consists of a mixture of two zero-mean Gaussian distributions. To ensure the positivity of $\boldsymbol{x}$ when needed, truncated Gaussian distributions are considered. Note that allowing one of the variances to tend to 0, we can also obtain non-differentiable Bernoulli-Gaussian or Bernoulli-truncated-Gaussian prior. While ultimately imposing positivity constraints is necessary for spectral unmixing, relaxing these constraints makes it easier to compute exact posterior moments, at least for low-dimensional problems. The posterior distribution of $\boldsymbol{x}$ is then given by

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

Note that $\boldsymbol{\Theta}$ is assumed to be fixed. However, it could be estimated within the EP framework using by including the EP algorithm within a variational expectation-maximization as in [10]. The posterior in Eq. (3) is intractable as soon as it is no longer Gaussian or a Gaussian mixture and numerical tools are needed to approximate its (marginal) moments. In the next section, we summarize the proposed approach based on EP to approximate the mean of $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta})$ and its covariance matrix.

III. Expectation Propagation with diagonal+LR constraints

A. 2-factor EP main principles

The basic idea behind EP is to approximate the intractable posterior distribution $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta})$ by a tractable distribution $Q(\boldsymbol{x})$ that leverages the factorization of $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\Theta})$. As in [7], [8], we use two unnormalized multivariate Gaussian factors $q_1(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_1, \boldsymbol{S}_1)$ and $q_0(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}_0, \boldsymbol{S}_0)$ to approximate the likelihood $f(\boldsymbol{y}|\boldsymbol{x})$ and the prior $f(\boldsymbol{x})$, respectively. This leads to $Q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{m}, \boldsymbol{S}) \propto q_1(\boldsymbol{x}) q_0(\boldsymbol{x})$ with

$$\boldsymbol{S}^{-1} = \boldsymbol{S}_1^{-1} + \boldsymbol{S}_0^{-1} \tag{4}$$

$$\boldsymbol{S}^{-1}\boldsymbol{m} = \boldsymbol{S}_1^{-1}\boldsymbol{m}_1 + \boldsymbol{S}_0^{-1}\boldsymbol{m}_0. \tag{5}$$

Note that to simplify the notation, normalised and unnormalised distributions are not distinguished here. With this factorization, EP operates by solving iteratively the two following KL divergence minimization problems

$$\min_{q_0} KL\left(q_1(\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{\Theta}) || Q(\boldsymbol{x})\right), \tag{6a}$$

$$\min_{q_1} KL\left(f(\boldsymbol{y}|\boldsymbol{x})\, q_0(\boldsymbol{x}) || Q(\boldsymbol{x})\right), \tag{6b}$$

subject to constraints on q_0 and q_1.

In practice, since $Q(\boldsymbol{x})$ is Gaussian, these minimization problems reduces to matching the means and precision matrices of the first argument of the KL divergences (referred to as tilted distributions) to those of the second argument, subject to constraints. In [7], these two problems where simplified by imposing $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ to be diagonal, leading to $\boldsymbol{S}$ being a diagonal approximation of the posterior covariance matrix. As mentioned in the introduction, we refer to this approach as diagonal EP (D-EP). While not imposing constraints on $\boldsymbol{S}$ is possible within EP for small dimensional problems, this induces a large number of sequential updates (instead of only two here) and it requires storing and handling large covariance matrices. Instead of comparing our method to a real full-EP method (F-EP), we use as comparison and refer to as F-EP* the method proposed in [7] that runs D-EP and then keeps the full covariance of the tilted distribution during the last update of the approximate likelihood.

To address scalability issues and still capture part of the posterior dependencies between the elements of $\boldsymbol{x}$, we propose to identify low-rank (LR) structures in the posterior covariance matrix. The true prior considered here is separable so we used the same property for $q_0(\boldsymbol{x})$, i.e., $\boldsymbol{S}_0$ is forced to be diagonal (and positive definite (PD)). A posteriori correlation is induced by the operator $\boldsymbol{A}$ in the likelihood. We thus propose to force $\boldsymbol{S}_1$ to have an isotropic + LR structure, i.e., $\boldsymbol{S}_1 = \sigma^2\boldsymbol{I} + \boldsymbol{B}\boldsymbol{B}^T$, where $\boldsymbol{B}$ is an $R \times K$ matrix with $K \leq R$ and K a user-defined rank, typically chosen to be much smaller than R. This structural constraint fits well within EP, as $\boldsymbol{S}_1^{-1}$ can be expressed as a sum of diagonal and a low-rank matrix [11], which in turn leads to S having a diagonal + LR structure, since $\boldsymbol{S}_0$ is diagonal. It also ensures that $\boldsymbol{S}_1$ is PD. Note that a diagonal+LR structure could also be used for $\boldsymbol{S}_1$, however it can lead to numerical instabilities and we restricted ourselves to an isotropic + LR structure here. Although this structural assumption generalises the variational approximation of the true posterior compared to D-EP, the new constraints on $\boldsymbol{S}_1$

