SAVED - Space Alternating Variational Estimation for Sparse Bayesian Learning with Parametric Dictionaries

Christo Kurisummoottil Thomas, Dirk Slock,
EURECOM, Sophia-Antipolis, France, Email: {kurisumm,slock}@eurecom.fr

***Abstract*—In this paper, we address the fundamental problem of sparse signal recovery in a Bayesian framework, where the received signal is a multi-dimensional tensor. We further consider the problem of dictionary learning, where the tensor observations are assumed to be generated from a Khatri-Rao structured dictionary matrix multiplied by the sparse coefficients. We consider a Bayesian approach using variational Bayesian (VB) inference. VB allows one to obtain analytical approximations to the posterior distributions of interest even when an exact inference of these distributions is intractable. We propose a novel fast algorithm called space alternating variational estimation with dictionary learning (SAVED), which is a version of VB(-SBL) pushed to the scalar level. Similarly, as for SAGE (space-alternating generalized expectation maximization) compared to EM, the component-wise approach of SAVED compared to sparse Bayesian learning (SBL) renders it less likely to get stuck in bad local optima and its inherent damping (more cautious progression) also leads to typically faster convergence of the non-convex optimization process. Simulation results show that the proposed algorithm has a faster convergence rate and lower mean squared error (MSE) compared to the alternating least squares based method for tensor decomposition.**

***Keywords*—** Sparse Bayesian Learning, Variational Bayes, Tensor Decomposition, Dictionary Learning, Alternating Optimization

I. Introduction

In many applications such as Multiple Input Multiple Output (MIMO) radar [1], massive MIMO channel estimation [2], image and video processing etc., the received signals are multidimensional (i.e tensors). Moreover, these signals can be represented as a low rank tensor. To fully exploit the structure of such signals, tensor decomposition methods such as CANDECOMP/PARAFAC (CP) [3], [4] or Canonical Polyadic Decomposition (CPD) [5] have been introduced. Explicitly accounting for this tensorial structure can be more beneficial than the matricized or vectorized representations of the data, since the matrix decomposition cannot fully exploit the multi-dimensional subspace structure of the data. In this paper, we consider a generalized problem in which the dictionary matrix can be factorized as a tensor product, which can be formulated as

$$\mathbf{Y} = \sum_{i=1}^{M} x_i \mathbf{A}_{1,i} \circ \mathbf{A}_{2,i} ... \circ \mathbf{A}_{N,i} + \mathbf{w}, \tag{1}$$

where $\circ$ represents the Hadamard product between two matrices, $\mathbf{Y} \in \mathcal{C}^{I_1 \times I_2 ... \times I_N}$ is the observations or data, $\mathbf{A}_{j,i} \in \mathcal{C}^{I_j}, \mathbf{A}_j = [\mathbf{A}_{j,1}, ..., \mathbf{A}_{j,M}]$ and the tensor is represented by $[\![\mathbf{A}_1, ..., \mathbf{A}_N; \mathbf{x}]\!]$ is called the measurement or the sensing or the dictionary matrix which is unknown, $\mathbf{x}$ is the M-dimensional sparse signal and $\mathbf{w}$ is the additive noise. $\mathbf{x}$ contains only K non-zero entries, with $K << M$ and thus the dictionary matrix to be learned allows a low rank representation. $\mathbf{w}$ is assumed to be a white Gaussian noise, $\mathbf{w} \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$. To address this problem when the dictionary matrix is known, a variety of algorithms such as the orthogonal matching pursuit [6], the basis pursuit method [7] and the iterative re-weighted l_1 and l_2 algorithms [8] exist in the literature. In a Bayesian setting, the aim is to calculate the posterior distribution of the parameters given some observations (data) and some a priori knowledge. The Sparse Bayesian Learning algorithm was first introduced by [9] and then proposed for the first time for sparse signal recovery by [10]. The SBL is developed around a sparsity-promoting prior for $\mathbf{x}$ (with precision parameter $\boldsymbol{\alpha}$), whose realizations are softly sparse in a sense that most entries are small in magnitude and close to zero.

To find the tensor factor matrices, the most popular solution is the alternating least squares method (ALS) [11], which iteratively optimizes one factor matrix at a time while keeping the others fixed. However, knowledge of tensor rank is a prerequisite to implement this algorithm and it takes large number of iterations to converge. Moreover, classical algorithms ignore the potential statistical knowledge of the factor matrices into account. In the literature, most of the existing works focus on estimation of parameterized dictionary. Most common approach is to reduce the parameter space to a fixed grid of points, thus restricting the solution space. [12] proposes off-grid sparse decomposition problem where the dictionary columns are specified by unknown continuous-valued parameters.

In the empirical Bayesian approach, an estimate of the hyper parameters $\boldsymbol{\alpha}, \gamma$ and sparse signal $\mathbf{x}$ is performed iteratively using Type II maximum likelihood method [13]. Recently approximate message passing (AMP) [14], generalized AMP and vector AMP [15]–[17], [17], [18] were introduced to compute the posterior distributions in a message passing framework and with less complexity. But it suffers from the limitation that only for i.i.d Gaussian $\mathbf{A}$, the algorithm is guaranteed to converge. An alternative approach to SBL is using variational approximation for Bayesian inference [19], [20]. Variational Bayesian (VB) inference tries to find an approximation of the posterior distribution which maximizes the variational lower bound on $\ln p(\mathbf{y})$. [21] introduces a Fast version of SBL by alternatingly maximizing the variational posterior lower bound with respect to single (hyper)parameters.

