LOW COMPLEXITY STATIC AND DYNAMIC SPARSE BAYESIAN LEARNING COMBINING BP, VB AND EP MESSAGE PASSING

Christo Kurisummoottil Thomas, Dirk Slock

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

ABSTRACT

Sparse Bayesian Learning (SBL) provides sophisticated (state) model order selection with unknown support distribution. This allows to handle problems with big state dimensions and relatively limited data by exploiting variations in parameter importance. The techniques proposed in this paper allow to handle the extension of SBL to time-varying states, modeled as diagonal first-order auto-regressive (DAR(1)) processes with unknown parameters to be estimated also. Adding the parameters to the state leads to an augmented state and a non-linear (at least bilinear) state-space model. The proposed approach, which applies also to more general non-linear models, uses a combination of belief propagation (BP), Variational Bayes (VB) or mean field (MF) techniques, and Expectation Propagation (EP) to approximate the posterior marginal distributions of the scalar factors. We propose Fisher Information Matrix analysis to determine the variable split between the use of BP and VB allowing to stay optimal in terms of Laplace approximation.

1. INTRODUCTION

The signal model for the recovery of a time varying sparse signal can be formulated as, $\boldsymbol{y}_t = \mathbf{A}^{(t)}\boldsymbol{x}_t + \boldsymbol{v}_t$, where $\boldsymbol{y}_t$ is the observations or data at time t, $\mathbf{A}^{(t)}$ is called the measurement or the sensing matrix which is known and is of dimension $N \times M$ with $N < M$. $\boldsymbol{x}_t$ contains only K non-zero entries, with $K << M$ and is modeled by a diagonal AR(1) (auto-regressive) process. In the static case, in Bayesian learning, the sparse Bayesian learning (SBL) algorithm was first proposed by [1, 2], which got extended to dynamic SBL in [3]. However, in order to render low complexity or low latency solutions, online processing algorithms (which process a small set of measurement vectors at any time) will be necessary. 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 filering.

In [4], they introduce a belief propagation (BP) based SBL algorithm which is more computationally efficient than the original algorithm. The authors use BP to infer the posterior pdf of $\boldsymbol{x}$ and the hyperparameters are estimated using the EM algorithm. The authors in [5] propose a message passing (MP) approach for inferring the posteriors combining BP and mean field (MF) approximations. MF is a special case of Variational Bayes (VB) in which the partitioning of variables is pushed to the scalar granularity. The advantages of the MF approach are that it always admits a convergent implementation while BP yields a good approximation of the posterior marginals if the factor graph has no cycles. The authors show that the MP fixed-point equations for a combination of BP and the MF approximation correspond to stationary points of one single constrained region-based free energy approximation and provide a clear rule stating how to couple the messages propagating in the BP and MF part. Hence, it is advantageous to apply BP and the MF approximation on the same factor graph in such a combination that their respective virtues can be exploited while overcoming their drawbacks (complexity for BP, potential suboptimality for MF). However, [6] does not treat at all the topic of how to split nodes between BP and MF. We also note that the approximate message passing algorithms [6, 7] suffer from the limitation that the large system limits assume i.i.d. Gaussian or right rotationally invariant $\mathbf{A}^{(t)}$, and the algorithms may exhibit convergence problems.

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

1.1. Contributions of this paper

- We propose new low complexity SBL algorithms for the static and dynamic cases, with joint hyperparameter estimation.
- Building on the framework of [5], we combine BP and MF approximations in such a way as to optimize the message passing framework, unlike most of the existing applications of the framework, which apply BP and MF to the variable subsets with discrete and continuous distributions resp.
- Using Fisher Information Matrix (FIM) analysis, we propose an optimal partitioning of the unknown parameters in the factor graph such that we can combine BP and (EP) VB in an efficient way, with low complexity and no suboptimality in terms of Laplace approximation (FIM).
- Various new algorithms in this paper are an application of these parameter partitioning and BP/VB split guidelines. For both a static (classic) compressed sensing model or a dynamic case with autoregressive evolution of the unknown x (corresponding to a classical linear state-space model apart from sparsity considerations). We furthermore show in Lemma 1, in another application of the FIM analysis, that identifiability of the hyperparameters (state space model parameters) requires smoothing (filtering is not sufficient). Although (regardless of sparsity) KF with joint parameter estimation has been the subject of many approaches over decades, this smoothing requirement has never been pointed out or certainly not been analyzed before.

2. DYNAMIC SBL SYSTEM MODEL

[1] Sparse signal $\boldsymbol{x}_t$ is modeled using an AR(1) process with a diagonal correlation coefficient matrix $\boldsymbol{F}$, which can be written as follows,

$$\begin{aligned} &\text{State Update: } \boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \\ &\text{Observation: } \boldsymbol{y}_t = \mathbf{A}^{(t)}\boldsymbol{x}_t + \boldsymbol{v}_t, \end{aligned} \tag{1}$$

where $\boldsymbol{x}_t = [x_{1,t},\ ...,\ x_{M,t}]^T$. Diagonal matrices $\boldsymbol{F}$ and $\boldsymbol{\Gamma}$ are defined with its elements, $\boldsymbol{F}_{i,i} = f_i, f_i \in (-1, 1)$ and

[1] Notations: The operator $(\cdot)^H$ represents the conjugate transpose or conjugate for a matrix or a scalar respectively. The operators tr$(\cdot)$ represents trace if a matrix. In the following, the pdf of a complex Gaussian random variable x with mean μ and variance σ^2 is given by $\mathcal{CN}(x; \mu, \nu)$. $KL(q||p)$ represents the Kullback-Leibler distance between the two distributions q, p. $\mathbf{A}_{n,:}$ or $\mathbf{A}_n$ represents the n^{th} row or n^{th} column of $\mathbf{A}$ respectively. blkdiag$(\cdot)$ represents blockdiagonal part of a matrix. diag$(\mathbf{X})$ or diag$(\boldsymbol{x})$ represents a vector obtained by the diagonal elements of the matrix $\mathbf{X}$ or the diagonal matrix obtained with the elements of $\boldsymbol{x}$ in the diagonal respectively.

$\mathbf{\Gamma} = \text{diag}(\boldsymbol{\alpha}), \boldsymbol{\alpha} = [\alpha_1, ...\alpha_M]$. Here α_i represents the inverse variance of $x_{i,t} \sim \mathcal{CN}(0, \frac{1}{\alpha_i})$. Further, $\boldsymbol{w}_t \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Lambda}^{-1})$, where $\mathbf{\Lambda}^{-1} = \mathbf{\Gamma}(\mathbf{I} - \boldsymbol{F}\boldsymbol{F}^H) = \text{diag}(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_M})$ and $\boldsymbol{v}_t \sim \mathcal{CN}(\mathbf{0}, \frac{1}{\gamma}\mathbf{I})$. $\boldsymbol{w}_t$ are the complex Gaussian mutually uncorrelated state innovation sequences. Hence we sparsify the prediction error variance $\boldsymbol{w}_t$ also, with the same support as $\boldsymbol{x}_0$ and henceforth enforces the same support set for $\boldsymbol{x}_t, \forall t$. $\boldsymbol{v}_t$ is independent of the $\boldsymbol{w}_t$ process. Although the above signal model seems simple, there are numerous applications such as 1) Bayesian adaptive filtering [8], 2) Wireless channel estimation: multipath parameter estimation as in [9]. In this case, $\boldsymbol{x}_t$ = FIR filter response, and $\mathbf{\Gamma}$ represents e.g. the power delay profile.

