BP-VB-EP BASED STATIC AND DYNAMIC SPARSE BAYESIAN LEARNING WITH KRONECKER STRUCTURED DICTIONARIES

Christo Kurisummoottil Thomas, Dirk Slock

EURECOM, Sophia-Antipolis, France, Email:{kurisumm,slock}@eurecom.fr

ABSTRACT

In many applications such as massive multi-input multi-output (MIMO) radar, massive MIMO channel estimation, speech processing, image and video processing, the received signals are tensors. In such applications, utilizing techniques from tensor algebra can be beneficial since it retains the tensorial structure in the received signal compared to processing on the matricized version of the same signal. Furthermore, the underlying parameters or states to be estimated are sparse in many of the above-said applications compared to the large system dimensions. In this paper, we propose techniques which allow handling the extension of sparse Bayesian learning (SBL) to time-varying states. Adding the parameters of the autoregressive process which is used to the model the time-varyings of the state leads to a non-linear (at least bilinear) state-space model. Belief propagation (BP) is a promising method to compute the minimum mean squared error (MMSE) or maximum a posteriori (MAP) estimates, but at the the expense of a high computational burden. However, inspired by a previous work on a combined BP and variational Bayes (VB) technique, we noted that using a combination of BP, VB, and expectation propagation (EP) can help to alleviate the computational complexity.

Index Terms— Sparse Bayesian Learning, Variational Bayes, Tensor Decomposition, Kronecker Structured Dictionary Learning, Belief Propagation

1. INTRODUCTION

The signal model for the recovery of a time varying sparse signal under Kronecker structured (KS) [1,2] dictionary matrix can be formulated as

$$\begin{array}{l}\text{Observation: } \mathbf{y}_t = (\mathbf{A}_1^{(t)} \otimes \mathbf{A}_2^{(t)} .... \otimes \mathbf{A}_N^{(t)})\mathbf{x}_t + \mathbf{v}_t, \\ \text{State Update: } \mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t, \end{array} \tag{1}$$

where $\mathbf{y}_t = vec(\mathbf{Y}_t)$, $\otimes$ represents the Kronecker product between two matrices, $vec(\cdot)$ representing the vectorized version of the tensor or matrix $(\cdot)$, $\mathbf{Y}_t \in \mathcal{C}^{I_1 \times I_2 ... \times I_N}$ is the observations or data at time t, $\mathbf{A}_{j,i}^t \in \mathcal{C}^{I_j}$, the factor matrix $\mathbf{A}_j^{(t)} = [\mathbf{A}_{j,1}^{(t)}, ..., \mathbf{A}_{j,P_j}^{(t)}]$ which is unknown and the tensor product is represented by $[\![\mathbf{A}_1^{(t)}, ..., \mathbf{A}_N^{(t)}; \mathbf{x}_t]\!]$, $\mathbf{x}_t$ is the $M(= \prod\limits_{j=1}^{N} P_j)$-dimensional sparse signal and $\mathbf{w}_t$ is the additive noise. We consider $\mathbf{A}_{j,i}^{(t)} = [1\, \mathbf{a}_{j,i}^{(t)\,T}]^T$, justification for which can be seen in our previous work [3]. We also define the measurement of dictionary matrix as, $\mathbf{A}^{(t)} = \mathbf{A}_1^{(t)} \otimes \mathbf{A}_2^{(t)} .... \otimes \mathbf{A}_N^{(t)} = \bigotimes\limits_{j=1}^{N} \mathbf{A}_j^{(t)}$. Sparse signal $\mathbf{x}_t$ is modeled using an AR(1) process with a diagonal correlation coefficient matrix $\mathbf{F}$. $\mathbf{x}_t$ contains only K non-zero entries (or significant coefficients), with $K << M$ and thus the dictionary matrix to be learned allows a low rank representation. $\mathbf{v}_t$ is assumed to be a white Gaussian noise, $\mathbf{v}_t \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$. The Sparse Bayesian Learning (SBL) algorithm was first introduced by [4] in the machine learning context and then proposed for the first time for sparse signal recovery by [5]. The efficiency of the SBL approach revolves around the hierarchical prior modeling and the Bayesian LASSO [6] is a special case of SBL with Gaussian-Exponential hierarchical prior (equivalent to Laplacian marginal for $\mathbf{x}_t$). Dynamic autoregressive SBL (DAR-SBL) considered here is a case of joint Kalman filtering (KF) with a linear time-invariant diagonal state-space model, and parameter estimation, which can be considered an instance of nonlinear filtering.

Prior work on dictionary learning (DL) [7–11] focus on either maximum likelihood based schemes, LS or K-SVD algorithms. Due to space limitations, we refer the readers to a more detailed discussion on the state of the art in DL to our previous work [3].

1.1. Contributions of this paper

- This paper is an extension of our previous work based on KS dictionaries [3] to the case of time varying sparse state vector. Moreover, we observed that variational Bayes (VB) at the scalar level (for $\mathbf{x}_t$) and at the column level for the factor matrices $\mathbf{A}_j^{(t)}$ is quite suboptimal [12] since the posterior covariance computed does not converge to the true posterior. Hence, this requires more accurate approximations like belief propagation (BP) which we propose here.
- Building on the framework of [13], we combine BP and MF approximations in such a way as to optimize the message passing (MP) framework. The main focus being on getting reduced complexity (VB accounts for the low complexity) algorithms without sacrificing much on the performance (BP for performance improvement).

