BP-VB-EP Based Static and Dynamic Sparse Bayesian Learning with Kronecker Structured Dictionaries

Christo Kurisummoottil Thomas, Dirk Slock

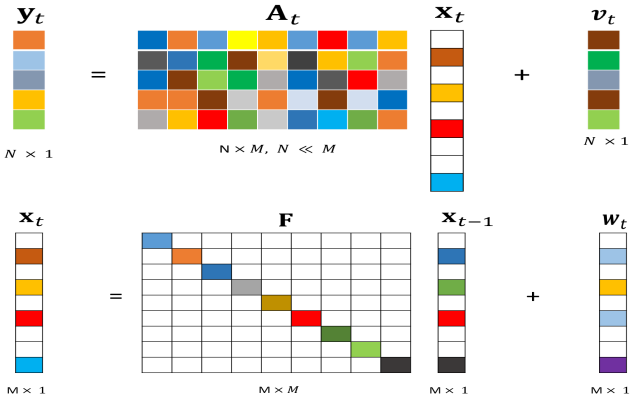
Communication Systems Department, EURECOM, Sophia Antipolis, France

ICASSP 2020, May 4-8,
Special Session:
"AMP and other Approximate Bayesian Inference Techniques"

Outline

Time Varying Sparse State Tracking

Sparse signal $\boldsymbol{x}_t$ is modeled using an AR(1) process with diagonal correlation coefficient matrix $\boldsymbol{F}$.



Define: $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$, $\boldsymbol{F} = \text{diag}(\mathbf{f})$.

f_i : correlation coefficient and $x_{i,t} \sim \mathcal{CN}(0, \frac{1}{\lambda_i})$. Further, $\boldsymbol{w}_t \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Gamma}^{-1} = \boldsymbol{\Lambda}^{-1}(\mathbf{I} - \boldsymbol{F}\boldsymbol{F}^H))$ and $\boldsymbol{v}_t \sim \mathcal{CN}(\mathbf{0}, \gamma^{-1}\mathbf{I})$. VB leads to Gaussian SAVE-Kalman Filtering (GS-KF).

Applications: Localization, Adaptive Filtering.

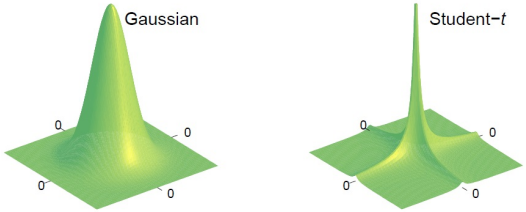
Sparsification of the Innovation Sequence (Sparse Bayesian Learning-SBL)

- Bayesian Compressed Sensing: 2-layer hierarchical prior for $\boldsymbol{x}$ as in [Tipping:JMLR01], inducing sparsity for $\boldsymbol{x}$.

$$p(x_{i,t}|\lambda_i) = \mathcal{N}(x_{i,t}; 0, \lambda_i^{-1}),\ p(\lambda_i/a, b) = \Gamma^{-1}(a) b^a \lambda_i^{a-1} e^{-b\lambda_i}$$

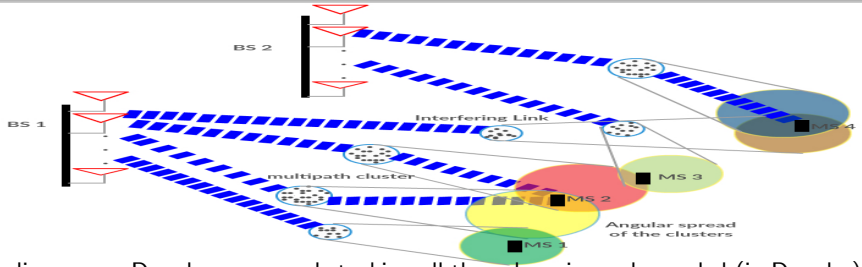
$\Rightarrow$ sparsifying Student-t marginal

$$p(x_{i,t}) = \frac{b^a \Gamma(a+\frac{1}{2})}{(2\pi)^{\frac{1}{2}}\Gamma(a)} (b + w_i^2/2)^{-(a+\frac{1}{2})}$$



- We apply the (Gamma) prior not to the precision of the state $\boldsymbol{x}$ but it's innovation $\boldsymbol{w}$, allowing to sparsify at the same time the components of x AND their variation in time (innovation).