In Bayesian compressive sensing, a two-layer hierarchical prior is assumed for the $\boldsymbol{x}$ as in [1]. The hierarchical prior is chosen such that it encourages the sparsity property of $\boldsymbol{x}_t$ or of the innovation sequences $\boldsymbol{v}_t$. The state update gets represented as, $p(\boldsymbol{x}_t/\boldsymbol{x}_{t-1}, \boldsymbol{F}, \mathbf{\Gamma}) = \prod_{i=1}^{M} \mathcal{CN}(f_i x_{i,t-1}, \frac{1}{\alpha_i})$. For the convenience of analysis, we reparameterize α_i in terms of λ_i and assume a Gamma prior for $\mathbf{\Lambda}$, $p(\mathbf{\Lambda}) = \prod_{i=1}^{M} p(\lambda_i/a, b) = \prod_{i=1}^{M} \Gamma^{-1}(a) b^a \lambda_i^{a-1} e^{-b\lambda_i}$. The inverse of noise variance γ is also assumed to have a Gamma prior, $p(\gamma/c, d) = \Gamma^{-1}(c) d^c \gamma_i^{c-1} e^{-d\gamma}$, such that the marginal pdf of $\boldsymbol{x}_t$ (student-t distribution) becomes more sparsity inducing than e.g., a Laplacian prior. The advantage is that the whole machinery of linear MMSE estimation can be exploited, such as e.g., the Kalman filter. But this is embedded in other layers making things eventually non-Gaussian. Now the likelihood distribution can be written as, $p(\boldsymbol{y}_t/\boldsymbol{x}_t, \gamma) = (2\pi)^{-N} \gamma^N e^{-\gamma ||\boldsymbol{y}_t - \mathbf{A}^{(t)}\boldsymbol{x}_t||^2}$. To make these priors non-informative (Jeffrey's prior), we choose them to be small values $a = c = b = d = 10^{-5}$. For the AR(1) coefficients f_k, we don't assume any prior distribution. We define the unknown parameter vector $\boldsymbol{\theta} = \{\boldsymbol{x}, \mathbf{\Lambda}, \gamma, \boldsymbol{F}\}$ and θ_i represents each scalar in $\boldsymbol{\theta}$.

3. COMBINED BP/MF APPROXIMATION

The fixed points of the standard BP algorithm are shown to be the stationary points of the Bethe free energy (BFE) [5]. However, for the MF approximation in variational Bayes, the approximate posteriors are shown to be converging to a local minimum of the MF free energy which is an approximation of the BFE. Moreover, we observe in [10, 11] that for estimation of the signals from interference corrupted observations, MF is a poor choice since it doesn't give the accurate posterior variance (posterior variance of x_i is observed to be independent of the error variances of other $x_l, l \neq i$). Assume that the posterior be represented as, $p(\boldsymbol{\theta}) = \frac{1}{Z} \prod_{a \in \mathcal{A}_{BP}} f_a(\boldsymbol{\theta}_a) \prod_{b \in \mathcal{A}_{MF}} f_b(\boldsymbol{\theta}_b)$, where $\mathcal{A}_{BP}, \mathcal{A}_{MF}$ represent the set of nodes belonging to the BP part and MF part respectively with $\mathcal{A}_{BP} \cap \mathcal{A}_{MF} = \emptyset$. Z represents the normalization variable. Throughout the paper, the vector $\boldsymbol{\theta}_i$ represents a subset of $\boldsymbol{\theta}$ and θ_i represents a scalar parameter in $\boldsymbol{\theta}$. $\mathcal{N}(i), \mathcal{N}(a)$ represent the number of neighbouring nodes of any variable node i or factor node a. $\mathcal{N}_{BP}(i)$ represents the number of neighbouring nodes of i which belong to the BP part, similarly $\mathcal{N}_{MF}(i)$ is defined. Also, we define $\mathcal{I}_{MF} = \bigcup_{a \in \mathcal{A}_{MF}} \mathcal{N}(a), \mathcal{I}_{BP} = \bigcup_{a \in \mathcal{A}_{BP}} \mathcal{N}(a)$. The resulting free energy obtained by the combination of BP and MF are written as below (Note that we use an abuse of notation and let $q_i(\theta_i)$ represents the belief about θ_i (the approximate posterior)),

$$\begin{aligned} F_{BP,MF} = & \sum_{a \in \mathcal{A}_{BP}} \sum_{\boldsymbol{\theta}_a} q_a(\boldsymbol{\theta}_a) \ln \tfrac{q_a(\boldsymbol{\theta}_a)}{f_a(\boldsymbol{\theta}_a)} - \sum_{a \in \mathcal{A}_{MF}} \sum_{\boldsymbol{x}_a} \prod_{i \in \mathcal{N}(a)} \\ & q_i(\theta_i) \ln f_a(\boldsymbol{\theta}_a) - \sum_{i \in \mathcal{I}} (|\mathcal{N}_{BP}(i)| - 1) \sum_{\theta_i} q_i(\theta_i) \ln q_i(\theta_i). \end{aligned} \tag{2}$$

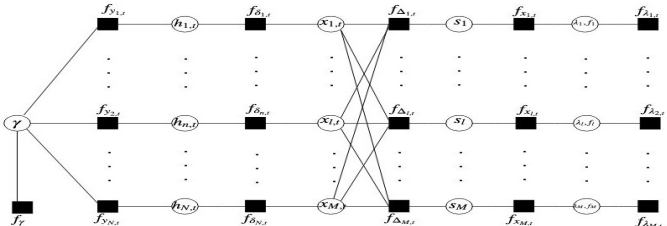

Fig. 1: Factor Graph for the dynamic SBL (at time t). Note that messages from the smoothing stage is not shown here.