A. Contributions of this paper

In this paper:

- We propose a novel Space Alternating Variational Estimation based SBL technique with dictionary learning called SAVED compared to our previous works SAVE [22]–[24] which was for a known dictionary matrix case.
- Numerical results suggest that our proposed solution has a faster convergence rate (and hence lower complexity) than (even) the existing fast SBL and performs better than the existing fast SBL algorithms in terms of reconstruction error in the presence of noise.

In the following, boldface lower-case and upper-case characters denote vectors and matrices respectively. The operators $tr(\cdot)$, $(\cdot)^T$, $(\cdot)^*$, $(\cdot)^H$ represents trace, transpose, conjugate and conjugate transpose respectively. A real Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Theta}$ is distributed as $x \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Theta})$. $diag(\cdot)$ represents the diagonal matrix created by elements of a row or column vector. The operator $< x >$ or $\mathrm{E}(\cdot)$ represents the expectation of x. $\|\cdot\|$ represents the Frobenius norm. All the random variables are complex here unless specified otherwise. $\mathbf{A}{\odot}\mathbf{B}$ and $\mathbf{A}{\circ}\mathbf{B}$ represents the Khatri-Rao product and Hadamard product between the matrices $\mathbf{A}, \mathbf{B}$ respectively. We represent $\bigodot\limits_{j=1}^{N} \mathbf{A}_j = \mathbf{A}_1 \odot \mathbf{A}_2 \odot ... \odot \mathbf{A}_N$,We represent $\bigcirc\limits_{j=1}^{N} \mathbf{A}_j = \mathbf{A}_1 \circ \mathbf{A}_2 \circ ... \circ \mathbf{A}_N$, where $\mathbf{A}_j \in \mathcal{C}^{I_j \times 1}$.

II. Hierarchical Probabilistic Model

In the following sections, we represent (1) using the tensor decomposition properties from [11]. Let $Y_{i_1,...,i_N}$ represents the $(i_1, i_2, ..., i_N)^{th}$ element of the tensor and $\mathbf{y} = [y_{1,1,...,1}, y_{1,1,...,2}...., y_{I_1,I_2,...,I_N}]^T$, then it can be verified that [25],

$$\mathbf{y} = (\mathbf{A}_1 \odot \mathbf{A}_2 ... \odot \mathbf{A}_N)\mathbf{x} + \mathbf{w}, \tag{2}$$

where $\odot$ represents the Khatri-Rao product between two matrices, $\mathbf{y} \in \mathcal{C}^{(\prod\limits_{i=1}^{N} I_i) \times 1}$ and we denote $\mathbf{A} = \mathbf{A}_1 \odot \mathbf{A}_2 ... \odot \mathbf{A}_N$. Since the sparsity measure (number of nonzero components) of $\mathbf{x}$ is unknown, the following VB-SBL algorithm performs automatic rank determination. In Bayesian compressive sensing, a two-layer hierarchical prior is assumed for the $\mathbf{x}$ as in [9]. The hierarchical prior is chosen such that it encourages the sparsity property of $\mathbf{x}$. $\mathbf{x}$ is assumed to have a Gaussian distribution parameterized by $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ ... \ \alpha_M]$, $\alpha_i > 0$ and real, where α_i represents the inverse variance or the precision parameter of x_i.

$$p(\mathbf{x}/\boldsymbol{\alpha}) = \prod_{i=1}^{M} p(x_i/\alpha_i) = \prod_{i=1}^{M} \mathcal{CN}(0, \alpha_i^{-1}). \tag{3}$$

Further a Gamma prior is considered over $\boldsymbol{\alpha}$,

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{M} p(\alpha_i/a, b) = \prod_{i=1}^{M} \Gamma^{-1}(a) b^a \alpha_i^{a-1} e^{-b\alpha_i}. \tag{4}$$

The inverse of noise variance, $\gamma > 0$ and real, is also assumed to have a Gamma prior, $p(\gamma) = \Gamma^{-1}(c) d^c \alpha_i^{c-1} e^{-d\gamma}$. Now the likelihood distribution can be written as,

$$p(\mathbf{y}/\mathbf{x}, \gamma) = (2\pi)^{-N} \gamma^N e^{-\gamma ||\mathbf{y} - \mathbf{A}\mathbf{x}||^2}. \tag{5}$$

Note that no prior is assumed for $\mathbf{A}_{j,i}$ (deterministic). Further we consider certain factor matrices to be structured and they are considered to be Vandermonde matrices. So we write,

$$\mathbf{A}_{j,i} = [1, e^{j\phi_{j,i}}, ..., e^{j(I_j - 1)\phi_{j,i}}]^T, \ \forall j = 1, ..., S, \tag{6}$$

where $\phi_{j,i}$ represents the spatial frequency and $\mathbf{A}_j, \forall j = S+1, ..., N$ are considered to be unstructured, with $S <= N$. We consider certain factors to be unstructured because the parametric forms are uncertain. For eg. in massive MIMO channel estimation [26], the array response at the mobile station (MS) is not exploitable. Even the array response at the base station (BS) will typically require calibration to be exploitable. Doppler shifts are clear Vandermonde vectors. Delays could be more or less clear, if one goes to frequency domain in OFDM, and one only takes into account the range of subcarriers for which the Tx/Rx filters can be considered f-flat. Then over those subcarriers, it's also Vandermonde. Let $\mathbf{A}_{j,i}$ represents the i^{th} column of $\mathbf{A}_j$. For the unstructured factor matrices also, we consider $\mathbf{A}_{j,i} = [1 \, \mathbf{a}_{j,i}^H]^H$ and further $\mathbf{a}_{j,i}$ is unconstrained and deterministic (in all the Vandermonde cases, it is perfect, or in all cases of phasors). Assuming first entry to be 1 is even better than $\|\mathbf{A}_{j,i}\| = 1$ because $\|\mathbf{A}_{j,i}\| = 1$ still leaves a phase ambiguity. With first entry$= 1$, the factors are unique, up to permutation in the sum of terms of course.