2. SBL DATA MODEL

[1] Starting from the system model (1), diagonal matrices $\mathbf{F}$ (assumed to be deterministic) and $\mathbf{\Gamma}$ are defined with its elements, $\mathbf{F}_{i,i} = f_i, f_i \in (-1, 1)$ and $\mathbf{\Gamma}^{-1} = \text{diag}\,(\boldsymbol{\alpha}), \boldsymbol{\alpha} = [\alpha_1, ...\alpha_M]$. Here α_i represents the inverse variance of $x_{i,t} \sim \mathcal{N}(0, \frac{1}{\alpha_i})$. Further, $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^{-1})$, where $\mathbf{\Lambda}^{-1}$=$\mathbf{\Gamma}(\mathbf{I} - \mathbf{F}\mathbf{F}^T)$=diag$\,(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_M})$. $\mathbf{w}_t$ are the complex Gaussian mutually uncorrelated state innovation sequences. Hence we sparsify the prediction error variance $\mathbf{w}_t$ also, with the same support as $\mathbf{x}_0$ and henceforth enforces the same support set for $\mathbf{x}_t, \forall t$. $\mathbf{v}_t$ is independent of the $\mathbf{w}_t$ process. The above signal model finds numerous applications including but not limited to 1) Bayesian adaptive filtering [14, 15],

EURECOM's research is partially supported by its industrial members: ORANGE, BMW, Symantec, SAP, Monaco Telecom, iABG, and by the projects DUPLEX (French ANR), MASS-START (French FUI) and EU ITN project SPOTLIGHT.

[1] In this paper, boldface lower-case and upper-case characters denote vectors and matrices respectively. The operators tr$(\cdot)$, $(\cdot)^T$, $(\cdot)^*$, $(\cdot)^H$, $\|\cdot\|$ represent trace, transpose, conjugate, conjugate transpose and Frobenius norm, respectively. A complex Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Theta}$ is denoted as $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Theta})$. diag$\,(\cdot)$ represents the diagonal matrix created by elements of a row or column vector. The operator $< x >$ or E$(\cdot)$ represents the expectation of x. $\mathbf{I}_N$ represents the identity matrix with dimension N. $\delta(\cdot)$ represents the Kronecker delta function.

2) Wireless channel estimation: multipath parameter estimation as in [16–18]. In Bayesian compressive sensing, a two-layer hierarchical prior is assumed for the $\mathbf{x}_t$ as in [4]. The hierarchical prior is chosen such that it encourages the sparsity property of $\mathbf{x}_t$ or of the innovation sequences $\mathbf{v}_t$. The state update gets represented as, $p(\mathbf{x}_t/\mathbf{x}_{t-1},\mathbf{F},\mathbf{\Gamma})=\prod\limits_{i=1}^{M}\mathcal{N}(f_i x_{i,t-1},\frac{1}{\alpha_i})$. For the convenience of analysis, we reparameterize α_i in terms of λ_i. Since $\lambda_i \propto \alpha_i$, $1/\lambda_i$'s are also sparse and hence we can assume a Gamma prior for $\mathbf{\Lambda}$, $p(\mathbf{\Lambda}) = \prod\limits_{i=1}^{M} Gamma(\lambda_i;a,b)$, where, $Gamma(\lambda_i;a,b) = \prod\limits_{i=1}^{M}\Gamma^{-1}(a)b^a\lambda_i^{a-1}e^{-b\lambda_i}$, such that the marginal pdf of $\mathbf{x}_t$ (student-t distribution) becomes more sparsity inducing than e.g., a Laplacian prior. The inverse of noise variance γ is distributed as, $Gamma(\gamma;c,d)$. The advantage is that the whole machinery of linear MMSE estimation can be exploited, such as e.g., the KF. But this is embedded in other layers making things eventually non-Gaussian. Now the likelihood distribution can be written as, $p(\mathbf{y}_t/\mathbf{x}_t,\gamma) \propto \gamma^N e^{-\gamma\left\|\mathbf{y}_t-\mathbf{A}^{(t)}\mathbf{x}_t\right\|^2}$. To make these priors non-informative (Jeffrey's prior), we choose them to be small values $a=c=b=d=10^{-5}$. We define the unknown parameter vector $\boldsymbol{\theta}_t=\{\mathbf{x}_t,\mathbf{\Lambda},\gamma,\mathbf{F},\mathbf{A}^{(t)}\}$.

We also remark that the techniques proposed here (with KS DL) are an instance of gridless compressed sensing, for e.g. [19]. To further elucidate the application of the system model in (1), we consider the wireless channel estimation in the case of millimeter wave or massive MIMO systems. The time varying channel impulse response $\mathbf{H}^{(t)}$ has per path a rank one contribution in four dimensions (Tx and Rx spatial multi-antenna dimensions, delay spread and Doppler spread). In the frequency domain

$$\begin{aligned}&\text{vec}(\mathbf{H}^{(t)}) = \textstyle\sum_{i=1}^{N_p} x_{i,t}\,\mathbf{h}_t(\psi_i)\otimes\mathbf{h}_r(\phi_i)\otimes\mathbf{v}_f(\tau_i(t))\otimes\mathbf{v}_t(f_i)\\ &= \mathbf{A}(\boldsymbol{\theta})^{(t)}\mathbf{x}_t\end{aligned}$$

where $\mathbf{v}_f(.)$, $\mathbf{v}_t(.)$ are appropriate Vandermonde vectors (possibly subsampled in the case of $\mathbf{v}_f(.)$). Hence we get a sum of rank one $4D$ tensors. We emphasize that the presented algorithm do not exploit parametric forms, because those parametric forms are uncertain. For eg., considering the massive MIMO channel estimation problem [20], the array response at the mobile station (MS) is not exploitable in practice. Even the array response at the base station (BS) will typically require calibration to be exploitable or may have mutual coupling effects which are unknown. Doppler shifts can be clearly represented as 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 frequency-flat. Then over those subcarriers, it's also Vandermonde. We do limit our discussions in this paper to the case of known $\mathbf{A}^{(t)}$. However, note that here we do consider the extension of the SAVE and BP based algorithm discussed herein to the case of unknown $\mathbf{A}^{(t)}$ under the dynamic case (static SBL is a special case).