Tensor Representation (Channel Tracking in MaMIMO OFDM)

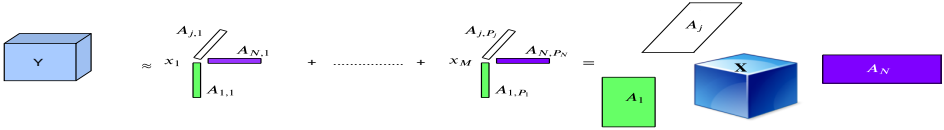


- Sampling across Doppler space and stacking all the subcarrier and sampled (in Doppler) elements as a vector

$$\text{vec}(\mathbf{H}(t)) = \sum_{i=1}^{L} x_{i,t}\, \mathbf{h}_t(\phi_i) \otimes \mathbf{h}_r(\theta_i) \otimes \boldsymbol{v}_f(\tau_i) \otimes \boldsymbol{v}_t(f_i) = \mathbf{A}(t)\boldsymbol{x}_t$$

- 4-D Tensor model, Delay, Doppler and Tx/Rx spatial dimensions.
- Array response itself: Kronecker structure in the case of polarization or in the case of $2D$ antenna arrays with separable structure [Sidiropoulos:icassp18].
- User mobility changes the scattering geometry and path coefficients.
- Tensor based KF proposed here avoids the off-grid basis mismatch issues.

Kronecker Structured Tensor Models



- Tensor signals appear in many applications: massive multi-input multi-output (MIMO) radar, massive MIMO (MaMIMO) channel estimation, speech processing, image and video processing.
- Retaining the tensorial structure in the received signal can be beneficial compared to processing its unstructured matrix version.
- The signal model for the recovery of a time varying sparse signal under Kronecker structured (KS) dictionary matrix can be formulated as

$$\text{Observation: } \boldsymbol{y}_t = \underbrace{(\mathbf{A}_1(t) \otimes \mathbf{A}_2(t) .... \otimes \mathbf{A}_N(t))}_{\mathbf{A}(t)} \boldsymbol{x}_t + \boldsymbol{v}_t,\ \boldsymbol{y}_t = vec(\mathbf{Y}_t)$$

$$\text{State Update: } \boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t,$$

$\mathbf{Y}_t \in \mathcal{C}^{I_1 \times I_2 \cdots \times I_N}$ is the observations or data at time t, $\mathbf{A}_{j,i}(t) \in \mathcal{C}^{I_j}$, the factor matrix $\mathbf{A}_j(t) = [\mathbf{A}_{j,1}(t), ..., \mathbf{A}_{j,P_j}(t)]$ and the overall unknown parameters are $[\![\mathbf{A}_1(t), ..., \mathbf{A}_N(t); \boldsymbol{x}_t]\!]$, $\boldsymbol{x}_t$ is $M(= \prod_{j=1}^{N} P_j)$-dimensional sparse center tensor and $\boldsymbol{w}_t$, $\boldsymbol{v}_t$ are the state or measurement noise.

Our Contribution Here Compared to State of the Art

- Extension of our previous work[1] [ThomasSlock:ICASSP2019] on Kronecker Structured (KS) dictionary learning (DL) - called as SAVED-KS DL (Space Alternating Variational Estimation for KS DL) to the case of time varying sparse state vector.
- Variational Bayes (VB) at the scalar level i.e mean field (MF) for $\boldsymbol{x}_t$ and at the column level for the factor matrices in the measurement matrix is quite suboptimal since the posterior covariance computed does not converge to the true posterior [ThomasSlock:DSW2018]. One potential solution: belief propagation (BP).
- Also, better variational free energy (VFE) approximation, e.g. Belief Propagation (BP) or Expectation Propagation (EP) instead of MF.
- Inspired by the framework of [RieglerFleury:TIT2013], we combine BP and MF approximations in such a way as to optimize the message passing (MP) framework.
- The main focus: reduced complexity (e.g. VB leads to the low complexity) algorithms without sacrificing much on the performance (BP for performance improvement).
- Dynamic autoregressive SBL (DAR-SBL): a case of joint Kalman filtering (KF) with a linear time-invariant diagonal state-space model, and parameter estimation $\Longrightarrow$ an instance of nonlinear filtering.