The beliefs have to satisfy the following normalization and marginalization constraints,

$$\begin{aligned} & \sum_{\theta_i} q_i(\theta_i) = 1, \ \forall i \in \mathcal{I}_{MF} \setminus \mathcal{I}_{BP}, \ \sum_{\boldsymbol{\theta}_a} q_a(\boldsymbol{\theta}_a) = 1, \forall a \in \mathcal{A}_{BP}, \\ & q_i(\theta_i) = \sum_{\boldsymbol{\theta}_a \setminus \theta_i} q_a(\boldsymbol{\theta}_a), \ \forall a \in \mathcal{A}_{BP}, i \in \mathcal{N}(a). \end{aligned} \tag{3}$$

Let $m_{a \to i}$ represents the message passed from any factor node a to variable node i and $n_{i \to a}$ represents the message passed from any variable node i to factor node a. The fixed point equations corresponding to the constrained optimization of (2) can be written as follows [5],

$$\begin{aligned} q_i(\theta_i) &= z_i \prod_{a \in \mathcal{N}_{BP}(i)} m^{BP}_{a \to i}(\theta_i) \prod_{a \in \mathcal{N}_{MF}(i)} m^{MF}_{a \to i}(\theta_i), \\ n_{i \to a}(\theta_i) &= \prod_{a \in \mathcal{N}_{BP}(i) \setminus a} m_{a \to i}(\theta_i) \prod_{a \in \mathcal{N}_{MF}(i)} m_{a \to i}(\theta_i), \\ m^{MF}_{a \to i}(\theta_i) &== \exp(< \ln f_a(\theta_a) > \prod_{j \in \mathcal{N}(a) \setminus i} n_{j \to a}(\theta_j)), \\ m^{BP}_{a \to i}(\theta_i) &= (\int \prod_{j \in \mathcal{N}(a) \setminus i} n_{j \to a}(\theta_j) f_a(\theta_a) \prod_{j \neq i} d\theta_j), \end{aligned} \tag{4}$$

where $<>_q$ represents the expectation w.r.t distribution q.

3.1. BP-MF based Static SBL

The figure 1 represents the factor graph (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 $h_{n,t} = \mathbf{A}^{(t)}_{n,:}\boldsymbol{x}_t, s_{l,t} = f_l x_{l,t-1}$ and the hard constraint factor nodes, $f_{\delta_{n,t}} = \delta(h_{n,t} - \mathbf{A}^{(t)}_{n,:}\boldsymbol{x}_t), \forall n \in [1:N], t, f_{\Delta_{l,t}} = \delta(s_{l,t} - f_l x_{l,t-1}), \forall l \in [1:M], t$. For the static case, the system model will be $\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{v}$, so $f_l = 0, \lambda_l = \alpha_l, \forall l$. We omit subscript t for simplicity. The message $m_{f_{\delta_n} \to x_l}$ from the hard factor f_{δ_n} to variable node x_l is computed by the BP rule with the incoming messages to the node, $n_{h_n \to f_{\delta_n}}(h_n) = m_{f_{y_n} \to h_n}(h_n)$ and $n_{x_{l'} \to f_{\delta_n}}(x_{l'}), \forall l' \neq l$, later defined in (6). So $m_{f_{\delta_n} \to x_l}(x_l) = \int f_{\delta_n} n_{h_n \to f_{\delta_n}}(h_n) \prod_{l' \neq l} n_{x_{l'} \to f_{\delta_n}}(x_{l'}) \prod_{l' \neq l} dx_{l'}$. 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 [4] and thus it suffices to represent them by the mean and variance of the beliefs. With the hard constraints, the equivalent observation model can be written as,

$$\begin{aligned} & y_n - \textstyle\sum_{l' \neq l}^{M} A_{n,l'} \widehat{x}_{l',n} = A_{n,l} x_l + \sum_{l' \neq l}^{M} A_{n,l'} \widetilde{x}_{l',n} + v_n, \text{ where,} \\ & \widetilde{x}_{l',n} \sim \mathcal{CN}(0, \nu_{l',n}), \text{ and } m_{f_{\delta_n} \to x_l} \propto \mathcal{CN}(x_l; \widehat{x}_{n,l}, \nu_{n,l}), \\ & \widehat{x}_{n,l} = A^{-1}_{n,l}(y_n - p_n + A_{n,l}\widehat{x}_{l,n}), \ p_n = \textstyle\sum_{l'=1}^{M} A_{n,l'} \widehat{x}_{l',n}, \\ & \nu_{n,l} = |A_{n,l}|^{-2}(\widehat{\gamma}^{-1} + \nu_n - |A_{n,l}|^2 \nu_{l,n}), \ \nu_n = \textstyle\sum_{l'=1}^{M} |A_{n,l'}|^2 \nu_{l',n}. \end{aligned} \tag{5}$$

We define $d_l = (\sum_{n=1}^{N} \nu^{-1}_{n,l})^{-1}, r_l = d_l(\sum_{n=1}^{N} \frac{\widehat{x}_{n,l}}{\nu_{n,l}})$. Given the messages, $m_{f_{\delta_n} \to x_l}(x_l)$, the belief $q(x_l)$ can be obtained as $(f_{\lambda_i}(\lambda_i) = p(\lambda_k/a, b))$, $q(x_l) \propto f_{\lambda_i}(\lambda_i) \prod_{n=1}^{N} m_{f_{\delta_n} \to x_l} \propto \mathcal{CN}(x_l; \widehat{x}_l, \sigma_l^2)$,

$$\text{where } \sigma_l^{-2} = \lambda_l + d_l^{-1},\ \widehat{x}_l = \tfrac{r_l}{1+d_l\sigma_l^{-2}}. \tag{6}$$

One remark here is that compared to our previous work using VB [10], combining BP and MF gives a more accurate approximation of the error variance as shown in (6), where σ_l^2 incorporates the effect of all $\sigma_{l'}^2, l' \neq l$. Since the factor node $f_{\delta_n} \in \mathcal{A}_{BP}$, the message $n_{x_l \to f_{\delta_n}}(x_l)$ from variable node x_l to f_{δ_n} is updated by the BP rule as follows,

$$\begin{aligned} &n_{x_{l,t}\to f_{\delta_n}}(x_l) = \tfrac{q(x_l)}{m_{f_{\delta_n}\to x_l}(x_l)} \propto \mathcal{CN}(x_l; \widehat{x}_{l,n}, \nu_{l,n}), \\ &\text{where, } \nu_{l,n}^{-1} = (\sigma_l^{-2} - \nu_{n,l}^{-1}),\ \widehat{x}_{l,n} = \nu_{l,n}(\tfrac{\widehat{x}_l}{\sigma_l^2} - \tfrac{\widehat{x}_{n,l}}{\nu_{n,l}}). \end{aligned} \tag{7}$$

3.2. Dynamic BP-MF-EP based SBL

The joint distribution of all the observations and parameters can be written as, $p(\boldsymbol{y}_t, \boldsymbol{\theta}/\boldsymbol{y}_{1:t-1}) = p(\boldsymbol{y}_t/\boldsymbol{\theta})p(\boldsymbol{\theta}/\boldsymbol{y}_{1:t-1})$, where $p(\boldsymbol{\theta}/\boldsymbol{y}_{1:t-1})$ denotes the predictive distribution. Similar as in KF, first we compute the posterior distribution of θ_i given the observations till $(t-1)$, which is called as the prediction stage. Since the correlation coefficient matrix $\boldsymbol{F}$ is diagonal, all the $x_{i,t}$ are decoupled in the state update model and we exploit this fact to predict the states and the hyperparameters in the state update model using MF.