complicate the problems in Eq. (6). We now summarize how the two problems in Eq. (6) are solved.

B. Update of $q_0(\boldsymbol{x})$

Computing the mean and covariance of the tilted distribution $q_1(\boldsymbol{x})f(\boldsymbol{x}|\boldsymbol{\Theta})$ is either intractable or computationally expensive for high-dimensional problems. However since $\boldsymbol{S}_0$ is diagonal, its diagonal elements can be updated sequentially as in traditional, sequential EP. These R updates reduce to matching the moments of univariate tilted distributions and a univariate Gaussian. An alternative approach is to use parallel EP to update all the elements of $\boldsymbol{S}_0$ simultaneously, however it does not significantly reduce the overall computational time since R univariate cavity distributions still need to be computed. Note that a standard damping strategy [7], [12] is also used to prevent potential oscillations between successive iterations.

C. Update of $q_1(\boldsymbol{x})$

The main methodological contribution of this work is the update of the appproximate likelihood $q_1(\boldsymbol{x})$, which becomes challenging as R increases. Indeed the precision matrix of the cavity $f(\boldsymbol{y}|\boldsymbol{x})\,q_0(\boldsymbol{x})$ is given by $\boldsymbol{\Sigma}^{-1} = \boldsymbol{A}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{A}^T + \boldsymbol{S}_0^{-1}$, and its inversion can be prohibitively expensive for general matrices $\boldsymbol{A}$. Updating $\boldsymbol{m}_1$ once $\boldsymbol{S}_1$ has been updated is straightforward and not discussed here due to space constraints (see [10] for details). We now discuss the update of $\boldsymbol{S}_1$. It can be shown that this update reduces to solving

$$\min_{\sigma^2,\boldsymbol{B}} \quad \mathrm{tr}(\boldsymbol{S}^{-1}\boldsymbol{\Sigma}) + \log|\boldsymbol{S}|, \tag{7}$$

with $\boldsymbol{S}^{-1} = (\sigma^2\boldsymbol{I} + \boldsymbol{B}\boldsymbol{B}^T)^{-1} + \boldsymbol{S}_0^{-1}$. To the best of our knowledge, the problem above does not have a closed-form solution but it can be solved by a blockwise coordinate descent algorithm based on a log-determinant semidefinite program [11]. However, such an approach becomes expensive as R increases. First, instead of computing and storing the tilted covariance $\boldsymbol{\Sigma}$, it is approximated by generating N_s random samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_s}$ from $N(0, \boldsymbol{\Sigma})$ (which does not require the computation of $\boldsymbol{\Sigma}$ [13]) and by computing $\hat{\boldsymbol{\Sigma}}$, the sample covariance matrix of the samples. The new problem becomes

$$\min_{\sigma^2,\boldsymbol{B}} \quad C = \mathrm{tr}(\boldsymbol{S}^{-1}\hat{\boldsymbol{\Sigma}}) + \log|\boldsymbol{S}|, \tag{8}$$

subject to $\sigma^2 > 0$ and $\boldsymbol{B} \in \mathbb{R}^{R\times K}$. This problem is simpler to solve as it can be shown that C can be interpreted using the following Bayesian model

$$\begin{cases} \boldsymbol{x}_i & \sim & \mathcal{N}(0, D_0), \quad (i = 1, \ldots, N_s), i.i.d. \\ \boldsymbol{z}_i & \sim & \mathcal{N}(0, I), \quad (i = 1, \ldots, N_s), i.i.d. \\ \boldsymbol{u}_i|\boldsymbol{x}_i, \boldsymbol{z}_i & \sim & \mathcal{N}(\boldsymbol{x}_i + \boldsymbol{B}\boldsymbol{z}_i, \sigma^2\boldsymbol{I}), \quad (i = 1, \ldots, N_s), i.i.d. \end{cases}$$