3. COMBINED BP/MF APPROXIMATION USING VARIATIONAL FREE ENERGY OPTIMIZATION

The time index t is omitted here for simplicity. In this section, we develop an MP framework combining BP and VB which are derived from variational free energy (VFE) which is a fundamental quantity in statistical physics. We refer the readers to [21] for a detailed technical discussion on VFE and the theoretical formulation behind the derivation of MP expressions. We instead brief the results here. Assume that the posterior be factorized as, $p(\boldsymbol{\theta}) = \frac{1}{Z}\prod\limits_{a\in\mathcal{A}_{BP}} f_a(\boldsymbol{\theta}_a)\prod\limits_{b\in\mathcal{A}_{MF}} f_b(\boldsymbol{\theta}_b)$, where $\mathcal{A}_{BP},\mathcal{A}_{MF}$ represent the set of nodes belonging to the BP part and MF part, respectively. $\mathcal{N}_{BP}(i)$ represents the number of neighbouring nodes of i which belong to the BP part, similarly $\mathcal{N}_{MF}(i)$ is defined. $\mathcal{N}(i)$, $\mathcal{N}(a)$ represent the number of neighbouring nodes of any variable node i or factor node a, respectively. Let $m_{a\rightarrow i}$ represents the message passed from any factor node a to variable node i and $n_{i\rightarrow a}$ represents the message passed from any variable node i to factor node a. The fixed point equations corresponding to the constrained optimization VFE can be written as follows [13] ($<>_q$ represents the expectation w.r.t distribution q)

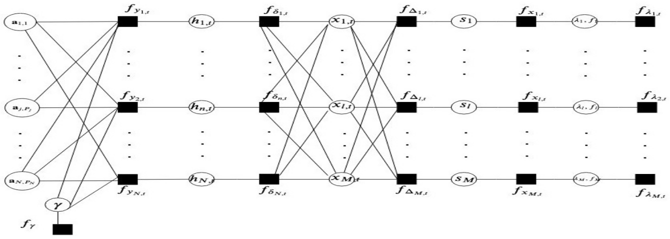

Fig. 1. Factor Graph (FG) for the dynamic SBL (at time t). Note that messages from the smoothing stage is not shown here.

$$\begin{aligned}
q_i(\theta_i) &= z_i \prod_{a\in\mathcal{N}_{BP}(i)} m^{BP}_{a\rightarrow i}(\theta_i)\prod_{a\in\mathcal{N}_{MF}(i)} m^{MF}_{a\rightarrow i}(\theta_i),\\
n_{i\rightarrow a}(\theta_i) &= \prod_{b\in\mathcal{N}_{BP}(i)\setminus a} m_{b\rightarrow i}(\theta_i)\prod_{b\in\mathcal{N}_{MF}(i)} m_{b\rightarrow i}(\theta_i),\\
m^{MF}_{a\rightarrow i}(\theta_i) &= \exp(<\ln f_a(\theta_a)>_{\prod\limits_{j\in\mathcal{N}(a)\setminus i} n_{j\rightarrow a}(\theta_j)}),\\
m^{BP}_{a\rightarrow i}(\theta_i) &= (\int\prod_{j\in\mathcal{N}(a)\setminus i} n_{j\rightarrow a}(\theta_j) f_a(\theta_a)\prod_{j\neq i} d\theta_j),
\end{aligned}\tag{2}$$

3.1. Dynamic BP-MF-EP based SBL

The figure 1 represents the FG (note that static case is a special case with the state update nodes being not present), where it is divided into two disjoint subsets $\mathcal{A}_{BP} = f_{\delta_{n,t}}\forall n,l,t$ and $\mathcal{A}_{MF}$ represents rest of the factor or variable nodes. To combine BP and MF, we introduce the new variables $h_{n,t} = \mathbf{A}^{(t)}_{n,:}\mathbf{x}_t, s_{l,t} = f_l x_{l,t-1}$ and the hard constraint factor nodes, $f_{\delta_{n,t}} = \delta(h_{n,t} - \mathbf{A}^{(t)}_{n,:}\mathbf{x}_t), \forall n\in[1:N], t$, and $f_{\Delta_{l,t}} = \delta(s_{l,t} - f_l x_{l,t-1}), \forall l\in[1:M], t$. We can compute $m_{f_{\delta_{n,t}\rightarrow x_{l,t}}}(x_{l,t}) = \int f_{\delta_{n,t}} n_{h_{n,t}\rightarrow f_{\delta_{n,t}}}(h_{n,t})\prod\limits_{l'\neq l} n_{x_{l',t}\rightarrow f_{\delta_{n,t}}}(x_{l',t})\prod\limits_{l'\neq l} dx_{l',t}$. For notational brevity, we denote subscript (l,n) or (n,l) to represent the messages passed from l to n or viceversa. All the messages (beliefs or continuous pdfs) passed between them can be shown to be Gaussian [22] and thus it suffices to represent them by the mean and variance of the beliefs. The joint distribution of all the observations and parameters can be written as, $p(\mathbf{y}_t,\boldsymbol{\theta}_t/\mathbf{y}_{1:t-1}) = p(\mathbf{y}_t/\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t/\mathbf{y}_{1:t-1})$, where $p(\boldsymbol{\theta}_t/\mathbf{y}_{1:t-1})$ denotes the predictive distribution. Similar as in KF, first we compute the posterior distribution of $\theta_{i,t}$ given the observations till $(t-1)$, which is called as the prediction stage.