[1] abscd

Outline

Variational Free Energy (VFE) Framework

- Intractable joint posterior distribution of the parameters $\theta = \{\boldsymbol{x}, \mathbf{A}, \mathbf{f}, \boldsymbol{\lambda}, \gamma\}$.
- Actual posterior: $p(\boldsymbol{\theta}) = \frac{1}{Z} \underbrace{\prod_{a \in \mathcal{A}_{BP}} f_a(\boldsymbol{\theta}_a) \prod_{b \in \mathcal{A}_{MF}} f_b(\boldsymbol{\theta}_b)}_{\text{factor nodes}}$, where $\mathcal{A}_{BP}, \mathcal{A}_{MF}$ = set of variable nodes belonging to the BP/MF part with $\mathcal{A}_{BP} \cap \mathcal{A}_{MF} = \emptyset$.
- The whole $\boldsymbol{\theta}$ is partitioned into the set θ_i, and we want to approximate the true posterior $p(\boldsymbol{\theta})$ by an approximate posterior $q(\boldsymbol{\theta}) = \prod_i q_i(\theta_i)$. The individual variable portions θ_i are the variable nodes in a factor graph.
- $\mathcal{N}_{BP}(i)$, $\mathcal{N}_{MF}(i)$– the set of neighbouring factor nodes of variable node i which belong to the BP/MF part.
- $\mathcal{I}_{MF} = \bigcup_{a \in \mathcal{A}_{MF}} \mathcal{N}(a)$, $\mathcal{I}_{BP} = \bigcup_{a \in \mathcal{A}_{BP}} \mathcal{N}(a)$. $\mathcal{N}(a)$– the set of neighbouring variable nodes of any factor node a.
- The resulting Free Energy (Entropy – Average Energy) obtained by the combination of BP and MF are written as below (let $q_i(\theta_i)$ represents the belief about θ_i (the approximate posterior)),

$$F_{BP,MF} = \sum_{a \in \mathcal{A}_{BP}} \sum_{\boldsymbol{\theta}_a} q_a(\boldsymbol{\theta}_a) \ln \frac{q_a(\boldsymbol{\theta}_a)}{f_a(\boldsymbol{\theta}_a)} - \sum_{b \in \mathcal{A}_{MF}} \sum_{\mathbf{x}_b} \prod_{i \in \mathcal{N}(b)} q_i(\theta_i) \ln f_b(\boldsymbol{\theta}_b) - \sum_{i \in \mathcal{I}} (|\mathcal{N}_{BP}(i)| - 1) \sum_{\theta_i} q_i(\theta_i) \ln q_i(\theta_i).$$

Message Passing Expressions

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

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

- The fixed point equations corresponding to the constrained optimization of the approximate VFE can be written as follows [RieglerFleuryTIT2013],

$$q_i(\theta_i) = z_i \prod_{a \in \mathcal{N}_{BP}(i)} m^{BP}_{a \to i}(\theta_i) \prod_{a \in \mathcal{N}_{MF}(i)} m^{MF}_{a \to i}(\theta_i), \implies \text{Product of incoming beliefs}$$

$$n_{i \to a}(\theta_i) = \prod_{c \in \mathcal{N}_{BP}(i) \setminus a} m_{c \to i}(\theta_i) \prod_{d \in \mathcal{N}_{MF}(i)} m_{d \to i}(\theta_i), \implies \text{variable to factor nodes}$$

$$m^{MF}_{a \to i}(\theta_i) = \exp(< \ln f_a(\theta_a) >_{\prod_{j \in \mathcal{N}(a) \setminus i} n_{j \to a}(\theta_j)}),$$

$$m^{BP}_{a \to i}(\theta_i) = (\int \prod_{j \in \mathcal{N}(a) \setminus i} n_{j \to a}(\theta_j) f_a(\theta_a) \prod_{j \neq i} d\theta_j), \implies \text{factor to variable nodes}$$