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

Assuming that the belief $q(\gamma)$ at time t, of noise precision γ is known, the message $m_{f_{y_{n,t}}\to h_{n,t}}(h_{n,t})$ from the factor node $f_{y_{n,t}} \in \mathcal{A}_{MF}$ is calculated using the MF rule $m_{f_{y_{n,t}}\to h_{n,t}}(h_{n,t}) =< \exp\left(\ln f_{y_{n,t}}(h_{n,t},\gamma)\right) >_{q(\gamma)}$, which becomes, $m_{f_{y_{n,t}}\to h_{n,t}}(h_{n,t}) \propto \mathcal{CN}(h_{n,t}; y_{n,t}, \widehat{\gamma}_t^{-1})$. Here $\widehat{\gamma}_t =< \gamma >_{q(\gamma)}$. For more detailed derivation, we refer to our paper [12]. Now the mean and variance of the message passed from $f_{\Delta_{l,t}}$ to the variable node $x_{l,t}$ can be computed as,

$$\begin{aligned} &\widehat{x}_{l,t|t-1} = \widehat{f}_{l|t-1}\widehat{x}_{l,t-1|t-1},\ \sigma^2_{l,t|t-1} = |\widehat{f}_{l|t-1}|^2\sigma^2_{l,t-1|t-1} + \\ &\sigma^2_{f_l|t-1}(|\widehat{x}_{l,t-1|t-1}|^2 + \sigma^2_{l,t-1|t-1}) + \widehat{\lambda}^{-1}_{l|t-1}. \end{aligned} \tag{8}$$

$m_{f_{\Delta_{l,t}}\to x_{l,t}}(x_{l,t})$ is not a tractable distribution and thus using EP [13], we project it into the class of Gaussian distribution (ϕ), where the projection operator can be represented as $\text{Proj}_\phi[p] = \arg\min_{q\in\phi} KL(p||q)$. This leads to moment matching (approximated $q \in \mathcal{CN}(x;\mu,\nu)$ has the same mean and variance as p). So we approximate,

$$m_{f_{\Delta_{l,t}}\to x_{l,t}}(x_{l,t}) \approx \mathcal{CN}(x_{l,t}; \widehat{x}_{l,t|t-1}, \sigma^2_{l,t|t-1}).$$

3.2.2. Measurement Update Stage

In the measurement update stage, the posterior for $\boldsymbol{x}_t$ is inferred using BP as in Section 3.1.and we represent the messages by $\widehat{x}^{(t)}_{n,l}, \nu^{(t)}_{n,l}$ and the beliefs by $\widehat{x}_{l,t|t}, \sigma^2_{l,t|t}$. In the measurement stage, the prior for $x_{k,t}$ gets replaced by the belief from the prediction stage and thus the term r_l need to be rewritten as, $r_{l,t}{=}d_{l,t}(\sum_{n=1}^{N} \frac{\widehat{x}^{(t)}_{n,l}}{\nu^{(t)}_{n,l}} + \frac{\widehat{x}_{l,t|t-1}}{\sigma^2_{l,t|t-1}})$.

3.2.3. Lag-1 Smoothing Stage

We show in Lemma 1 that KF is not enough to adapt the hyper parameters, 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). All the hyperparameters λ_l, f_l, γ belong to $\mathcal{A}_{MF}$. Note that the notations $\widehat{f}_{k|t}, \widehat{\lambda}_{k|t}, \widehat{\gamma}_t$ refers to mean of the posteriors (which is equal to the LMMSE point estimates) for the respective hyperparameters at time t and $\sigma^2_{f_k|t}$ represents the posterior variance of f_k at time t. For the smoothing stage, we use BP with Gaussian Markov Random Fields (GMRF) based factorization. GMRF refers to the representation of BP [14], when the underlying Gaussian distribution is expressed in terms of pairwise connections between scalar variables $x_{i,t}$. Substituting the state update equation into the observation model (1), we obtain the system model for the smoothing stage as follows,

$$\boldsymbol{y}_t = \mathbf{A}^{(t)}\boldsymbol{F}\boldsymbol{x}_{t-1} + \widetilde{\boldsymbol{v}}_t,\ \text{where } \widetilde{\boldsymbol{v}}_t = \mathbf{A}^{(t)}\boldsymbol{w}_t + \boldsymbol{v}_t, \tag{9}$$

where $\widetilde{\boldsymbol{v}}_t \sim \mathcal{CN}(0, \widetilde{\boldsymbol{R}}_t)$ with $\widetilde{\boldsymbol{R}}_t = \mathbf{A}^{(t)}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{(t)\,H} + \frac{1}{\gamma}\mathbf{I}$. The joint distribution can be factorized as, $p(\boldsymbol{y}_t, \boldsymbol{\theta}/\boldsymbol{y}_{1:t-1}) = p(\boldsymbol{y}_t/\boldsymbol{\theta})p(\boldsymbol{x}_{t-1}/\boldsymbol{y}_{1:t-1})p(\boldsymbol{F}, \boldsymbol{\Lambda}, \gamma \mid \boldsymbol{y}_{1:t-1})$.

$$\begin{aligned} &\ln p(\boldsymbol{y}_t, \boldsymbol{\theta}/\boldsymbol{y}_{1:t-1}) = \tfrac{-1}{2}\ln\det\widetilde{\boldsymbol{R}}_t - |f_i|^2|x_i|^2\mathbf{A}_{\bar{i}}^{(t)\,T}\widetilde{\boldsymbol{R}}_t^{-1}\mathbf{A}_{\bar{i}}^{(t)} \\ &+2\Re(f_i^H x_i^H \mathbf{A}_{\bar{i}}^{(t)\,H}\widetilde{\boldsymbol{R}}_t^{-1}(\boldsymbol{y}_t - \mathbf{A}_{\bar{i}}^{(t)}\boldsymbol{F}_{\bar{i}}\boldsymbol{x}_{\bar{i},t})) + c_f, \end{aligned} \tag{10}$$

where c_f being the terms independent of f_i, $\mathbf{A}_{\bar{i}}^{(t)}, \boldsymbol{x}_{\bar{i},t}$ represents the matrix $\mathbf{A}^{(t)}$ or the vector $\boldsymbol{x}_t$ with i^{th} column or element removed. Note that we propose to compute $\widehat{\boldsymbol{R}}_t$ by substituting the point estimates of $\boldsymbol{\Lambda}, \gamma$. We also define $\widehat{\boldsymbol{F}}_{\bar{i}|t} = \text{diag}(\widehat{f}_{j|t}, j \neq i)$ with i^{th} element removed. Further applying the MF rule from (4), we write the mean and variance of the resulting Gaussian distribution as,

$$\begin{aligned} &\sigma^{-2}_{f_i|t} = (|\widehat{x}_{i,t-1|t}|^2 + \sigma^2_{i,t-1|t})\mathbf{A}_i^{(t)\,T}\widetilde{\boldsymbol{R}}_t^{-1}\mathbf{A}_i^{(t)}, \\ &\widehat{f}_{i|t} = \sigma^2_{f_i|t}\widehat{x}^H_{i,t-1|t}\mathbf{A}_i^{(t)\,H}\widetilde{\boldsymbol{R}}_t^{-1}(\boldsymbol{y}_t - \mathbf{A}_{\bar{i}}^{(t)}\widehat{\boldsymbol{F}}_{\bar{i}|t}\widehat{\boldsymbol{x}}_{\bar{i},t-1|t}). \end{aligned} \tag{11}$$

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_k, 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 $\boldsymbol{x}_t$ are considered jointly in the factor graph. 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.