3.1.1. Diagonal AR(1) (DAR(1)) Prediction Stage

Since there is no coupling between the scalars in the state update (1), it is enough to update the prediction stage using MF. However, the interation between $x_{l,t}$ and f_l requires Gaussian projection, using expectation propagation (EP). For more detailed derivation, we refer to our previous work [12] due to space limitations.

3.1.2. Measurement Update (Filtering) Stage

For the measurement update stage, the posterior for $\mathbf{x}_t$ is inferred using BP. Note that we represent the mean of the messages by

$\widehat{x}^{(t)}_{n,l}, \nu^{(t)}_{n,l}$. The mean and variance of the beliefs computed at $x_{l,t}$ are denoted by $\widehat{x}_{l,t|t}, \sigma^2_{l,t|t}$. In the measurement stage, the prior for $x_{l,t}$ gets replaced by the belief from the prediction stage. We refer to our previous work [21] for detailed derivations and expressions for the messages. We define $d_{l,t} = (\sum\limits_{n=1}^{N} \nu^{(t)\,-1}_{n,l})^{-1}, r_{l,t} = d_{l,t}(\sum\limits_{n=1}^{N} \frac{\widehat{x}^{(t)}_{n,l}}{\nu^{(t)}_{n,l}} + \frac{\widehat{x}_{l,t|t-1}}{\sigma^2_{l,t|t-1}})$. Given the messages, $m_{f_{\delta_{n,t}} \to x_{l,t}}(x_{l,t})$, the belief $q(x_{l,t})$ can be obtained as $(f_{\lambda_i}(\lambda_i) = p(\lambda_k/a, b))$, $q(x_{l,t}) \propto f_{\lambda_i}(\lambda_i) \prod\limits_{n=1}^{N} m_{f_{\delta_{n,t}} \to x_{l,t}} \propto \mathcal{N}(x_{l,t}; \widehat{x}_{l,t|t}, \sigma^2_{l,t|t})$, where

$$\sigma^{-2}_{l,t|t} = \lambda_{l,t} + d^{-1}_{l,t}, \ \widehat{x}_{l,t|t} = \tfrac{r_{l,t}}{1+d_{l,t}\sigma^{-2}_{l,t|t}}. \tag{3}$$

One remark here is that compared to our previous work using VB [23], combining BP and MF gives a more accurate approximation of the error variance as shown in (3), where $\sigma^2_{l,t|t}$ incorporates the effect of all $\sigma^2_{l',t|t}, l' \neq l$.

3.1.3. *Lag-1 Smoothing Stage*

We show in [21, Lemma 1] that KF is not enough to adapt the hyperparameters, instead we need at least a lag 1 smoothing (i.e. the computation of $\widehat{x}_{k,t-1|t}, \sigma^2_{k,t-1|t}$ through BP). For the smoothing stage, we use BP with Gaussian Markov Random Fields (GMRF) based factorization. GMRF refers to the representation of BP [24], when the underlying Gaussian distribution is expressed in terms of pairwise connections between scalar variables $x_{i,t}$. We skip the detailed derivation and instead refer to our paper [21]. Applying the MF rule from (2), the resulting Gaussian distribution has mean, $\sigma^{-2}_{f_i|t}$ and variance, $\widehat{f}_{i|t}$, the detailed derivations for which are in [21, Section 3.2.3]. The entire algorithm (a combination of BP, MF and EP, we call it as Combined BP-MF-EP DAR-SBL) is described in Algorithm 1. Also we remark that for the estimation of λ_l, γ, we follow the same approach as in our paper [12] and we refer to it for more details. One remark here is that another version called as Combined Vector BP-MF-EP DAR-SBL follows immediately from the derivations for Algorithm 1, where all the components of $\mathbf{x}_t$ are considered jointly in the FG. Even though the performance will be higher (as observed in the simulations) for the vector case, it comes at the cost of a higher complexity due to the matrix inversion involved. Note that in Algorithm 1, we introduce temporal averaging for certain quantities (represented by $<>_{|t}$) in hyperparameter estimates and β being the temporal weighting coefficient which is less than one, see [12] for more details. For the KS DL, the algorithm remains same as in our previous work [3], which is denoted as space alternating variational estimation with Kronecker structured DL (SAVED-KS DL).