Indeed, it turns out that minimizing the log-marginal posterior

$$\begin{aligned} \min_{\sigma^2,B} \quad L &= \sum_{i=1}^{N_s} \log\left(f(\boldsymbol{x}_i|\boldsymbol{u}_i = 0)\right) \\ &= \sum_{i=1}^{N_s} \log\left(\int f(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{u}_i = 0)\mathrm{d}\boldsymbol{z}_i\right) \end{aligned} \tag{9}$$

is equivalent to minimizing C. Instead of direct minimization of L, we use an EM-based method with $\{\boldsymbol{z}_i\}_i$ as latent variables. The resulting EM algorithm, not detailed here due to space constraints, is very similar to that used for probabilistic PCA [14] due to nature of the model $f(\boldsymbol{u}_i|\boldsymbol{x}_i, \boldsymbol{z}_i)$. The updates of $\boldsymbol{B}$ and σ^2 are performed with a gradient-descent method.

D. Overall EP method

The EP algorithm is initialised as follows: $(\boldsymbol{m}_0, \boldsymbol{S}_0)$ is set to match the prior mean and covariance. The algorithm then starts with the update of $q_1(\boldsymbol{x})$. Damping is applied during the update of $q_0(\boldsymbol{x})$ but not during the update of $q_1(\boldsymbol{x})$ to preserve the structure of $\boldsymbol{S}_1$. The algorithm iterates until the $(\boldsymbol{m}, \boldsymbol{S})$ converges or a maximum number of iterations is reached.

IV. Simulation and Results

In this section, we primarily assess the performance of the new algorithm, referred to as LR-EP algorithm, in terms of posterior mean and covariance estimation. First, we simulate data according to a Bayesian model whose posterior moments can be computed analytically, at least in low dimensions. We then investigate a sparse spectral unmixing problems where the posterior moments cannot be computed analytically.

A. Sparse Gaussian mixture models

We first explore the performance of our proposed algorithm with four different types of matrix $\boldsymbol{A}$:

a) i.i.d. Gaussian matrix, with $L = R = 10$,
b) a Toeplitz matrix whose first column is $[1, \ldots, L]^T$, with $L = R = 10$.
c) a quasi rank-1 matrix: $\boldsymbol{A} = \boldsymbol{r}\cdot\boldsymbol{1_R}^T + \beta\boldsymbol{A_0}$, where $\beta = 10^{-3}$, $\boldsymbol{r} \in \mathbb{R}^{L\times 1}$ and $\boldsymbol{A_0} \in \mathbb{R}^{\boldsymbol{L}\times\boldsymbol{R}}$ are i.i.d. standard Gaussian matrix, with $L = R = 10$.
d) Downsampled subset of the USGS 1995 spectral library used in [9] and in our unmixing experiments, with $L = R = 10$.
e) Same as d) but with $L = 100$ and $R = 10$.

The first two matrices are relatively well-conditioned and exhibit weak to no correlation structures. Conversely, other matrices are ill-conditioned and exhibit stronger correlation structure. As more spectral bands are added in e) compared to d) we expect stronger correlation to appear in the posterior distribution. We generated the non-zero elements of x from a standard normal distribution and set randomly 3 out of 10 elements to 0. We then simulated data with a signal-to-noise ratio (SNR) level of 30 dB. To build the exact and approximate Bayesian models, we use a Gaussian mixture priors, i.e., $f(\boldsymbol{x}_r|\boldsymbol{\theta}_r = \{v_0, v_1, \pi\}) = \pi\mathcal{N}(\boldsymbol{x}; 0, v_0\boldsymbol{I}) + (1-\pi)\mathcal{N}(\boldsymbol{x}; 0, v_1\boldsymbol{I})$, where $(v_0, v_1, \pi) = (1, 1e^{-3}, 0.73)$. This yields a true posterior that is a Gaussian mixture of 2^R components, allowing us to obtain closed-form expressions for its mean and covariance in the low-dimensional regime.

We compare the estimation accuracy of the posterior means and covariance matrices using EP with different covariance structure constraints (i.e., D-EP, LR-EP and F-EP*) and VAFC [15], as a VB equivalent of our LR-EP method. We set the

user-defined rank of VAFC is set to $K = 1$ and use $K = 1$ and $K = 5$ for LR-EP. The maximum number of iterations for all EP algorithms is 100. The posterior mean (resp. covariance) accuracy is measured using the relative root mean squared error (RMSE) (resp. the relative log-euclidean distance). The results are summarized in Tables I and II. The values provided are averaged over 10 realizations and the values in brackets indicate the corresponding standard deviations.