Algorithm 1: Combined BP-MF-EP DAR-SBL

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 $\boldsymbol{\Sigma}_{t-1|t-1} = \text{diag}(\sigma^2_{l,t|t-1})$.

for $t = 1 : T$ do

Prediction Stage:

1. Compute $\widehat{x}_{l,t|t-1}, \sigma^2_{l,t|t-1}$ from (8).

Filtering Stage:

1. Compute $\widehat{x}^{(t)}_{n,l}, \nu^{(t)}_{n,l}$ from (5) and update $\widehat{x}_{l,t|t}, \sigma^{-2}_{l,t|t}$ from (6).
2. Compute $\nu^{(t)}_{l,n}, \widehat{x}^{(t)}_{l,n}$ from (7). 3. Continue steps 1) to 2) until convergence.

Smoothing Stage:

Initialization: $\boldsymbol{\Sigma}^{(0)}_{t-1|t} = \boldsymbol{\Sigma}_{t-1|t-1}, \widehat{\boldsymbol{x}}^{(0)}_{t-1|t} = \widehat{\boldsymbol{x}}_{t-1|t-1}$. Define $\mathbf{B}^{(t)} = \boldsymbol{F}^T\mathbf{A}^{(t)\,T}\widetilde{\boldsymbol{R}}_t^{-1}\mathbf{A}^{(t)}\boldsymbol{F} + \boldsymbol{\Sigma}_{t-1|t-1}, \mathbf{h}_t = \boldsymbol{F}^T\mathbf{A}^{(t)\,T}\widetilde{\boldsymbol{R}}_t^{-1}\boldsymbol{y}_t$.

1. $P_{i,j} = \frac{-B^{(t)\,2}_{i,j}}{B^{(t)}_{i,i} + \sum_{k\in\mathcal{N}(i)\setminus j} P_{k,i}}$, $\mu_{i,j} = (h_{i,t} + \sum_{k\in\mathcal{N}(i)\setminus j} P_{k,i}\mu_{k,i}), \forall i,j.$
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: $x'_{k,t} = x_{k,t} - f_k x_{k,t-1}, \zeta_t = \beta\zeta_{t-1} + (1-\beta) < \left\|\boldsymbol{y}_t - \mathbf{A}^{(t)}\boldsymbol{x}_t\right\|^2 >$**):**

1. Compute $\widehat{f}_{l|t}, \sigma^2_{f_l|t}$ from (11), $\widehat{\gamma}_t = \frac{c+N}{(\zeta_t+d)}$ and $\lambda_{l|t} = \frac{(a+1)}{(<|x'_{k,t}|^2>_{|t}+b)}$.

---

4. OPTIMAL PARTITIONING OF BP AND MF NODES

In this section, we show that the partitioning of BP and MF nodes can be characterized through the computation of FIM $= \text{E}(\frac{\partial \ln p(\boldsymbol{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \frac{\partial \ln p(\boldsymbol{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}^H)$. For our analysis, we will allude briefly to an extended concept of Cramer-Rao bound (CRB), the mismatched CRB ($mCRB$) [15] of VB ($mCRB_{VB}$), which is a version of the CRB under model misspecification, and corresponds to the Laplace approximation covariance. Let CRB corresponds to the proper Bayesian CRB and $mCRB_{BP}$ refers to the $mCRB$ for the BP.

Theorem 1 *If the parameter partitioning in VB is such that the different parameter blocks are decoupled at the level of Fisher Information Matrix, then VB is not suboptimal in terms of (mismatched) Cramer-Rao Bound. 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.*

$$\begin{aligned} &mCRB_{BP} = blkdiag(CRB) = blkdiag(FIM^{-1}),\\ &mCRB_{VB} = (blkdiag(FIM))^{-1},\\ &So, mCRB_{BP} = mCRB_{VB}\, if FIM = blkdiag(FIM). \end{aligned} \tag{12}$$

Proof: We briefly outline the proof here. Laplace approximation refers to the evaluation of marginal likelihood or free energy using Laplace's method [16]. This is equivalent to a Gaussian approximation of the posterior $q(\boldsymbol{\theta}_i/\boldsymbol{y})$ around a maximum a posteriori (MAP) estimate ($\boldsymbol{\theta}_i^{(0)}$), motivated by the fact that in the asymptotic limit (large amount of data or high SNR), the posterior approaches a Gaussian around the MAP point. Under the Laplace approximation, the belief becomes $q(\boldsymbol{\theta}_i) = \mathcal{CN}(\boldsymbol{\theta}_i^{(0)}, \boldsymbol{\Sigma}_i^{(0)})$. Further we evaluate the free energy [14] (F denotes the free energy and $L = \ln p(\boldsymbol{y}, \boldsymbol{\theta})$),

$$\begin{aligned} &F = L(\boldsymbol{\theta}^{(0)}) + \tfrac{1}{2}\sum_{i=1}^{M}(G_i + \ln\det \boldsymbol{\Sigma}_i^{(0)} + k_i \ln(2\pi e)),\\ &\ln q_i(\boldsymbol{\theta}_i) = L(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{\bar{i}}) + \tfrac{1}{2}\sum_{j=1, j\neq i}^{M} G_i,\ G_i = \mathrm{tr}\{\boldsymbol{\Sigma}_i \tfrac{\partial}{\partial \boldsymbol{\theta}_i}(\tfrac{\partial L}{\partial \boldsymbol{\theta}_i})^H\}. \end{aligned} \tag{13}$$

Here k_i refers to the number of scalars in $\boldsymbol{\theta}_i$ and $\ln(2\pi e)$ is the entropy of a Gaussian random variable. Now by differentiating $\ln q_i(\boldsymbol{\theta}_i)$ w.r.t the posterior covariance, we obtain the approximate covariances as,

$$\boldsymbol{\Sigma}_i = -(\tfrac{\partial}{\partial \boldsymbol{\theta}_i}(\tfrac{L(\boldsymbol{\theta}^{(0)})}{\partial \boldsymbol{\theta}_i})^H)^{-1} = (\mathrm{blkdiag}(FIM))^{-1} \tag{14}$$

The posterior covariance in (14) is computed by evaluating the Hessian at the variational mode or maximum a posteriori (MAP) point. This variational mode can be obtained as $\boldsymbol{\theta}_i^{(0)} = \max_{\boldsymbol{\theta}_i} \ln q(\boldsymbol{\theta}_i)$. In the Laplace approximation, all pdfs are Gaussian with CRB (portions) as covariance and LMMSE estimates as means. So in the too fine partitioning case, the VB partitioning is applied to the FIM, taking a too fine blockdiagonal part, and since that partitioning is finer than the blockdiagonal FIM structure, then the inverse of the too fine blockdiagonal part of the FIM does not give the correct CRB. So $mCRB_{VB} \neq CRB$. So the nodes in the factor graph are decided based on the partitioning of the blocks in the FIM block diagonal structure, such that the $mCRB_{VB} = CRB$. Here ends the proof.