We define the unfolding operation on an N^{th} order tensor $\mathbf{Y} = [\![\mathbf{A}_1, ..., \mathbf{A}_N; \mathbf{x}]\!]$ as [11] ($\mathbf{Y}^{(n)}$ is of size $I_n \times \prod\limits_{i=1, i \neq n}^{N} I_i$ below, $\mathbf{X} = \mathrm{diag}\,(\mathbf{x})$),

$$\mathbf{Y}^{(n)} = \mathbf{A}_n \mathbf{X} (\mathbf{A}_N \odot \mathbf{A}_{N-1} ... \mathbf{A}_{n+1} \odot \mathbf{A}_{n-1} ... \odot \mathbf{A}_1)^T. \tag{7}$$

A. Variational Bayesian Inference

The computation of the posterior distribution of the parameters is usually intractable. In order to address this issue, in variational Bayesian framework, the posterior distribution $p(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}/\mathbf{y})$ is approximated by a variational distribution $q(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A})$ that has the factorized form:

$$q(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}) = q_\gamma(\gamma) \prod_{i=1}^{M} q_{x_i}(x_i) \prod_{i=1}^{M} q_{\alpha_i}(\alpha_i) \prod_{i=1}^{M} \prod_{j=1}^{N} q_{\mathbf{a}_{j,i}}(\mathbf{a}_{j,i}) \tag{8}$$

Variational Bayes compute the factors q by minimizing the Kullback-Leibler distance between the true posterior distribution $p(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}/\mathbf{y})$ and the $q(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A})$. From [19],

$$KLD_{VB} = KL\left(p(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}/\mathbf{y}) || q(\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A})\right) \tag{9}$$

The KL divergence minimization is equivalent to maximizing the evidence lower bound (ELBO) [20]. To elaborate on this, we can write the marginal probability of the observed data as,

$$\ln p(\mathbf{y}) = L(q) + KLD_{VB}, \text{ where,}$$
$$L(q) = \int q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}, \;\; KLD_{VB} = -\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}/\mathbf{y})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{10}$$

where $\boldsymbol{\theta} = \{\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}\}$ and θ_i represents each independent factor in $\boldsymbol{\theta}$. Since $KLD_{VB} \geq 0$, it implies that $L(q)$ is a lower bound on $\ln p(\mathbf{y})$. Moreover, $\ln p(\mathbf{y})$ is independent of $q(\boldsymbol{\theta})$ and therefore maximizing $L(q)$ is equivalent to minimizing KLD_{VB}. This is called as ELBO maximization and doing this in an alternating fashion for each variable in $\boldsymbol{\theta}$ leads to,

$$\begin{aligned} &\ln(q_i(\theta_i)) =< \ln p(\mathbf{y},\boldsymbol{\theta}) >_{k\neq i} + c_i, \\ &p(\mathbf{y},\boldsymbol{\theta}) = p(\mathbf{y}/\mathbf{x},\boldsymbol{\alpha},\gamma) p(\mathbf{x}/\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\gamma). \end{aligned} \tag{11}$$

Here $<>_{k\neq i}$ represents the expectation operator over the distributions q_k for all $k \neq i$.

III. SAVED Sparse Bayesian Learning

In this section, we propose a Space Alternating Variational Estimation with dictionary learning (SAVED) based alternating optimization between each elements of $\boldsymbol{\theta}$. For SAVED, not any particular structure of $\mathbf{A}$ is assumed, in contrast to AMP which performs poorly when $\mathbf{A}$ is not i.i.d or sub-Gaussian. Based on a quadratic loss function, the Bayesian estimator of a parameter is the posterior mean; we therefore define the variational Bayesian estimators of parameters $\boldsymbol{\theta}$ as the means of the variational approximation to the posterior distribution. The joint distribution can be written as,

$$\begin{aligned} &\ln p(\mathbf{y},\boldsymbol{\theta}) = N \ln\gamma - \gamma\, ||\mathbf{y} - \mathbf{A}\mathbf{x}|\,|^2 + \\ &\sum_{i=1}^{M} \left(\ln\alpha_i - \alpha_i |x_i|^2\right) + \sum_{i=1}^{M} \left((a-1)\ln\alpha_i + a\ln b - b\alpha_i\right) \\ &+ (c-1)\ln\gamma + c\ln d - d\gamma + \text{constants}, \end{aligned} \tag{12}$$

In the following, $c_{x_i}, c'_{x_i}, c_{\alpha_i}, c_{a_{ji}}$ and c_γ represents normalization constants for the respective pdfs.