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

Gaussian BP-MF-EP KF

- Proposed Method: Alternating optimization between non linear KF for the sparse states (plus the hyperparameters) and BP for dictionary learning (DL).
- **Diagonal AR(1) (DAR(1)) Prediction Stage:** Since there is no coupling between the scalars in the state update , it is enough to update the prediction stage using MF. However, the interaction between $x_{l,t}$ and f_l requires Gaussian projection, using expectation propagation (EP). More details in [ThomasSlock:Asilomar2019].
- y_n- factor node, x_l- variable node. (l,n) or (n,l) to represent the messages passed from l to n or viceversa. Gaussian messages from y_n to x_l parameterized by mean $\widehat{x}_{n,l}^{(t)}$ and variance $\nu_{n,l}^{(t)}$.
- **Measurement Update (Filtering) Stage:** For the measurement update stage, the posterior for $\boldsymbol{x}_t$ is inferred using BP. In the measurement stage, the prior for $x_{l,t}$ gets replaced by the belief from the prediction stage. We define $d_{l,t}=(\sum_{n=1}^{N} {\nu_{n,l}^{(t)}}^{-1})^{-1}$, $r_{l,t}=d_{l,t}(\sum_{n=1}^{N} \frac{\widehat{x}_{n,l}^{(t)}}{\nu_{n,l}^{(t)}} + \frac{\widehat{x}_{l,t|t-1}}{\sigma^2_{l,t|t-1}})$.

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

Lag-1 Smoothing Stage for Correlation Coefficient $\mathbf{f}$

$$\boldsymbol{y}_t = \mathbf{A}(t)\boldsymbol{F}\boldsymbol{x}_{t-1} + \widetilde{\boldsymbol{v}}_t, \text{ where } \widetilde{\boldsymbol{v}}_t = \mathbf{A}(t)\boldsymbol{w}_t + \boldsymbol{v}_t,\ \widetilde{\boldsymbol{v}}_t \sim \mathcal{CN}(0, \widetilde{\boldsymbol{R}}_t)$$

- We show in Lemma 1 of [ThomasSlock:Asilomar2019] 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.
- Gaussian Posterior for $\boldsymbol{x}_t$:

$$\begin{aligned}
&\sigma^{-2,(i)}_{k,t-1|t} = (\widehat{f}^2_{k|t} + \sigma^2_{f_k|t})\mathbf{A}^H_k(t)\widetilde{\boldsymbol{R}}_t^{-1}\mathbf{A}_k(t) + \sigma^{-2}_{k,t-1|t-1}, \\
&< x^{(i)}_{k,t-1|t} > = \sigma^{2,(i)}_{k,t-1|t}(\widehat{f}^H_{k|t}\mathbf{A}^H_k(t)\widetilde{\boldsymbol{R}}_t^{-1}(\boldsymbol{y}_t - \mathbf{A}_{\overline{k}}(t)\boldsymbol{F}_{\overline{k}|t} < \boldsymbol{x}^{(i-1)}_{\overline{k},t-1|t} >) + \tfrac{\widehat{x}_{k,t-1|t-1}}{\sigma^2_{k,t-1|t-1}}).
\end{aligned}$$

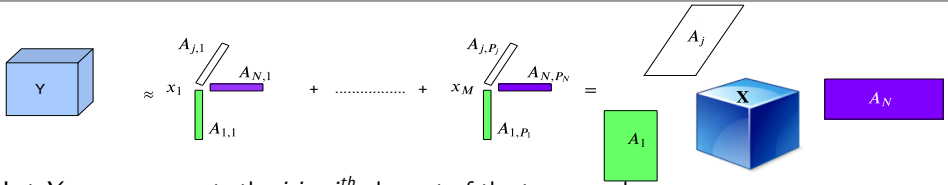
- Applying the MF rule, the resulting Gaussian distribution for $\mathbf{f}$ has mean, $\sigma^{-2}_{f_i|t}$ and variance, $\widehat{\mathbf{f}}_{i|t}$, the detailed derivations for which are in [ThomasSlock:Asilomar2019].