TABLE I
RELATIVE RMSE OF POSTERIOR MEAN FOR DIFFERENT METHODS

	D-EP	***LR-EP*** ($K = 1$)	***LR-EP*** ($K = 5$)	***F-EP****	***VAFC*** ($K = 1$)
a	**0.0098** (0.0094)	0.0110 (0.0105)	<u>0.0106</u> (0.0104)	**0.0098** (0.0092)	0.1890 (0.0266)
b	0.0164 (0.0289)	**0.0079** (0.0035)	<u>0.0083</u> (0.0045)	0.0293 (0.0604)	0.2997 (0.0367)
c	0.0426 (0.0188)	**0.0152** (0.0027)	**0.0152** (0.0031)	<u>0.0255</u> (0.0318)	0.4040 (0.1297)
d	0.2560 (0.4740)	<u>0.0222</u> (0.0053)	**0.0207** (0.0048)	0.0798 (0.1128)	0.2910 (0.0407)
e	7.0×10^{-5} (4.5×10^{-5})	2.0×10^{-5} (8.3×10^{-6})	<u>1.8×10^{-5}</u> (1.0×10^{-5})	$\mathbf{1.0\times10^{-5}}$ (1.5×10^{-5})	0.4513 (0.1947)

Remarks. The best (resp. second best) results are bold (resp. underlined), which also applies to the other tables.

TABLE II
RELATIVE LOG-EUCLIDEAN ERROR OF POSTERIOR COVARIANCE FOR DIFFERENT METHODS

Types	***D-EP***	***LR-EP*** ($K = 1$)	***LR-EP*** ($K = 5$)	***F-EP****	***VAFC*** ($K = 1$)
a	0.3718 (0.0291)	0.3357 (0.0308)	<u>0.3345</u> (0.0325)	**0.3155** (0.0308)	0.7989 (0.0292)
b	**0.4645** (0.0182)	0.4854 (0.0203)	<u>0.4782</u> (0.0191)	0.6637 (0.0208)	1.3711 (0.0790)
c	<u>0.4185</u> (0.0929)	**0.7006** (0.1107)	0.7017 (0.1120)	8.7330 (0.8792)	22.2471 (2.2678)
d	<u>0.5458</u> (0.0348)	0.6874 (0.0695)	**0.6834** (0.0812)	0.9496 (0.0545)	2.1849 (0.0791)
e	0.1808 (0.0038)	**0.1030** (0.0144)	<u>0.1031</u> (0.0145)	0.2302 (0.0360)	0.4171 (0.0751)

Overall in Tables I and II, VAFC is less accurate than EP in estimating both the mean and marginal variance, probably because of the mode-seeking behavior. It should also be noted that we observed that VAFC can converge slowly and convergence might be difficult to assess. Among the EP algorithms, LR-EP is the most flexible model. It is expected to give better results than D-EP or F-EP*. The LR-EP results are at least in par with those of D-EP, e.g., for case a), and generally better than D-EP for cases b)-d). For e), all the EP methods estimate very well the posterior means and LR-EP provides better covariance estimates. It should be noted that we observed that the performance of D-EP, F-EP* and, to a lesser extent LR-EP, depends on the amount of damping applied to stabilize EP. LR-EP seems overall more stable than D-EP when the posterior exhibits strong correlation. Fig. 1 shows examples of covariance matrices estimated via EP. Using $K = 1$, LR-EP is able to capture most of the dependencies between covariates, while D-EP only captures marginal variances.