Update of $q_{x_i}(x_i)$**:** Using (11), $\ln q_{x_i}(x_i)$ turns out to be quadratic in x_i and thus can be represented as a Gaussian distribution as follows,

$$\begin{aligned} &\ln q_{x_i}(x_i) = - < \gamma > \Big\{ < \|\mathbf{y} - (\bigodot_{j=1}^{N} \mathbf{A}_{j,\bar{i}})\mathbf{x}_{\bar{i}}\|^2 > - \\ &(\mathbf{y} - < (\bigodot_{j=1}^{N} \mathbf{A}_{j,\bar{i}}) >< \mathbf{x}_{\bar{i}} >)^H < (\bigodot_{j=1}^{N} \mathbf{A}_{j,i}) > x_i - \\ &x_i^H < (\bigodot_{j=1}^{N} \mathbf{A}_{j,i})^H > (\mathbf{y} - < (\bigodot_{j=1}^{N} \mathbf{A}_{j,\bar{i}}) >< \mathbf{x}_{\bar{i}} >) + \\ &< (\bigodot_{j=1}^{N} \mathbf{A}_{j,i})^H (\bigodot_{j=1}^{N} \mathbf{A}_{j,i}) > |x_i|^2 \Big\} - < \alpha_i > |x_i|^2 + c_{x_i} \\ &= -\tfrac{1}{\sigma_i^2} \left(x_i - \widehat{x}_i\right)^2 + c'_{x_i}. \end{aligned} \tag{13}$$

Note that we split $\mathbf{A}\mathbf{x}$ as, $\mathbf{A}\mathbf{x} = (\bigodot_{j=1}^{N} \mathbf{A}_{j,i}) x_i + (\bigodot_{j=1}^{N} \mathbf{A}_{j,\bar{i}}) \mathbf{x}_{\bar{i}}$, where $(\bigodot_{j=1}^{N} \mathbf{A}_{j,i})$ represents the i^{th} column of $\mathbf{A}$, $(\bigodot_{j=1}^{N} \mathbf{A}_{j,\bar{i}})$ represents the matrix with i^{th} column of $\mathbf{A}$ removed, x_i is the i^{th} element of $\mathbf{x}$, and $\mathbf{x}_{\bar{i}}$ is the vector without x_i. $\mathbf{A}_{j,\bar{i}}$ represents $\mathbf{A}_j$ with i^{th} column of $\mathbf{A}_j$ removed. Using the property of the Khatri-Rao products that $(\bigodot_{j=1}^{N} \mathbf{A}_j)^H (\bigodot_{j=1}^{N} \mathbf{A}_j) = \bigcirc_{j=1}^{N} \mathbf{A}_j^H \mathbf{A}_j$ derived in [27, Lemma 1], we simplify $(\bigodot_{j=1}^{N} \mathbf{A}_{j,i})^H (\bigodot_{j=1}^{N} \mathbf{A}_{j,i}) = (\bigcirc_{j=1}^{N} \mathbf{A}_{j,i}^H \mathbf{A}_{j,i}) = \prod_{j=1}^{N} \|\mathbf{A}_{j,i}\|^2$, $(\bigodot_{j=1}^{N} \mathbf{A}_{j,i})^H (\bigodot_{j=1}^{N} \mathbf{A}_{j,\bar{i}}) = (\bigcirc_{j=1}^{N} \mathbf{A}_{j,i}^H \mathbf{A}_{j,\bar{i}})$. Clearly, the mean and the variance of the resulting Gaussian distribution becomes,

$$\begin{aligned} \sigma_i^2 &= \frac{1}{<\gamma> \prod_{j=1}^{N} <\|\mathbf{A}_{j,i}\|^2> + <\alpha_i>}, \;\; < x_i >= \\ \widehat{x}_i &= \sigma_i^2 ((\bigodot_{j=1}^{N} \widehat{\mathbf{A}}_{j,i})^H \mathbf{y} - (\bigcirc_{j=1}^{N} \widehat{\mathbf{A}}_{j,i}^H \widehat{\mathbf{A}}_{j,\bar{i}}) < \mathbf{x}_{\bar{i}} >) < \gamma >, \end{aligned} \tag{14}$$

where $\widehat{x}_i$ represents the point estimate of x_i and $\widehat{\mathbf{A}}_{j,i} = [1 < \mathbf{a}_{j,i}^H >]^H$, $< \mathbf{a}_{j,i} >$ being the mean of $\mathbf{a}_{j,i}$ which follows from the below derivation for $\mathbf{a}_{j,i}$.

Update of $q_{\mathbf{a}_{j,i}}(\mathbf{a}_{j,i})$**:** For convenience of the analysis, we define the following terms. $\mathbf{X} = \operatorname{diag}\mathbf{x}$, $\bigodot_{k=N,k\neq j}^{1} \mathbf{A}_k = \mathbf{A}_N \odot \mathbf{A}_{N-1} ... \mathbf{A}_{j+1} \odot \mathbf{A}_{j-1} ... \odot \mathbf{A}_1, \mathbf{V}_j =< \mathbf{X} > (\bigodot_{k=N,k\neq j}^{1} < \mathbf{A}_k >)^T, \mathbf{W}_j =< \mathbf{X} (\bigodot_{k=N,k\neq j}^{1} \mathbf{A}_k)^T (\bigodot_{k=N,k\neq j}^{1} \mathbf{A}_k)^* \mathbf{X}^H >$. We go back to the tensorial representation to derive $q_{\mathbf{a}_{j,i}}(\mathbf{a}_{j,i})$. The variational approximation leads to the following Gaussian distribution for the vector $\mathbf{a}_{j,i}$,

$$\begin{aligned} &\ln q_{\mathbf{a}_{j,i}}(\mathbf{a}_{j,i}) = - < \gamma > \|\mathbf{Y} - [\![\mathbf{A}_1, ..., \mathbf{A}_N; \mathbf{x}]\!]\|^2 = - \\ &< \gamma > \operatorname{tr}\{-\mathbf{Y}^{(j)} \mathbf{V}_j^H \mathbf{A}_j^H + \mathbf{A}_j \mathbf{V}_j \mathbf{Y}^{(j)} + \mathbf{A}_j \mathbf{W}_j \mathbf{A}_j^H\} + c_{\mathbf{a}_{ji}} \end{aligned} \tag{15}$$