Algorithm 1 Combined BP-MF-EP DAR-SBL with KS DL

```
Initialization: $\widehat{f}_{l|0}, \widehat{\lambda}_{l|0} = \frac{a}{b}, \widehat{\gamma}_0 = \frac{c}{d}, \widehat{x}_{l,0|0} = 0, \sigma^2_{l,0|0} = 0, \forall l$. Define $\mathbf{\Sigma}_{t-1|t-1} = \text{diag}\,(\sigma^2_{l,t|t-1})$.
for $t = 1 : T$ do
Prediction Stage: 1. From [12], $\widehat{x}_{l,t|t-1} = \widehat{f}_{l|t-1}\widehat{x}_{l,t-1|t-1}$, $\sigma^2_{l,t|t-1} = |\widehat{f}_{l|t-1}|^2\sigma^2_{l,t-1|t-1} + \sigma^2_{f_l|t-1}(|\widehat{x}_{l,t-1|t-1}|^2 + \sigma^2_{l,t-1|t-1}) + \widehat{\lambda}^{-1}_{l|t-1}$.
Filtering Stage:
    1. Compute $\widehat{x}^{(t)}_{n,l}, \nu^{(t)}_{n,l}$ from [21, eq. (5)] and update $\widehat{x}_{l,t|t}, \sigma^{-2}_{l,t|t}$ from (3).
    2. Compute $\nu^{(t)}_{l,n}, \widehat{x}^{(t)}_{l,n}$ from [21, eq. (7)].   3. Continue steps 1) to 2) until convergence.
Smoothing Stage:
Initialization: $\mathbf{\Sigma}^{(0)}_{t-1|t} = \mathbf{\Sigma}_{t-1|t-1}, \widehat{\mathbf{x}}^{(0)}_{t-1|t} = \widehat{\mathbf{x}}_{t-1|t-1}$. Define $\mathbf{B}^{(t)} =< \mathbf{F}^T\mathbf{A}^{(t)\,T}\widetilde{\mathbf{R}}^{-1}_t\mathbf{A}^{(t)}\mathbf{F} > +\mathbf{\Sigma}_{t-1|t-1}, \mathbf{h}_t =< \mathbf{F}^T\mathbf{A}^{(t)\,T} > \widetilde{\mathbf{R}}^{-1}_t\mathbf{y}_t$.
    1. $P_{i,j} = \frac{-B^{(t)\,2}_{i,j}}{B^{(t)}_{i,i} + \sum_{k\in\mathcal{N}(i)\backslash j} P_{k,i}}$, $\mu_{i,j} = (h_{i,t} + \sum_{k\in\mathcal{N}(i)\backslash j} P_{k,i}\mu_{k,i}), \forall i, j$.
    2. $\sigma^{-2}_{i,t-1|t} = B^{(t)}_{i,i} + \sum_{k\in\mathcal{N}(i)} P_{k,i}$, $\widehat{x}_{i,t-1|t} = \sigma^2_{i,t-1|t}(h_{i,t} + \sum_{k\in\mathcal{N}(i)} P_{k,i}\mu_{k,i})$
Estimation of hyperparameters (Define: $x'_{k,t} = x_{k,t} - f_k x_{k,t-1}, \zeta_t = \beta\zeta_{t-1} + (1-\beta) < \left\|\mathbf{y}_t - \mathbf{A}^{(t)}\mathbf{x}_t\right\|^2 >$), $b'_t = (< \left|x'_{k,t}\right|^2 >_{|t} + b)$.
    1. Compute $\widehat{f}_{l|t}, \sigma^2_{f_{l|t}}$ from [21, eq. (11)], $\widehat{\gamma}_t = \frac{c+N}{(\zeta_t+d)}$ and $\lambda_{l|t} = \frac{(a+1)}{b'_t}$.
SAVED-KS DL: $\widehat{\mathbf{a}}_{ji} = (\mathbf{b}_j)_{\overline{1}}$,  $\mathbf{b}_j = (\mathbf{Y}^{(j)} < \mathbf{X}^{(j)} >< (\bigotimes_{k=N,k\neq j}^{1} \mathbf{A}_k)^T >)_i$,
$\mathbf{\Upsilon}_{j,i} = \beta_{j,i}\mathbf{I}$, $\beta_{j,i} = \text{tr}\{(\bigotimes_{k=N,k\neq j}^{1} < \mathbf{A}^T_k\mathbf{A}^*_k >) < \mathbf{X}^{(j)\,T}\mathbf{X}^{(j)} >\}$.
```

3.2. Suboptimality of SAVED-KS DL and Joint VB

First, we define the unfolding operation on an N^{th} order tensor $\mathbf{Y}_t = [\![\mathbf{A}^{(t)}_1, ..., \mathbf{A}^{(t)}_N; \mathbf{x}]\!]$ as [25] ($\mathbf{Y}^{(n)}_t$ is of size $I_n \times \prod\limits_{i=1,i\neq n}^{N} I_i$)

$$\mathbf{Y}^{(n)}_t = \mathbf{A}^{(t)}_n\mathbf{X}^{(n)}_t(\mathbf{A}^{(t)}_N \otimes \mathbf{A}^{(t)}_{N-1}...\mathbf{A}^{(t)}_{n+1} \otimes \mathbf{A}^{t}_{n-1}... \otimes \mathbf{A}^{(t)}_1)^T. \tag{4}$$

From the expression for the error covariance in the estimation of the factor $\mathbf{a}_{ji}$ (tr$\{(\bigotimes\limits_{k=N,k\neq j}^{1} < \mathbf{A}^T_k\mathbf{A}^*_k >) < \mathbf{X}^{(j)\,T}\mathbf{X}^{(j)} >\}\mathbf{I}$), it is clear that it does not take into account the estimation error in the other columns of $\mathbf{A}_j$. The columns of $\mathbf{A}_j$ can be correlated, for e.g. if we consider two paths (say i, j) with same DoA but with different delays, the delay responses $\mathbf{v}_f(\tau_i(t))$ and $\mathbf{v}_f(\tau_i(t))$ may be correlated. However, since it is not clear how to model this dependency, we indeed keep it as a future work. This suboptimality in the error covariance estimate using SAVED-KS resulting from the correlation between the columns, can be avoided by using a joint VB [3]. The joint VB estimates (mean and covariance) can be obtained as

