SPACE ALTERNATING VARIATIONAL ESTIMATION AND KRONECKER STRUCTURED DICTIONARY LEARNING

Christo Kurisummoottil Thomas, Dirk Slock

kurisumm@eurecom.fr, slock@eurecom.fr

Motivation

Why SAVED-KS ?

- VB: analytical approximations to the posterior distributions of interest even when exact inference of these distributions is intractable.
- Propose a novel fast algorithm called space alternating variational estimation with Kronecker structured dictionary learning (SAVED-KS), which is a version of VB(-SBL) pushed to the scalar level.
- The component-wise approach of SAVE compared to SBL renders it less likely to get stuck in bad local optima and its inherent damping (more cautious progression) also leads to typically faster convergence of the non-convex optimization process
- Unstructured KS dictionary matrices learning

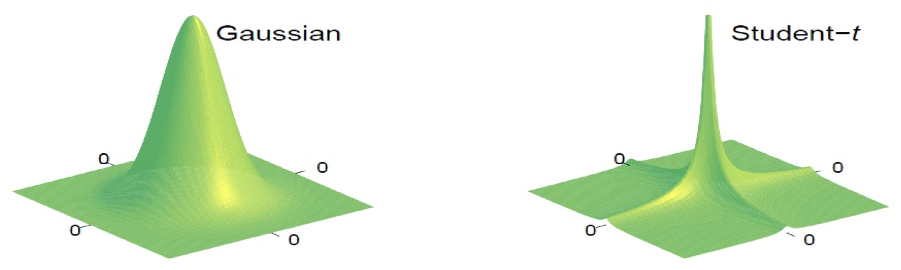
Sparse Bayesian Learning

- Bayesian Compressed Sensing: 2-layer hierarchical prior for $\mathbf{x}$ as in [Tipping:JMLR01, WipfRao:TSP04], inducing sparsity for $\mathbf{x}$.

$p(x_i|\alpha_i) = \mathcal{N}(0, \alpha_i^{-1})$, $p(\alpha_i/a, b) = \Gamma^{-1}(a) b^a \alpha_i^{a-1} e^{-b\alpha_i}$

$\Rightarrow$ sparsifying Student-t marginal

$$p(x_i) = \frac{b^a \Gamma(a+\frac{1}{2})}{(2\pi)^{\frac{1}{2}}\Gamma(a)} (b + x_i^2/2)^{-(a+\frac{1}{2})}$$



- SBL which is geared towards compressed sensing assuming sparse unknown vectors. It's just that the SBL approach works well in the case of relatively limited date (for a given state-space dimension) in which case estimation emphasis is given to large unknowns and small unknowns get more or less ignored.
- We apply the (Gamma) prior to the precision of the state x, allowing to sparsify the components of x.

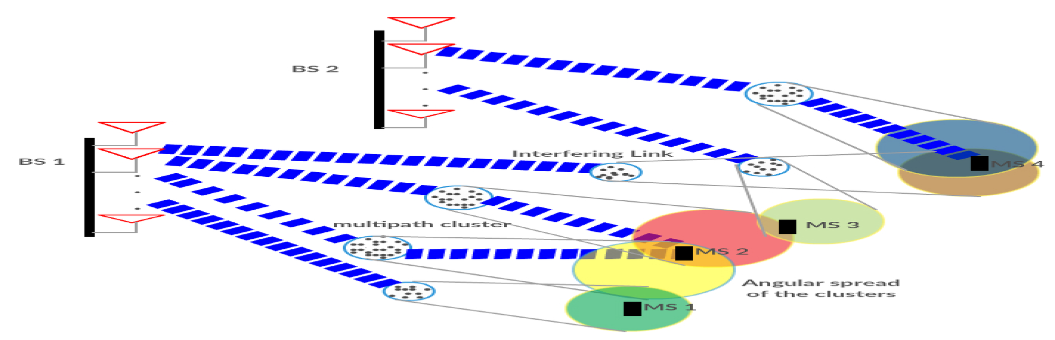
Application: Massive MIMO Channel Estimation