4.1. Optimal Partitioning for Static SBL:

We define $\mathbf{J}_{\boldsymbol{\theta}_i\boldsymbol{\theta}_j} = \mathrm{E}(\frac{\partial \ln p(\boldsymbol{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i}\frac{\partial \ln p(\boldsymbol{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}^H)$, which represents the part of the FIM which shows the correlation of $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j$. For brevity of notation, we denote $\mathbf{J}_{\boldsymbol{\theta}_i\boldsymbol{\theta}_i} = \mathbf{J}_{\boldsymbol{\theta}_i}$. First we consider the static case when $f_l = 0, \forall l$. We omit the index t for simplicity. $f_{\alpha_i}(\alpha_i) = p(\alpha_i \mid a, b), \alpha_i = \lambda_i$ represents the prior distribution of the precision parameter α_i which is chosen as Gamma.

$$FIM = \mathbf{J}_s = \begin{bmatrix} \mathbf{A}^H\mathbf{A} + \mathbf{I} & & \mathbf{0}_M \\ \mathbf{0}_M & \mathbf{J}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & \mathbf{0}_M \\ \mathbf{0}_M & \mathbf{0}_M & \mathbf{J}_{\gamma\gamma} \end{bmatrix} \tag{15}$$

The non block diagonal elements of the FIM are crosscorrelation as follows, $\mathbf{J}_{\gamma\boldsymbol{x}} = \mathrm{E}(\frac{\partial \ln p(\boldsymbol{y},\boldsymbol{x},\gamma,\boldsymbol{\Gamma})}{\partial \gamma}\frac{\ln p(\boldsymbol{y},\boldsymbol{x},\gamma,\boldsymbol{\Gamma})}{\partial \boldsymbol{x}}^H) = (N/\gamma - \boldsymbol{v}^H\boldsymbol{v})(\gamma\mathbf{A}^H\boldsymbol{v} - \boldsymbol{\Gamma}x) = 0$. Similarly the crosscorrelation between $\boldsymbol{x}$ and $\boldsymbol{\Gamma}$ will be zero and also for $\boldsymbol{\Gamma}$ and γ. The cross correlations are zero because of zero mean circularly symmetric complex Gaussian variables because 3rd order moments of zero mean $\boldsymbol{v}$ and $\boldsymbol{x}$ are zero. Thus the resulting FIM will be block diagonal. In this block diagonal structure, the crosscorrelation matrix $\mathbf{J}_{\boldsymbol{x}\boldsymbol{x}} = \mathbf{A}^H\mathbf{A} + \mathbf{I}$ will be full and thus requires the estimation of $\boldsymbol{x}$ using BP, while scalar factors which are decoupled γ, α_i can be estimated using MF. This explains the optimality of our BP-MF partitioning as shown in the Figure 1.

4.2. Optimal Partitioning for DAR-SBL:

In this section we formulate the optimal partitioning between VB and BP for the dynamic SBL case. Here we need to consider the FIMs recursively, i.e. FIM of the time update stage followed by the measurement stage. For the time update stage, we abbreviate $p(x_{k,t}, x_{k,t-1}, f_k, \lambda_k/\boldsymbol{y}_{1:t-1}) = p$ for convenience here. $\ln p = \ln\lambda_k - \lambda_k|x_{k,t} - f_k x_{k,t-1}|^2 - \sigma^{-2}_{k,t-1|t-1}|x_{k,t-1} - \widehat{x}_{k,t-1|t-1}|^2 + \sum_{k=1}^{M}\ln q_{\lambda_k}(\lambda_k)$. The measurement FIM (15) is the prior FIM for the next time update. Thus it follows that BP is needed for the inference of $\boldsymbol{x}_t$ and MF for γ. One remark here is that the prior $\boldsymbol{x}_t$ covariance for the measurement update is the inverse FIM of the time update and is diagonal here.

Lemma 1 *The AR(1) model parameters require (at least lag* 1*) smoothing for identifiability.*

Proof: Considering, with augmented state $\boldsymbol{\theta}_t = [\boldsymbol{x}_t; \mathbf{f}; \mathrm{diag}(\boldsymbol{\Lambda}); \gamma]$ ($3M + 1$ dimensional), we obtain the FIM, $\mathbf{J}_t = \mathrm{blkdiag}(\mathbf{J}_{\boldsymbol{x},t}, \mathbf{J}_{\boldsymbol{F},t}, \mathbf{J}_{\boldsymbol{\Lambda},t}, \mathbf{J}_{\gamma,t})$. In [17], Tichavský et al. derived an elegant recursive approach to calculate the FIM recursions for a general discrete-time nonlinear filtering problem. Based on a similar derivation, we arrive at the following recursions for the sequence $\mathbf{J}_{\boldsymbol{\theta}_i,t}$ of posterior information submatrices for estimating $\boldsymbol{\theta}_i$,

$$\begin{aligned} &\mathbf{J}_{\boldsymbol{x},t} = \boldsymbol{\Lambda} + \gamma\mathbf{A}^{(t)\,H}\mathbf{A}^{(t)} + \boldsymbol{\Lambda}\boldsymbol{F}(\boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^H + \mathbf{J}_{x,t-1})^{-1}\boldsymbol{\Lambda}\boldsymbol{F}^H,\\ &\mathbf{J}_{\boldsymbol{F},t} = \mathbf{J}_{\boldsymbol{F},t} + \mathbf{D} - \mathbf{J}_{\boldsymbol{F}\boldsymbol{x},t}(\boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^H + \mathbf{J}_{\boldsymbol{F},t-1})^{-1}\mathbf{J}^T_{\boldsymbol{x}\boldsymbol{F},t}\\ &\text{with } \mathbf{D} = (\mathbf{I} - \boldsymbol{F}\boldsymbol{F}^H)^{-1},\ \mathbf{J}_{\boldsymbol{x}\boldsymbol{F},t} = \boldsymbol{F}\boldsymbol{\Lambda}[\mathbf{J}_{\boldsymbol{x},t} + \boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}]^{-1}\mathbf{J}_{\boldsymbol{x}\boldsymbol{F}},\\ &\mathbf{J}_{\boldsymbol{\Lambda},t} = \mathbf{D} - \mathbf{D}(\mathbf{D} + \mathbf{J}_{\boldsymbol{\Lambda},t-1})^{-1}\mathbf{D} \text{ with } \mathbf{D} = \boldsymbol{\Lambda}^{-2},\ \mathbf{J}_{\gamma,t} = N/\gamma^2. \end{aligned} \tag{16}$$