We used here the fact that [11] $\|\mathbf{A}\|^2 = \operatorname{tr}\{\mathbf{A}^{(k)} (\mathbf{A}^{(k)})^H\}$ for a tensor $\mathbf{A}$. $\mathbf{A}_j \mathbf{W}_j \mathbf{A}_j^H$ can be written as, $\operatorname{tr}\{\mathbf{A}_{j,i} \mathbf{W}_j \mathbf{A}_{j,i}^H\} +$ “terms independent of $\mathbf{a}_{j,i}$”, which gets simplified as $\operatorname{tr}\{\mathbf{W}_j\} \|\mathbf{a}_{j,i}\|^2 +$ “terms independent of $\mathbf{a}_{j,i}$”. Finally, the mean and variance of the resulting Gaussian distribution can be written as,

$$\begin{aligned} &< \mathbf{a}_{j,i} >= \widehat{\mathbf{a}}_{ji} = (\mathbf{b_j})_{\bar{\mathbf{1}}}, \;\; \mathbf{b_j} = \mathbf{Y^{(j)}} (\bigodot_{\mathbf{k=N,k\neq j}}^{\mathbf{1}} < \mathbf{A_k} >)^* \mathbf{c_j}, \\ &\boldsymbol{\Upsilon}_{j,i} = \beta_{j,i} \mathbf{I}, \;\; \beta_{j,i} = \operatorname{tr}\{< \mathbf{X} (\bigcirc_{k=1,k\neq j}^{N} \mathbf{A}_k^T \mathbf{A}_k^*) \mathbf{X}^H >\} \end{aligned} \tag{16}$$

where $\mathbf{c}_j = [0, .., 0, x_i, 0, ..0]^H$ represents the vector with all zeros except at i^{th} location. Also, we can write $< \|\mathbf{A}_{j,i}\|^2 >= 1 + \|\widehat{\mathbf{a}}_{j,i}\|^2 + \beta_{j,i}(I_j - 1)$, which gets used in (14). For computing $\beta_{j,i}$, we use the independence assumption of the variational distributions q of $x_i, \mathbf{A}_k$. So terms of the form $< \mathbf{A}_{k,i}^T \mathbf{A}_{k,i}^* >= \|\mathbf{A}_{k,i}\|^2$ and $< \mathbf{A}_{k,i}^T \mathbf{A}_{k,i}^* >=< \mathbf{A}_{k,i}^T >< \mathbf{A}_{k,j}^* >, < x_i x_j^* >=< x_i >< x_j^* >, \forall i \neq j$.

Update of $q_{\alpha_i}(\alpha_i), q_\gamma(\gamma)$**:** The variational approximation leads to a Gamma distribution for the $q_{\alpha_i}(\alpha_i), q_\gamma(\gamma)$. The detailed derivations can be found in our paper [22]. They are parameterized by just one quantity which is the mean of the Gamma distribution, given by,

$$< \alpha_i >= \frac{a+\frac{1}{2}}{(<|x_i|^2>+b)}, \text{ where } < |x_i|^2 >= |\widehat{x}_i|^2 + \sigma_i^2.$$
$$< \gamma >= \frac{c+\frac{N}{2}}{(<\left\|\mathbf{y} - (\bigodot_{j=1}^{N} \mathbf{A}_j)\mathbf{x}\right\|^2 >+d)}, \tag{17}$$

where, $< \|\mathbf{y} - (\bigodot_{j=1}^{N} \mathbf{A}_j)\mathbf{x}\|^2 >= \|\mathbf{y}\|^2 - 2\mathbf{y}^H(\bigodot_{j=1}^{N} < \widehat{\mathbf{A}}_j >)\widehat{\mathbf{x}} + \text{tr}((\bigcirc_{j=1}^{N} < \mathbf{A}_j^H \mathbf{A}_j >)(\widehat{\mathbf{x}}\widehat{\mathbf{x}}^H + \mathbf{\Sigma}))$, $\mathbf{\Sigma} = \text{diag}\,(\sigma_1^2, \sigma_2^2, ..., \sigma_M^2)$, $\widehat{\mathbf{x}} = [\widehat{x}_1, \widehat{x}_2, ..., \widehat{x}_M]^H$. From (14), it can be seen that the estimate $\widehat{\mathbf{x}}$ converges to the L-MMSE equalizer, $\widehat{\mathbf{x}} = (\mathbf{A}^H\mathbf{A} + \frac{1}{<\gamma>}\mathbf{\Sigma}^{-1})^{-1}\mathbf{A}^H\mathbf{y}$.