$$\begin{aligned} &\mathbf{M}^T_j = \widehat{\mathbf{A}}^T_{\overline{1},j} =< \gamma > \mathbf{\Psi}^{-1}_j\mathbf{B}^T_j, \\ &\mathbf{\Psi}_j = (< \gamma >< \mathbf{X}^{(j)}(\bigotimes_{k=N,k\neq j}^{1} < \mathbf{A}^T_k\mathbf{A}^*_k >)\mathbf{X}^{(j)\,T} >), \end{aligned} \tag{5}$$

where $\mathbf{V}_j =< \mathbf{X}^{(j)} >< (\bigotimes\limits_{k=N,k\neq j}^{1} \mathbf{A}_k)^T >$ and $\mathbf{B}_j$ is defined as with the first row of $(\mathbf{Y}^{(j)}\mathbf{V}^T_j)$ removed. However, the joint VB involves a matrix inversion and is not recommended for large system dimensions. Nevertheless, it is possible to estimate each columns of $\mathbf{A}_j$ by BP, since each column estimate can be expressed as the solution of a linear system of equation from (5), $\widehat{\mathbf{a}}^T_{j,i} = \mathbf{\Psi}^{-1}_j\mathbf{b}_{j,i}$. $\mathbf{b}_{j,i}$ represents the i^{th} column of $\mathbf{B}^T_j$. The message passing expressions under BP (using a GMRF based FG) for the factor matrices can be written as

$$\begin{aligned} &\zeta_{m,n} = -(\mathbf{\Psi}_j)^2_{m,n}/(\zeta_{m,m} + \sum_{k\in\mathcal{N}(m)\backslash j} \zeta_{k,m}), \\ &\kappa_{m,n} = \frac{(\zeta_{m,m}\kappa_{m,m}) + \sum_{k\in\mathcal{N}(m)\backslash j} \zeta_{k,m}\kappa_{k,m}}{(\mathbf{\Psi}_j)_{m,n}}, \end{aligned} \tag{6}$$

where we initialize $\zeta_{m,m} = (\mathbf{\Psi}_j)_{m,m}$, $\kappa_{m,m} = \frac{(\mathbf{b}_{j,i})_m}{(\mathbf{\Psi}_j)_{m,m},\zeta_{m,n}} = 0, \kappa_{m,n} = 0$. Finally the mean (κ_m) and variance (ζ_m) of the posterior belief can be computed as

$$\begin{aligned} &\kappa_m = \frac{\zeta_{m,m}\kappa_{m,m} + \sum_{k\in\mathcal{N}(m)} \zeta_{k,m}\kappa_{k,m}}{\zeta_{m,m} + \sum_{k\in\mathcal{N}(m)} \zeta_{k,m}}, \\ &\zeta_m = \zeta_{m,m} + \sum_{k\in\mathcal{N}(m)} \zeta_{k,m}. \end{aligned} \tag{7}$$

We remark that, the above BP based low complexity scheme for KS DL represents a major innovation compared to our previous work [3], apart from the extension to the dynamic SBL case.

4. OPTIMAL PARTITIONING OF THE MEASUREMENT STAGE AND KS DL

In [3], we derived the Fisher Information Matrix (FIM) for the KS DL where the sparse vector is static. Here, we reuse the FIM expressions to derive the optimal partitioning of the variables in the measurement stage. We refer to our paper [26, Lemma 1], where the main message was that if the parameter partitioning in VB is such that the different parameter blocks are decoupled at the level of FIM, then VB is not suboptimal in terms of (mismatched) Cramer-Rao Bound (mCRB). More detailed overview on mCRB can be found in [27]. If a finer partitioning granularity is used (such as up to scalar level as in MF), then VB becomes quite suboptimal, which can be alleviated by using BP instead.

Lemma 1. *For the measurement stage, an optimal partitioning is to apply BP for the sparse vector $\mathbf{x}_t$ and VB (SAVED-KS) for the columns of the factor matrices $\mathbf{A}^{(t)}_{j,i}$ assuming the vectors $\mathbf{A}^{(t)}_{j,i}$ are independent and have zero mean. However, if the columns of $\mathbf{A}^{(t)}_j$ are correlated, then a joint VB, with the posteriors of the factor matrices assumed independent, should be done for an optimal performance.*

Proof: Let $\mathbf{F}^{(i)}_j = [\mathbf{0}_{I_j \times I_j(i-1)} \; \mathbf{I}_{I_j} \; \mathbf{0}_{I_j \times I_j(P_j-i)}]$, $\boldsymbol{\Phi}_{j,t} = vec(\mathbf{A}^{(t)}_j)$. We observe that we can separate the contributions of $\mathbf{A}^{(t)}$ and $\mathbf{x}_t$ in (1) as, $\mathbf{y}_t = \underbrace{(\sum\limits_{r=1}^{M} x_{r,t}\mathbf{F}_r)}_{\mathbf{F}(\mathbf{x}_t)} \underbrace{(\bigotimes\limits_{j=1}^{N} \boldsymbol{\Phi}_{j,t})}_{\mathbf{f}(\boldsymbol{\Phi}_t)} + \mathbf{w}_t$. We define, $\mathbf{F}_r = \bigotimes\limits_{p_{ji},\forall j} \mathbf{F}^{(p_{ji})}_j, r = \sum\limits_{j=1}^{N} (p_{ji}-1)J_j + p_{Ni}, J_j = \prod\limits_{r=j+1}^{N} P_r$.