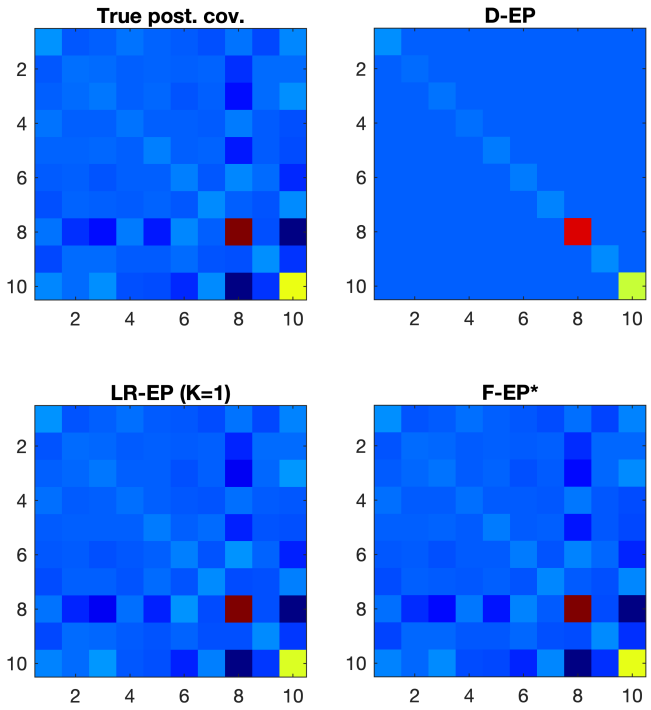

Fig. 1. Example of true and estimated covariance matrices for the matrix of case e) (spectral library). All the subplots share the same color scale.

B. Sparse spectral unmixing

We now investigate the harder, sparse unmixing problem and assess our LR-EP method for this higher-dimensional problem where $\boldsymbol{x}$ is subject to positivity constraints a priori. We selected $R = 50$ signatures from a pruned USGS 1995 library $\boldsymbol{A}$ with $L = 224$ spectral channels by selecting signatures that are relatively similar (the minimum and maximum spectral angle between columns of $\boldsymbol{A}$ are $4.44°$ and $52.79°$ respectively) (see Fig. 2). The abundances of the first 5 endmembers are generated independently according to a truncated Gaussian distribution with hidden mean and variance equal to 0 and 1, respectively. The remaining 45 abundances are set to 0, resulting in sparse mixtures. In these experiments we use mixtures of truncated Gaussian priors, i.e., $f(\boldsymbol{x}_r|\boldsymbol{\theta}_r = \{v_0, v_1, \pi\}) = \pi\mathcal{N}_+(\boldsymbol{x}; 0, v_0\boldsymbol{I}) + (1-\pi)\mathcal{N}_+(\boldsymbol{x}; 0, v_1\boldsymbol{I})$ and set $(v_0, v_1, \pi) = (1, 1e^{-3}, 0.12)$.

In addition to the competing methods already introduced, we also use two classical pixel-wise linear unmixing methods: the non-negative least squares (NNLS) algorithm [16], based on constrained maximum likelihood estimation, and SUnSAL [9], which relies on MAP estimation using an exponential abundance prior (resulting in an ℓ_1-norm based penalty). The regularization parameter of SUnSAL was tuned by cross-validation to minimize the average abundance estimation errors. VAFC is not considered as it was not providing satisfactory results in the previous experiments.

Due to the additional positivity constraints, it is no longer possible to compute exact posterior moments so we use the relative RMSE between the actual and estimated abundance vectors as main metric (see Table III). It should be noted

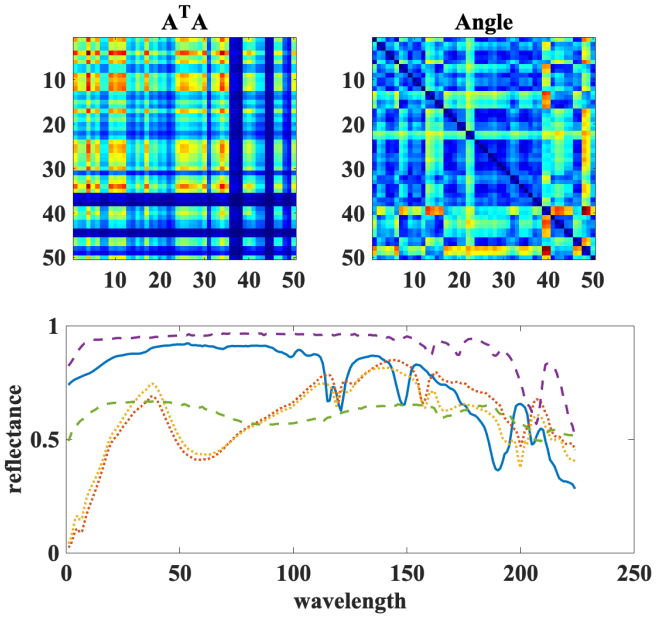

Fig. 2. The strong correlations between the endmembers in $\boldsymbol{A}$ pose a challenge for abundance estimation. The bottom plot depicts, the five active endmembers of the pruned USGS 1995 library.