A. Joint VB for $\mathbf{A}_j$

In this section, we treat the columns of the factor matrix $\mathbf{A}_j$ jointly in the approximate posterior using VB. We also define for the convenience of the analysis, $\mathbf{A}_j = [\mathbf{1}\,\mathbf{A}^{\mathbf{H}}_{\overline{\mathbf{1}},\mathbf{j}}]^{\mathbf{H}}$, where $\mathbf{A}_{\overline{1},j}$ represents all other rows except the first and $\mathbf{1}$ represents a column vector (of size M) with all ones.

$$\ln q_{\mathbf{A}_j}(\mathbf{A}_j) = - < \gamma >< \|\mathbf{Y} - [\![\mathbf{A}_1, ..., \mathbf{A}_N; \mathbf{x}]\!]\|_F^2 >= \text{tr}\{-\mathbf{Y}^{(j)}\mathbf{V}_j^H\mathbf{A}_j^H - \mathbf{A}_j\mathbf{V}_j\mathbf{Y}^{(j)\,H} + \mathbf{A}_j\mathbf{W}_j\mathbf{A}_j^H\} + c_{\mathbf{A}_j}, \tag{18}$$

Defining $\mathbf{B}_j$ as with the first column of $(\mathbf{Y}^{(j)}\mathbf{V}_j^H)$ removed. So $\text{tr}\{-\mathbf{Y}^{(j)}\mathbf{V}_j^H\mathbf{A}_j^H\} = \sum_{i=1}^{M}(\mathbf{Y}^{(j)} < (\bigodot_{k=1,k\neq j}^{N} \mathbf{A}_k)^* >< \mathbf{X}^H >)_{1,i} + \text{tr}\{\mathbf{B}_j\mathbf{A}^H_{\overline{1},j}\}$. Now consider the term $\mathbf{A}_j\mathbf{W}_j\mathbf{A}_j^H = [\mathbf{1}\,\mathbf{A}^H_{\overline{1},j}]^H\mathbf{W}_j[\mathbf{1}\,\mathbf{A}^{\mathbf{H}}_{\overline{\mathbf{1}},\mathbf{j}}]$ which can be simplified as,

$$[\mathbf{1}\,\mathbf{A}^H_{\overline{1},j}]^H\mathbf{W}_j[\mathbf{1}\,\mathbf{A}^H_{\overline{1},j}] = \begin{bmatrix} \mathbf{1}^H\mathbf{W}_j\mathbf{1} & \mathbf{1}^H\mathbf{W}_j\mathbf{A}^H_{\overline{1},j} \\ \mathbf{A}_{\overline{1},j}\mathbf{W}_j\mathbf{1} & \mathbf{A}_{\overline{1},j}\mathbf{W}_j\mathbf{A}^H_{\overline{1},j} \end{bmatrix}. \tag{19}$$

Finally (18) gets simplified as,

$$\ln q_{\mathbf{A}_j}(\mathbf{A}_j) =< \gamma > \text{tr}\{\mathbf{B}_j\mathbf{A}^H_{\overline{1},j}\} + < \gamma > \text{tr}\{\mathbf{A}_{\overline{1},j}\mathbf{B}_j^H\} - < \gamma > \text{tr}\{\mathbf{A}_{\overline{1},j} < \mathbf{X}(\bigcirc_{k=1,k\neq j}^{N} < \mathbf{A}_k^T\mathbf{A}_k^* >)\mathbf{X}^H > \mathbf{A}^H_{\overline{1},j}\} \tag{20}$$

It can be seen that (20) corresponds to the functional form of a circularly-symmetric complex matrix normal distribution [28]. This can be represented for a random matrix $\mathbf{X} \in \mathbf{C}^{n\times p}$ as $p(\mathbf{X}) \propto \exp(-\text{tr}\{\mathbf{\Psi}^{-1}(\mathbf{X} - \mathbf{M})^H\boldsymbol{\phi}^{-1}(\mathbf{X} - \mathbf{M})\})$, which is denoted as $\mathcal{CMN}(\mathbf{X} \mid \mathbf{M}, \boldsymbol{\phi}, \mathbf{\Psi})$. Thus the variational approximation for $\mathbf{A}_{\overline{1},j}$ gets represented as $\mathcal{CMN}(\mathbf{A}_{\overline{1},j} \mid \mathbf{M}_j, \mathbf{I}_{I_j}, \mathbf{\Psi}_j)$.

$$\begin{aligned} \mathbf{M}_j &= \widehat{\mathbf{A}}_{\overline{1},j} =< \gamma > \mathbf{B}_j\mathbf{\Psi}_j \\ \mathbf{\Psi}_j &= (< \gamma >< \mathbf{X} \bigcirc_{k=1,k\neq j}^{N} < \mathbf{A}_k^T\mathbf{A}_k^* > \mathbf{X}^H >)^{-1} \end{aligned} \tag{21}$$

Note that $vec(\widehat{\mathbf{A}}_{\overline{1},j}) \sim \mathcal{N}(vec(\mathbf{M}_j), \mathbf{\Psi}_j \otimes \mathbf{I}_M)$, so the terms of the form $< \|\mathbf{A}_{j,i}\|^2 >$ in (14) becomes, $< \|\mathbf{A}_{j,i}\|^2 >= 1 + \|\mathbf{M}_{j,i}\|^2 + (\mathbf{\Psi}_j)_{i,i}$. $(\mathbf{\Psi}_j)_{i,i}$ is the i^{th} diagonal element of $\mathbf{\Psi}_j$ and $\mathbf{M}_{j,i}$ represents the i^{th} column of $\mathbf{M}_j$. Also, we can represent $\mathbf{A}_j^H\mathbf{A}_j = \mathbf{1}\mathbf{1}^H + \mathbf{M}_j^H\mathbf{M}_j + I_j\mathbf{\Psi}_j$.