$\mathbf{J(\Phi_t, x_t)} = [\mathbf{J(\Phi_t)} \; \mathbf{J(x_t)}]$, $\mathbf{J(\Phi_t)} = [\mathbf{J(\Phi_{1,t})} \, ... \, \mathbf{J(\Phi_{N,t})}]$ where, $\mathbf{J(\Phi_{j,t})} = \mathbf{F(x_t)}(\mathbf{\Phi_{1,t}} \otimes ...\mathbf{I_{I_jP_j}} .... \otimes \mathbf{\Phi_{N,t}})$,

$\mathbf{J(x_t)} = [\, \mathbf{F_1}(\bigotimes\limits_{\mathbf{j=1}}^{\mathbf{N}} \mathbf{\Phi_{j,t}}), \, ..., \, \mathbf{F}_{\mathrm{M}}(\bigotimes\limits_{\mathbf{j=1}}^{\mathbf{N}} \mathbf{\Phi_{j,t}}))\,]$.

Further, the FIM for the case of SBL can be derived as [3]

$$FIM = \begin{bmatrix} \mathrm{E}(\gamma)\mathbf{J(\Phi_t)^T J(\Phi_t)} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{E}(\gamma)\mathbf{J(x_t)^T J(x_t)} + \mathrm{E}(\mathbf{\Gamma}^{-1}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & a\,\mathrm{E}(\mathbf{\Gamma}^{-2}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (N+c-1)\,\mathrm{E}(\gamma^{-2}) \end{bmatrix} \tag{8}$$

Here, $\gamma\mathbf{J(x_t)^T J(\Phi_t)} = \mathbf{0}$, since $\mathbf{x}_t$ is zero mean. If the all columns of $\mathbf{A}^{(t)}_j$ are independent and zero mean, then $\mathrm{E}(\mathbf{J(\Phi_t)^T J(\Phi_t)})$ becomes a diagonal matrix with no coupling between the free variables of any two different columns of the factor matrices. However, if any factor matrix is $\mathbf{A}^{(t)}_j$ is correlated, it is suboptimal to factorize the columns of $\mathbf{A}^{(t)}_j$ independently in the approximate posterior. Hence, in this case, a joint VB method (which has higher complexity) would be optimal to estimate the posterior distributions and this indeed justify the superior performance of joint VB approach described in Section 3.2.

5. SIMULATION RESULTS

For the observation model, the parameters chosen are $N = 256, M = 200$. For the simulations, we consider a $3-D$ tensor with dimensions $(4, 8, 8)$ and the number of non-zero elements of $\mathbf{x}_t$ or the rank of the tensor (no of non-zero elements of $\mathbf{x}_t$) is fixed to be $K = 16$. All signals are considered to be real in the simulation. All the elements of the factor matrix $\mathbf{A}^{(t)}_j$ (time varying) are generated i.i.d. from a Gaussian distribution with mean 0 and variance 1. The rows of $\mathbf{A}^{(t)}$ are scaled by $\sqrt{16}$ so that the signal part of any scalar observation has unit variance. Taking the SNR to be 20dB, the variance of each element of $\mathbf{v}_t$ (Gaussian with mean 0) is computed as 0.01.

Consider the state update, $\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t$. To generate $\mathbf{x}_0$, the first 16 elements are chosen as Gaussian (mean 0 and variance 1) and then the remaining elements of the vector $\mathbf{x}_0$ are put to zero. Then the elements of $\mathbf{x}_0$ are randomly permuted to distribute the 30 non-zero elements across the whole vector. The diagonal elements of $\mathbf{F}$ are chosen uniformly in $[0.9, 1)$. Then the covariance of $\mathbf{w}_t$ can be computed as $\mathbf{\Gamma}(\mathbf{I} - \mathbf{F}\mathbf{F}^T)$. Note that $\mathbf{\Gamma}$ contains the variances of the elements of $\mathbf{x}_t$ (including $t = 0$), where for the non-zero elements of $\mathbf{x}_0$ the variance is 1. Following observations can be made from the simulations. In Figure 2, which is for static SBL case with DL, there is substantial improvement in NMSE compared to our previous work [3]. Our proposed low complexity algorithm using BP has similar performance as that of joint VB which has higher complexity. In Figure 3, we evaluate the performance of the proposed BP-MF-EP DAR SBL and show that the parameter estimation benefits from BP. "MF DAR-SBL" refers to the sub-optimal version with no BP and only MF for filtering or smoothing of $\mathbf{x}_t$. Also we show the drastic improvement in performance with lag-1 smoothing for hyperparameter estimation compared to just using filtering.

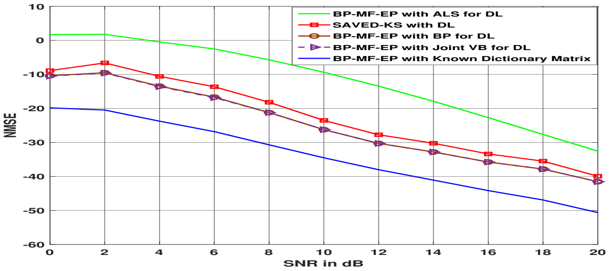

Fig. 2. Static SBL: NMSE as a function of N.

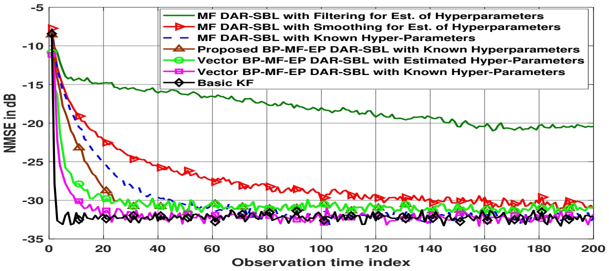

Fig. 3. DAR-SBL: NMSE as a function of time.