that adding the positivity constraints here makes the posteriors highly non-Gaussian, which can cause stability issues during the EP updates. This is particularly visible as the SNR increases, and the EP standard deviations in Table III increase. For the lowest SNR, LR-EP provides slightly better results than SUnSAL and D-EP, which illustrates the potential benefits of EP and variational inference for problems with large uncertainties. Additional work will however be required to investigate the performance degradation for high SNRs. It is reasonable to assume that as the SNR increases, the non-Gaussian modes of the posterior distribution become more separated and this makes local convergence of EP more likely, irrespective of the structural constraints applied to the approximating factors. It would be interesting to assess if damping the likelihood update or improving the prior update e.g., using stochastic EP [15] could improve the LR-EP results. Another interesting idea would be to use annealing for the likelihood factor.

TABLE III
RELATIVE RMSE OF POSTERIOR MEAN WITH DIFFERENT SNR LEVELS WITH THE EXTENDED SPECTRAL LIBRARY

SNR	***10dB***	***30dB***
NNLS	0.1774 (0.0745)	0.0706 (0.0334)
SUnSAL	0.1534 (0.0427)	**0.0573** (0.0287)
D-EP	0.1426 (0.0077)	0.2461 (0.2422)
LR-EP (K=5)	**0.1326** (0.0168)	0.1132 (0.0776)
LR-EP (K=10)	0.1354 (0.0179)	0.2601 (0.2384)
F-EP*	0.1405 (0.0120)	0.2672 (0.4009)

V. CONCLUSIONS

In this paper, we investigated the use of low-rank structures with the Expectation-Propagation framework for linear inverse problems. Using such structures, we showed that in contrast to most scalable methods relying on (local) mean-field approximations, EP can capture partially correlations. Moreover, in the examples where FFVB could be applied (i.e., Section IV-A), EP performed better. Still, EP does not benefit from general convergence guarantees and can either 1) provide poor estimates or 2) suffer from stability issues if the structure of the approximating factors is not chosen carefully. Those issues seem to be more frequent when the target distributions become highly non-Gaussian and when using Gaussian approximations becomes inappropriate. Future work include the investigation of more robust EP-like inference schemes, such as stochastic or average EP, and the consideration of more flexible covariance constraints for multidimensional problems.

REFERENCES

[1] G. O. Roberts and R. L. Tweedie, "Exponential convergence of Langevin distributions and their discrete approximations," *Bernoulli*, vol. 2, no. 4, pp. 341–363, 1996.
[2] R. M. Neal, "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011, pp. 113–160.
[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] T. Salimans and D. A. Knowles, "Fixed-form variational posterior approximation through stochastic linear regression," *Bayesian Analysis*, vol. 8, no. 4, pp. 837–882, 2013.
[5] M.-N. Tran, T.-N. Nguyen, and V.-H. Dao, "A practical tutorial on variational Bayes," *arXiv:2103.01327 [stat.CO]*, 2021.
[6] T. Minka, "Power EP," Microsoft Research, Cambridge, Tech. Rep. MSR-TR-2004-149, 2004.
[7] 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.
[8] Z. Li, Y. Altmann, J. Chen, S. Mclaughlin, and S. Rahardja, "Sparse linear spectral unmixing of hyperspectral images using expectation-propagation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.
[9] J. M. Bioucas-Dias and M. A. T. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2010, pp. 1–4.
[10] D. Yao, S. McLaughlin, and Y. Altmann, "Fast scalable image restoration using total variation priors and expectation propagation," *IEEE Transactions on Image Processing*, vol. 31, pp. 5762–5773, 2022.
[11] Y. Wu, Y. Qin, and M. Zhu, "High-dimensional covariance matrix estimation using a low-rank and diagonal decomposition," *Canadian Journal of Statistics*, vol. 48, no. 2, pp. 308–337, 2020.
[12] A. Vehtari, A. Gelman, T. Sivula, P. Jylänki, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, and C. P. Robert, "Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 577–629, 2020.
[13] 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.
[14] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.
[15] Y. Li, J. M. Hernández-Lobato, and R. E. Turner, "Stochastic expectation propagation," *Advances in neural information processing systems*, vol. 28, 2015.
[16] D. Heinz and Chein-I-Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 529–545, 2001.