B. Estimation of the Parametric Dictionary

Considering the structured factor matrices which are Vandermonde, we estimate the spatial frequencies using least sqaures method. We define the operator $\ln(\widehat{\mathbf{A}}_{j,i})$ as the elementwise natural logarithm, $\widehat{\mathbf{A}}_{j,i}$ being the unstructured estimate. Now, the spatial frequencies can be estimated as follows,

$$\begin{aligned} &\widehat{\phi}_{j,i} = -j(\mathbf{d}_{I_j}^H\mathbf{d}_{I_j})^{-1}\mathbf{d}_{I_j}\ln(\mathbf{A}_{j,i}), \forall j = 1, ..., S, \\ &\text{where}, \mathbf{d}_{I_j} = [0, 1, ..., I_j - 1]^T. \end{aligned} \tag{22}$$

C. Computational Complexity

For our proposed SAVED, it is evident that we don't need any matrix inversions compared to [18], [21]. Update of all the variable $\mathbf{x}, \boldsymbol{\alpha}, \gamma$ involves simple addition and multiplication operations. We introduce the following variables, $\mathbf{e} = \mathbf{y}^T < \mathbf{A} >$ and $\mathbf{E} =< \mathbf{A}^T\mathbf{A} >$. $\mathbf{e}, \mathbf{E}$ and $||\mathbf{y}|\,|^2$ can be precomputed, so only computed once. We also introduce the following notations, $\mathbf{x}_{i-} = [x_1...x_{i-1}]^T, \mathbf{x}_{i+} = [x_{i+1}...x_M]^T$. Also we represent $\gamma^t =< \gamma >$, $\alpha_i^t =< \alpha_i >$, $x_i^t = \widehat{x}_i$ and $\mathbf{\Sigma}^t = \mathbf{\Sigma}$ in the following sections, where t represents the iteration stage.

Algorithm 1 SAVED SBL Algorithm

```
Given: $\mathbf{y}, \mathbf{A}, M, N$.
Initialization: $a, b, c, d$ are taken to be very low, on the order of $10^{-10}$. $\alpha_i^0 = a/b, \forall i, \gamma^0 = c/d$ and $\sigma_i^{2\,,0} = \frac{1}{||\mathbf{A}_i||^2\gamma^0+\alpha_i^0}, \mathbf{x}^0 = \mathbf{0}$.
At iteration $t + 1$,
    1) Update $\sigma_i^{2\,,t+1}$, $\widehat{x}_i^{t+1}, \forall i$ from (14) using $\mathbf{x}_{i-}^{t+1}$ and $\mathbf{x}_{i+}^t$.
    2) Update $\widehat{\mathbf{A}}_{j,i}, \forall i, j$ from (16) or $\mathbf{A}_j$ from (21).
    3) Compute $< x_i^{2\,,t+1} >$ from (17) and update $\alpha_i^t$.
    4) Update the noise variance, $\gamma^{t+1}$ from (17).
    5) Continue steps $1 - 4$ till convergence of the algorithm.
    6) Compute the spatial frequencies for the parametric components $\mathbf{A}_j, j = 1, ..., S$ from (22).
```

IV. Simulation Results

In this section, we present the simulation results to validate the performance of our SAVED SBL algorithm (Algorithm 1) compared to state of the art solutions. We compare our algorithm with the classical ALS [11]. For the simulations, we consider a $3-D$ tensor with dimensions $(4, 4, 4)$ and the number of non-zero elements of $\mathbf{x}$ or the rank of the tensor (no of non-zero elements of $\mathbf{x}$) is fixed to be 4. All the elements of the dictionary matrix $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ and non-zero elements of $\mathbf{x}$ are generated i.i.d complex Gaussian, $\mathcal{CN}(0, 1)$ and the singular values are modified to convert the matrices such that they have a particular condition number $(= 2)$. This is done to ensure that the system identifiability is not affected by the Krushkal ill-conditioning [11]. Normalized Mean Square Error (NMSE) is defined as $NMSE = \frac{1}{M}\,||\widehat{\mathbf{x}} - \mathbf{x}|\,|^2$, $\widehat{\mathbf{x}}$ represents the estimated value, $NMSE_{dB} = 10\log 10(NMSE)$. In Figure 1, we depict the normalized MSE (NMSE) performance of our proposed SAVED algorithm with the classical ALS algorithm which doesn't utilize any statistical information

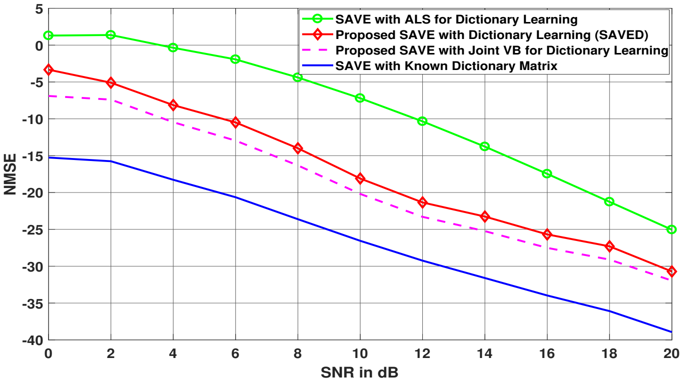

Fig. 1. NMSE vs SNR in dB.