Note that $if \mathbf{J}_{\boldsymbol{x}\boldsymbol{F},-1} = 0$, then $\mathbf{J}_{\boldsymbol{x}\boldsymbol{F},t} = 0, \forall t \geq 0$. FIM recursions show that filtering may be enough for the estimation of AR(1) parameters. However, closely looking at the expressions for $\widehat{f}_{k|t}$ derived in our work [12, eq. (24-25)] shows that $\widehat{f}_{k|t} = f_k$. This implies that we need to know the true f_k to estimate it, in the joint estimation framework. Further to prove the unidentifiability, we use the concept of global identifiability provided in [18].

$$\begin{aligned} &p(\mathbf{f}/\boldsymbol{x}_t, \boldsymbol{y}_t) = p(\boldsymbol{y}_t/\boldsymbol{x}_t)p(\boldsymbol{x}_t/\mathbf{f})p(\mathbf{f})/p(\boldsymbol{y}_t, \boldsymbol{x}_t)\\ &= p(\boldsymbol{x}_t/\mathbf{f})p(\mathbf{f})/\textstyle\int p(\boldsymbol{x}_t/\mathbf{f})p(\mathbf{f})d\mathbf{f} = p(\mathbf{f}/\boldsymbol{x}_t) \end{aligned} \tag{17}$$

The above expression (17) suggests that posterior of $\mathbf{f}$ given $\boldsymbol{x}_t$ does not depend on $\boldsymbol{y}_t$ or in other words the observations doesn't provide any extra information about $\mathbf{f}$ other than the prior $p(\mathbf{f}/\boldsymbol{x}_t)$ and hence $\mathbf{f}$ is globally not identifiable. This proves the Lemma. (17) also shows that $\mathbf{f}, \boldsymbol{x}_t$ are coupled in the estimation unlike the decoupling property shown by the FIM analysis.

Few remarks follows: Th $mCRB$ analysis in Theorem 1 indicates that the $\boldsymbol{x}$ part needs to be treated jointly, motivating joint VB or BP. We conjecture that whatever local identifiability analysis indicates as necessitating joint treatment for optimality requires indeed joint treatment. But local analysis may not capture all dependencies. The local analysis (recursive CRB) shows that filtering would be sufficient for local identifiability of $\mathbf{f}$ and that the f_i and the x_i are decoupled. However, global identifiability analysis reveals that filtering is not enough for identifiability of $\mathbf{f}$ and that the estimation of x_i and f_i is coupled. The gap between local and global analysis may perhaps be reflected in the observation that the hyperparameters could be estimated (in what corresponds to filtering) by Type-II Maximum Likelihood (ML) [19] (ie ML for hyperparameters, with the random parameters x integrated out). Such Type-II ML approach for hyperparameter estimation in the dynamic problem considered here will be investigated further in future work.

Corollary 1.1 *For the smoothing stage* (9)*, an optimal partitioning is to apply BP for estimation of the sparse vector, $\widehat{\boldsymbol{x}}_{t-1|t}$ and MF for the correlation coefficient $\boldsymbol{F}$.*

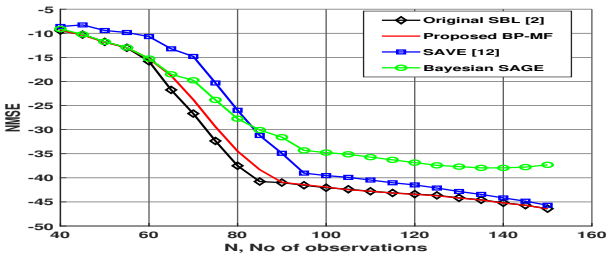

Fig. 2: Static SBL: NMSE as a function of N.

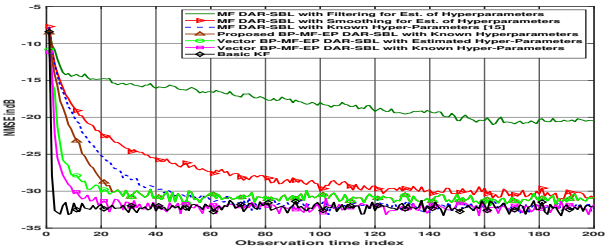

Fig. 3: DAR-SBL: NMSE as a function of time.

Proof: The FIM recursions for smoothing stage can be obtained as (detailed derivation is skipped due to space constraints), $\mathbf{J}_t = \text{blkdiag}(\mathbf{J}_{\boldsymbol{x},t}, \mathbf{J}_{\boldsymbol{F},t}, \mathbf{J}_{p,t})$, where $\mathbf{J}_{p,t}$ representing the information submatrix for the precision parameters $\boldsymbol{\Lambda}, \gamma$. We obtain $\mathbf{J}_{\boldsymbol{x},t} = \boldsymbol{F}^T \mathbf{A}^{(t)\,H} \widetilde{\boldsymbol{R}}_t^{-1} \mathbf{A}^{(t)} \boldsymbol{F} + \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{F}(\boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^H + \mathbf{J}_{\boldsymbol{x},t-1})^{-1}\boldsymbol{\Lambda}\boldsymbol{F}^H$, which is a full matrix.

$$
\begin{aligned}
&\mathbf{J}_{\boldsymbol{F},t} = \mathbf{J}_{\boldsymbol{F},t-1} + \boldsymbol{\Gamma}\text{diag}(\mathbf{A}^{(t)\,H}\widetilde{\boldsymbol{R}}_t^{-1}\mathbf{A}^{(t)}) + \mathbf{D} - \\
&\mathbf{J}_{\boldsymbol{F}\boldsymbol{x},t}(\mathbf{D} + \mathbf{J}_{\boldsymbol{F},t-1})^{-1}\mathbf{J}_{\boldsymbol{x}\boldsymbol{F},t}, \\
&\text{with } \mathbf{D} = (\mathbf{I} - \boldsymbol{F}\boldsymbol{F}^H)^{-1},\ \mathbf{J}_{\boldsymbol{x}\boldsymbol{F},t} = \boldsymbol{\Lambda}\boldsymbol{F}[\mathbf{J}_{\boldsymbol{x},t} + \boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}]^{-1}\mathbf{J}_{\boldsymbol{x}\boldsymbol{F},t}, \\
&\mathbf{J}_{p,t} = \begin{bmatrix} \mathbf{J}_{\boldsymbol{\Lambda},t} & \mathbf{J}_{\boldsymbol{\Lambda}\gamma,t} \\ \mathbf{J}_{\boldsymbol{\Lambda}\gamma,t} & \mathbf{J}_{\gamma\gamma} \end{bmatrix},\ \text{where, } \mathbf{J}_{\gamma\gamma} = \tfrac{1}{\gamma^4}\text{tr}\{\widetilde{\boldsymbol{R}}_t^{-1}\widetilde{\boldsymbol{R}}_t^{-1}\}, \\
&\mathbf{J}_{\boldsymbol{\Lambda},t} = \mathbf{C}_{\boldsymbol{\Lambda},t} + \mathbf{D} - \mathbf{D}(\mathbf{D} + \mathbf{J}_{\boldsymbol{\Lambda},t-1})^{-1}\mathbf{D} \text{ with } \mathbf{D} = \boldsymbol{\Lambda}^{-2}, \\
&(\mathbf{C}_{\boldsymbol{\Lambda},t})_{i,j} = \tfrac{1}{\lambda_i^2\lambda_j^2}\text{tr}\{\widetilde{\boldsymbol{R}}_t^{-1}\mathbf{A}_i^{(t)}\mathbf{A}_i^{(t)\,H}\mathbf{A}_j^{(t)}\mathbf{A}_j^{(t)\,H}\widetilde{\boldsymbol{R}}_t^{-1}\}, \\
&\mathbf{J}_{\boldsymbol{\Lambda}\gamma,t} = \mathbf{c}_{\boldsymbol{\Lambda}\gamma,t},\ (\mathbf{c}_{\boldsymbol{\Lambda}\gamma,t})_i = \tfrac{1}{\lambda_i^2\gamma^2}\text{tr}\{\widetilde{\boldsymbol{R}}_t^{-1}\mathbf{A}_i^{(t)}\mathbf{A}_i^{(t)\,H}\widetilde{\boldsymbol{R}}_t^{-1}\}.
\end{aligned} \tag{18}
$$