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

- $\widetilde{\boldsymbol{R}}_t = \mathbf{A}(t)\boldsymbol{\Lambda}^{-1}\mathbf{A}(t)^H + \frac{1}{\gamma}\mathbf{I}$.

Outline

Dictionary Learning using Tensor Signal Processing



- Let $Y_{i_1,\ldots,i_N}$ represents the $i_1 i_2 \ldots i_N^{th}$ element of the tensor and $\boldsymbol{y} = [y_{1,1,\ldots,1}, y_{1,1,\ldots,2}, \ldots, y_{I_1,I_2,\ldots,I_N}]^T$, then it can be verified that [Sidiropoulos:TSP17],
  $$\boldsymbol{y}_t = (\mathbf{A}_1(t) \otimes \mathbf{A}_2(t) \ldots \otimes \mathbf{A}_N(t))\boldsymbol{x}_t + \boldsymbol{w}_t, \boldsymbol{w} \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I}),$$
  Matrix Unfolding:$\mathbf{Y}^{(n)} = \mathbf{A}_n(t)\mathbf{X}^{(n)}(\mathbf{A}_N(t) \otimes \ldots \mathbf{A}_{n+1}(t) \otimes \mathbf{A}_{n-1}(t) \ldots \otimes \mathbf{A}_1(t))^T$.
- $\mathbf{A}_j(t)$ is of dimension, $I_j \times P_j$ and the resulting Tensor is $\mathcal{C}^{I_1 \times \ldots \times I_N}$.
- Retaining the Tensor structure in the dictionary matrix leads to better estimates than using the matricized version for $\mathbf{A}$ and learning it.
- Less free variables to be estimated in the Tensor structured case.
- Variational Bayesian Inference using the following approximate posterior

$$q(\boldsymbol{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}) = q_\gamma(\gamma) \prod_{i=1}^{M} q_{x_i}(x_i) \prod_{i=1}^{M} q_{\alpha_i}(\alpha_i) \prod_{j=1}^{N} \prod_{i=1}^{P_j} q_{\mathbf{A}_{j,i}}(\mathbf{A}_{j,i}), \implies \text{SAVED-KS DL Or}$$

$$q(\boldsymbol{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}) = q_\gamma(\gamma) \prod_{i=1}^{M} q_{x_i}(x_i) \prod_{i=1}^{M} q_{\alpha_i}(\alpha_i) \prod_{j=1}^{N} q_{\mathbf{A}_j}(\mathbf{A}_j) \implies \text{Joint VB for DL}$$

Suboptimality of SAVED-KS DL and Joint VB

- From the expression for the error covariance in the estimation of the factor $\mathbf{A}_{j,i}$ (SAVED-KS DL in [ThomasSlock:ICASSP19]) $(\mathrm{tr}\{(\bigotimes_{k=N,k\neq j}^{1} < \mathbf{A}_k^T(t)\mathbf{A}_k^*(t) >) < \mathbf{X}^{(j)\,T}\mathbf{X}^{(j)} >\}\mathbf{I})$, $\Longrightarrow$ it does not take into account the estimation error in the other columns of $\mathbf{A}_j(t)$. The columns of $\mathbf{A}_j(t)$ can be correlated, for e.g. if we consider two paths (say i, j) with same DoA but with different delays, the delay responses $\boldsymbol{v}_f(\tau_i(t))$ and $\boldsymbol{v}_f(\tau_i(t))$ may be correlated.
- The joint VB estimates (mean and covariance) can be obtained as

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

where $\mathbf{V}_j =< \mathbf{X}^{(j)} >< (\bigotimes_{k=N,k\neq j}^{1} \mathbf{A}_k(t))^T >$ and $\mathbf{B}_j$ is defined as with the first row of $(\mathbf{Y}^{(j)}\mathbf{V}_j^T)$ 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 $\mathbf{A}_j(t)$ by BP, since each column estimate can be expressed as the solution of a linear system of equation from (1), $\widehat{\mathbf{A}}_{j,i}^T(t) = \boldsymbol{\Psi}_j^{-1}\mathbf{b}_{j,i}$. $\mathbf{b}_{j,i}$ represents the i^{th} column of $\mathbf{B}_j^T$.