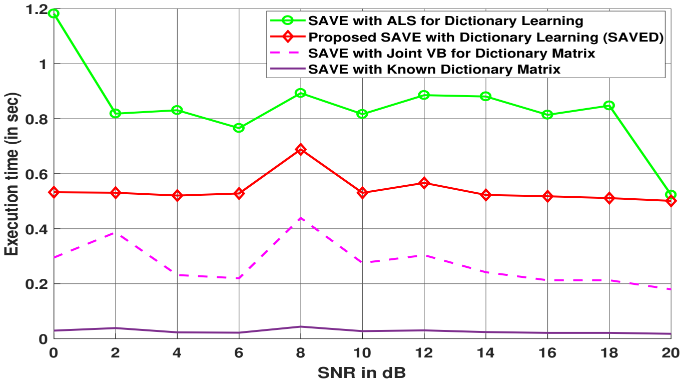

Fig. 2. Execution time in Matlab for the various algorithms.

about the dictionary or sparse coefficients. Our SAVED algorithm has much better reconstruction error performance compared to the ALS and our joint VB version performs better than the SAVED version, but comes with a higher computational complexity due to the matrix inversion. It is clear from Figure 2 that proposed SAVE approach has a faster convergence rate than the ALS.

V. Conclusion

We presented a fast SBL algorithm called SAVED, which uses the variational inference techniques to approximate the posteriors of the data and parameters. SAVE helps to circumvent the matrix inversion operation required in conventional SBL using EM algorithm. We showed that the proposed algorithm has a faster convergence rate and better performance in terms of NMSE than even the state of the art fast SBL solutions. Possible extensions to the current work might include: i) Convex combination of structured and unstructured KR factor matrices, for e.g., DoA response closeness to the vandermonde. iii) Asymptotic performance analysis and mismatched Cramer-Rao bounds for the SAVED algorithm.

References

[1] D. Nion and N. D. Sidiropoulos, "Tensor algebra and multidimensional harmonic retrieval in signal processing for MIMO radar ," *IEEE Trans. on Sig. Process.*, vol. 58, no. 11, Nov. 2010.

[2] C. Qian, X. Fu, N. D. Sidiropoulos, and Y. Yang, "Tensor-based parameter estimation of double directional massive MIMO channel with dual-polarized antennas," in *ICASSP*, 2018.

[3] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis," in *UCLA Working Papers in Phonetics, Available at http://publish.uwo.ca/harshman/wpppfac0.pdf.*, 1970.

[4] L. R. Tucker, "Implications of factor analysis of three-way matrices for measurement of change ," *Probl. Meas. Change*, 1963.

[5] M. Srensen, L. D. Lathauwer, P. Comon, S. Icart, and L. Deneire, "Canonical polyadic decomposition with a columnwise orthonormal factor matrix ," *SIAM J. Matrix Anal. Appl.*, vol. 33, no. 4, 2010.

[6] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit ," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, December 2007.

[7] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit ," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[8] D. Wipf and S. Nagarajan, "Iterative reweighted l_1 and l_2 methods for finding sparse solutions ," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329, April 2010.

[9] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[10] D. P. Wipf and B. D. Rao, "Sparse Bayesian Learning for Basis Selection ," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, August 2004.

[11] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications ," *SIAM Review*, vol. 51, no. 2, Aug. 2009.

[12] T. L. Hansen, M. A. Badiu, B. H. Fleury, and B. D. Rao, "A sparse Bayesian learning algorithm with dictionary parameter estimation ," in *8th IEEE Sens. Arr. and Multichanl. Sig. Process. Wkshp. (SAM),*, Spain, 2014.

[13] R. Giri and B. D. Rao, "Type I and type II bayesian methods for sparse signal recovery using scale mixtures," *IEEE Trans. on Sig. Process.*, vol. 64, no. 13, pp. 3418–3428, 2018.

[14] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing ," *PNAS*, vol. 106, no. 45, pp. 18 914–18 919, November 2009.

[15] R. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, August 2011, p. 21682172.

[16] S. Rangan, P. Schniter, and A. Fletcher, "On the convergence of approximate message passing with arbitrary matrices," in *Proc. IEEE Int. Symp. Inf. Theory*, 2014.

[17] ——, "Vector approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017.

[18] M. Al-Shoukairi, P. Schniter, and B. D. Rao, "GAMP-based low complexity sparse bayesian learning algorithm," *IEEE Transaction on Signal Processing*, vol. 66, no. 2, January 2018.

[19] M. J. Beal, "Variational algorithms for approximate Bayesian inference," in *Thesis, Univeristy of Cambridge, UK*, May 2003.

[20] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 131–146, November 2008.

[21] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, "Fast variational sparse bayesian learningwith automatic relevance determination for superimposed signals," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, December 2011.

[22] C. K. Thomas and D. Slock, "SAVE - Space alternating variational estimation for sparse Bayesian learning," in *IEEE Data Science Workshop*, June 2018.

[23] ——, "Gaussian variational Bayes Kalman filtering for dynamic sparse Bayesian learning," in *Intl. Conf. on Time Ser. and Forec.*, 2018.

[24] ——, "Space Alternating Variational Bayesian Learning for LMMSE Filtering," in *IEEE EUSIPCO*, September 2018.

[25] N. D. Sidiropoulos *et al.*, "Tensor decomposition for signal processing and machine learning ," *IEEE Trans. on Sig. Process.*, vol. 65, no. 13, July 2017.

[26] C. K. Thomas and D. Slock, "Variational Bayesian learning for channel estimation and transceiver determination," in *IEEE ITA Wkshp.*, Feb 2018.

[27] L. Cheng, Y.-C. Wu, and H. V. Poor, "Probabilistic tensor canonical polyadic decomposition with orthogonal factors ," *IEEE Trans. on Sig. Process.*, vol. 65, no. 3, Feb. 2017.

[28] A. K. Gupta and D. K. Nagar, "Matrix variate distributions," in *Boca Raton FL, USA: CRC Press*, 1999.