$(\mathbf{c}_{\boldsymbol{\Lambda}\gamma,t})_i$ represents the i^{th} element of the vector $\mathbf{c}_{\boldsymbol{\Lambda}\gamma,t}$. Here also, if $\mathbf{J}_{\boldsymbol{x}\boldsymbol{F},-1} = \mathbf{0}$, then $\mathbf{J}_{\boldsymbol{x}\boldsymbol{F},t} = \mathbf{0}, \forall t$. Thus the FIM for $\boldsymbol{x}_t$ is full and it follows from Theorem 1 that optimal partitioning is to apply BP for $\boldsymbol{x}_t$ and MF for the correlation coefficient $\boldsymbol{F}$ (since $\mathbf{J}_{\boldsymbol{F},t}$ is diagonal and also positive definite at any time instant t) in the smoothing stage. Here ends the proof.

5. SIMULATION RESULTS

For the observation model, the parameters chosen are $N = 100$, $M = 200$, $K = 30$. All signals are considered to be real in the simulation. All the elements of $\mathbf{A}^{(t)}$ (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 $\sqrt{30}$ so that the signal part of any scalar observation has unit variance. Taking the SNR to be 20dB, the variance of each element of $\boldsymbol{v}_t$ (Gaussian with mean 0) is computed as 0.01.

Consider the state update, $\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t$. To generate $\boldsymbol{x}_0$, the first 30 elements are chosen as Gaussian (mean 0 and variance 1) and then the remaining elements of the vector $\boldsymbol{x}_0$ are put to zero. Then the elements of $\boldsymbol{x}_0$ are randomly permuted to distribute the 30 non-zero elements across the whole vector. The diagonal elements of $\boldsymbol{F}$ are chosen uniformly in $[0.9, 1)$. Then the covariance of $\boldsymbol{w}_t$ can be computed as $\boldsymbol{\Gamma}(\mathbf{I} - \boldsymbol{F}\boldsymbol{F}^H)$. Note that $\boldsymbol{\Gamma}$ contains the variances of the elements of $\boldsymbol{x}_t$ (including $t = 0$), where for the non-zero elements of $\boldsymbol{x}_0$ the variance is 1. Following observations can be made from the simulations. In Figure 2, for SBL with estimated hyperparameters, there is substantial improvement in normalized MSE (NMSE) by using BP instead of MF method for estimating $\boldsymbol{x}$. Bayesian SAGE (Space Alternating Generalized EM) corresponds to the application of [9] to SBL. 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 $\boldsymbol{x}_t$. Also we show that there is a drastic improvement in performance with lag-1 smoothing for hyperparameter estimation compared to just using filtering.

6. CONCLUSIONS

We presented a fast SBL algorithm called BP-MF-EP DAR-SBL, which uses a combination of BP, MF and EP techniques to approximate the posteriors of the data and parameters and track a time varying sparse signal. BP-MF-EP DAR-SBL helps to circumvent the matrix inversion operation required in the original SBL algorithm. We propose for the first time in the literature an optimal way to select the partitioning of BP and MF nodes with CRB as a performance evaluation criteria. Future work include extension of the combined BP-MF framework for Kronecker structured dictionary learning [20].

7. REFERENCES

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

[2] D. P. Wipf and B. D. Rao, "Sparse Bayesian Learning for Basis Selection ," *IEEE Trans. on Sig. Proc.*, Aug. 2004.

[3] Z. Zhang and B. D. Rao, "Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning ," *IEEE J. of Sel. Topics in Sig. Process.*, Sept. 2011.

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

[5] E. Riegler et al., "Merging belief propagation and the mean field approximation: a free energy approach ," *IEEE Trans. on Info. Theo.*, vol. 59, no. 1, Jan. 2013.

[6] D. L. Donoho et al., "Message-passing algorithms for compressed sensing ," *PNAS*, vol. 106, Nov. 2009.

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

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

[9] Bernard H. Fleury et al., "Channel Parameter Estimation in Mobile Radio Environments Using the SAGE Algorithm ," *IEEE J. on Sel. Areas in Commun.*, vol. 17, no. 3, March 1999.

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

[11] C. K. Thomas and D. Slock, "Space alternating variational Bayesian learning for LMMSE filtering," in *EUSIPCO*, 2018.

[12] C. K. Thomas and D. Slock, "Gaussian variational Bayes Kalman filtering for dynamic sparse Bayesian learning," in *ITISE*, 2018.

[13] T. Minka, "A family of algorithms for approximate bayesian inference," in *Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, USA*, 2001.

[14] J. S. Yedidia et al., "Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms," *IEEE Trans. On Info. Theo.*, vol. 51, July 2005.

[15] S. Fortunati et al., "Performance bounds for parameter estimation under misspecified models," *IEEE Sig. Proc. Mag.*, Nov. 2017.

[16] R. E. Kass and A. E. Raftery, "Bayes factors ," *J. Am. Stat. Assoc.*, vol. 90, 1995.

[17] P. Tichavský, C. H. Muravchik, and A. Nehorai, "Posterior Cramer Rao Bounds for Discrete-Time Nonlinear Filtering," *IEEE Trans. On Sig. Process.*, vol. 46, May 1998.

[18] A. E. Gelfand and S. K. Sahu, "Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models," *Journ. of the Americ. Stat. Assoc.*, Mar. 1999.

[19] R. Giri and B. Rao, "Type I and Type II Bayesian Methods for Sparse Signal Recovery Using Scale Mixtures," *IEEE Trans. On Sig. Process.*, Jul. 2016.

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