Optimal Partitioning of the Measurement Stage and KS DL

Lemma

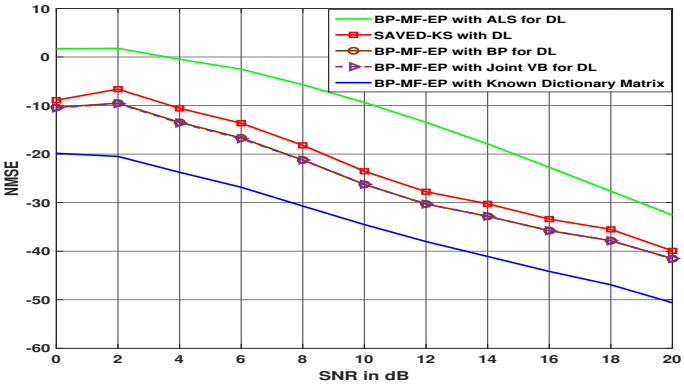
For the measurement stage, an optimal partitioning is to apply BP for the sparse vector $\boldsymbol{x}_t$ and VB (SAVED-KS) for the columns of the factor matrices $\mathbf{A}_{j,i}(t)$ assuming the vectors $\mathbf{A}_{j,i}(t)$ are independent and have zero mean. However, if the columns of $\mathbf{A}_j(t)$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

- Proof: Follows from [KalyanSlock:EUSIPCO2019, 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).
- $$\boldsymbol{y}_t = \underbrace{\Big(\sum_{r=1}^{M} x_{r,t}\boldsymbol{F}_r\Big)}_{\boldsymbol{F}(\boldsymbol{x}_t)} \underbrace{\Big(\bigotimes_{j=1}^{N} \boldsymbol{\Phi}_{j,t}\Big)}_{\mathbf{f}(\boldsymbol{\Phi}_t)} + \boldsymbol{w}_t.\ \mathbf{J}(\boldsymbol{\Phi}_t) = [\mathbf{J}(\boldsymbol{\Phi}_{1,t}) ..... \mathbf{J}(\boldsymbol{\Phi}_{N,t})]$$
- $$\text{where, } \mathbf{J}(\boldsymbol{\Phi}_{j,t}) = \boldsymbol{F}(\boldsymbol{x}_t)\big(\boldsymbol{\Phi}_{1,t} \otimes ... \mathbf{I}_{j_P j} .... \otimes \boldsymbol{\Phi}_{N,t}\big),$$

$$FIM = \begin{bmatrix} \mathrm{E}(\gamma)\mathbf{J}(\boldsymbol{\Phi}_t)^T\mathbf{J}(\boldsymbol{\Phi}_t) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{E}(\gamma)\mathbf{J}(\boldsymbol{x}_t)^T\mathbf{J}(\boldsymbol{x}_t) + \mathrm{E}(\boldsymbol{\Gamma}^{-1}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & a\mathrm{E}(\boldsymbol{\Gamma}^{-2}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (N+c-1)\mathrm{E}(\gamma^{-2}) \end{bmatrix}$$

Outline

BP-MF-EP Outperforms SAVED-KS DL



Static SBL: NMSE as a function of N.

- ALS- Alternating Least Squares.
- Exponential power delay profile for $\boldsymbol{x}_t$.
- 30 non zero elements in $\boldsymbol{x}_t$, same support across all time.
- Dimensions: 3-D Tensor $(4, 8, 8)$, with $M = 200$.

Conclusions and Thank You!

- Further advancements from [ThomasSlock:ICASSP19]: VB with a too fine variable partitioning is quite suboptimal.
- Better approximation is message passing based methods such as belief propagation (BP) and expectation propagation (EP)
- BP or EP message passing can be implemented using low complexity methods such as AMP/GAMP/VAMP, which are proven to be Bayes optimal under certain conditions on $\mathbf{A}$.
- AMP - Approximate message passing. We also derived an Generalized Vector AMP (GVAMP-SBL) SBL version to take care of diagonal power delay profile.
- Further work to be done on learning a combination of structured and unstructured Kronecker factor matrices.