6. CONCLUSION

We have presented here a low complexity algorithm for KS DL using a combination of BP, VB and EP. The motivation behind the proposed algorithm is to circumvent the suboptimality associated with incorrect posterior covariance computation for the columns of the factor matrices in a previous paper of ours which was entirely based on MF approximations. However, we are still unclear whether the proposed algorithm is robust enough to perform comparably with the variations in the model of KS $\mathbf{A}^{(t)}$, which is left as a future work.

7. REFERENCES

[1] M. F. Duarte and R. G. Baraniuk, “Kronecker compressive sensing ,” *Proceedings of the IEEE*, vol. 21, no. 2, Feb. 2012.

[2] Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, “Identifiability of Kronecker-structured dictionaries for tensor data ,” *IEEE Journ. Sel. Top. in Sig. Process.*, vol. 12, no. 5, Oct. 2018.

[3] C. K. Thomas and D. Slock, “Space alternating variational estimation and kronecker structured dictionary learning,” in *IEEE Intl. Conf. on Acous. Spee. and Sig. Process. (ICASSP)*, May 2019.

[4] M. E. Tipping, “Sparse Bayesian learning and the relevance vector machine,” *J. Mach. Learn. Res.*, vol. 1, 2001.

[5] D. P. Wipf and B. D. Rao, “Sparse Bayesian Learning for Basis Selection ,” *IEEE Trans. on Sig. Process.*, vol. 52, no. 8, August 2004.

[6] T. Park and G. Casella, “The Bayesian Lasso,” *J. Amer. Statist. Assoc.*, vol. 103, no. 482, Nov. 2008.

[7] M. S. Lewicki and T. J. Sejnowski, “Learning overcomplete respresentations ,” *Neural Computation*, vol. 12, no. 2, 2000.

[8] K. Skretting and K. Engang, “Recursive least squares dictionary learning algorithm ,” *IEEE Trans. on Sig. Process.*, vol. 58, Apr. 2010.

[9] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation ,” *IEEE Trans. on Sig. Process.*, vol. 54, no. 11, Nov. 2006.

[10] F. Roemer, G. D. Galdo, and M. Haardt, “Tensor-based algorithms for learning multidimensional separable dictionaries,” in *IEEE Intl. Conf. on Acous., Speech and Sig. process. (ICASSP)*, 2014.

[11] X. Ding, W. Chen, and I. J. Wassell, “Joint sensing matrix and sparsifying dictionary optimization for tensor compressive sensing. ,” *IEEE Trans. on Sig. Process.*, vol. 65, no. 4, Jul. 2017.

[12] C. K. Thomas and D. Slock, “Gaussian variational Bayes Kalman filtering for dynamic sparse Bayesian learning,” in *IEEE 5th Intl. Conf. on Time Ser. and Forecast. (ITISE)*, September 2018.

[13] E. Riegler, G. E. Kirkelund, C. N. Manchn, M. A. Badiu, and B. H. Fleury, “Merging belief propagation and the mean field approximation: a free energy approach ,” *IEEE Trans. on Info. Theo.*, vol. 59, no. 1, Jan. 2013.

[14] T. Sadiki and D. Slock, “Bayesian adaptive filtering: principles and practical approaches,” in *EUSIPCO*, 2004.

[15] J. B. S. Ciochina, C. Paleologu, “A family of optimized LMS-based algorithms for system identification,” in *Proc. EUSIPCO*, 2016.

[16] B. H. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, and K. I. Pedersen, “Channel Parameter Estimation in Mobile Radio Environments Using the SAGE Algorithm ,” *IEEE Journal on selected areas in communications*, vol. 17, no. 3, pp. 434–450, March 1999.

[17] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, “Compressed channel sensing: A new approach to estimating sparse multipath channels ,” *Proceedings of the IEEE*, vol. 98, no. 6, Jun. 2010.

[18] Q. Cheng, X. Fu, and N. D. Sidiropoulos., “Algebraic Channel Estimation Algorithms for FDD Massive MIMO systems ,” *arXiv preprint arXiv:1903.08938*, 2019.

[19] K. Ardah, A. L. de Almeida, and M. Haardt, “A Gridless CS Approach for Channel Estimation in Hybrid Massive MIMO Systems,” in *IEEE Intl. Conf. on Acous., Speech and Sig. Process. (ICASSP)*, Brighton, UK, 2019.

[20] C. K. Thomas and D. Slock, “Variational Bayesian learning for channel estimation and transceiver determination,” in *IEEE Info. Theo. and Appl. Wkshp. (ITA)*, Feb 2018.

[21] ——, “Low Complexity Static and Dynamic Sparse Bayesian Learning Combining BP, VB and EP Message Passing,” in *52nd IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2019.

[22] X. Tan and J. Li, “Computationally efficient sparse Bayesian learning via belief propagation ,” *IEEE Trans. on Sig. Proc.*, vol. 58, no. 4, Apr. 2013.

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

[24] J. S. Yedidia, W. T. Freeman, and Y. Weiss, “Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms,” *IEEE Trans. On Info. Theo.*, vol. 51, July 2005.

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

[26] C. K. Thomas and D. Slock, “Sparse Bayesian Learning for a Bilinear Calibration Model and Mismatched CRB,” in *IEEE EUSIPCO*, Sept. 2019.

[27] S. Fortunati, F. Gini, M. Greco, and C. Richmond, “Performance Bounds for Parameter Estimation under Misspecified Models [Fundamental findings and applications],” *IEEE Sig. Proc. Mag.*, Nov. 2017.