We get for the matrix impulse response of a time-varying frequency-selective MIMO channel $\mathbf{H}(t, \tau)$,

$$\mathbf{H}(t,\tau) = \sum_{i=1}^{N_p} A_i(t)\, e^{j2\pi f_i t}\, \mathbf{h}_r(\phi_i)\, \mathbf{h}_t^T(\psi_i)\, p(\tau - \tau_i) \;.$$

with N_p (specular) pathwise contributions where

Massive MIMO Channel Estimation



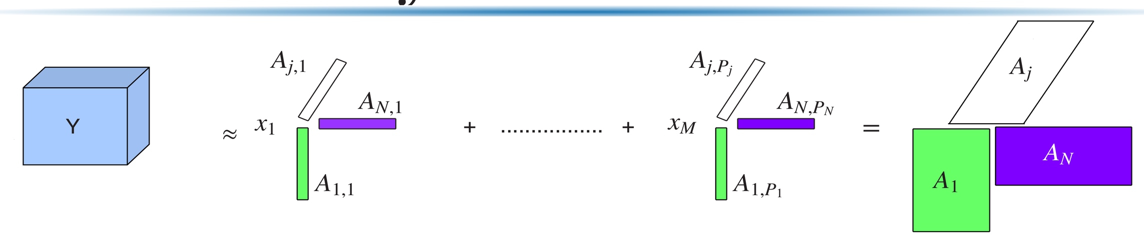
- A_i: complex attenuation, f_i: Doppler shift
- ψ_i: AoD (azimuth, elevation, polar), ϕ_i: AoA (azimuth, elevation, polarization)
- τ_i: path delay (ToA),$\mathbf{h}_t(.)$, $\mathbf{h}_r(.)$: $N_t/N_r \times 1$ Tx/Rx antenna array response,$p(.)$: pulse shape (Tx filter)

The channel impulse response $\mathbf{H}$ has per path a rank one contribution in four dimensions (Tx and Rx spatial multi-antenna dimensions, delay spread and Doppler spread). Hence, going to the frequency domain, we get

$$\mathrm{vec}(\mathbf{H}(1:t, f_1:f_2)) = \sum_{i=1}^{N_p} A_i\, \mathbf{h}_t(\psi_i) \otimes \mathbf{h}_r(\phi_i) \otimes \mathbf{v}_f(\tau_i) \otimes \mathbf{v}_t(f_i).$$

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. $\mathbf{h}_r$, $\mathbf{h}_t$: Kronecker structure in the case of polarization or in the case of $2D$ antenna arrays with separable structure [Sidiropoulos:icassp18].

System Model



Let $Y_{i_1,\ldots,i_N}$ represents the $i_1 i_2 \ldots i_N^{th}$ element of the tensor and $\mathbf{y} = [y_{1,1,\ldots,1}, y_{1,1,\ldots,2}, \ldots, y_{I_1,I_2,\ldots,I_N}]^T$, then it can be verified that [Sidiropoulos:TSP17],

$\mathbf{y} = (\mathbf{A}_1 \otimes \mathbf{A}_2 \ldots \otimes \mathbf{A}_N)\mathbf{x} + \mathbf{w}, \mathbf{w} \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$,

Matrix Unfolding:$\mathbf{Y}^{(n)} = \mathbf{A}_n \mathbf{X}^{(n)} (\mathbf{A}_N \otimes \ldots \mathbf{A}_{n+1} \otimes \mathbf{A}_{n-1} \ldots \otimes \mathbf{A}_1)^T$,

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

Variational Bayesian Inference

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

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

- Equivalent to maximizing the evidence lower bound (ELBO) [Tzikas:SPMag08], $\boldsymbol{\theta} = \{\mathbf{x}, \boldsymbol{\alpha}, \gamma, \mathbf{A}\}$.

$\ln p(\mathbf{y}) = L(q) + KLD_{VB}$, where,

$$L(q) = \int q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}, KLD_{VB} = -\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}/\mathbf{y})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$$

$$\ln(q_i(\theta_i)) = <\ln p(\mathbf{y}, \boldsymbol{\theta})>_{k \neq i} + c_i,$$

SAVED-KS Equations

Joint Distribution:

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

Gaussian q for x_i or for $\mathbf{a}_{ji}$ (Multivariate):

$$\sigma_i^2 = \frac{1}{<\gamma> \prod_{j=1}^{N} <\|\mathbf{A}_{j,p_{ji}}\|^2> + <\alpha_i>}, \quad \mathbf{C}_i = (\bigotimes_{j=1}^{N} \mathbf{A}_{j,p_{ji}}),$$

$$\widehat{x}_i = \sigma_i^2 ((<\mathbf{C}_i^H> \mathbf{y} - <(\mathbf{C}_i^H \mathbf{C}_{\bar{i}})><\mathbf{x}_{\bar{i}}>) <\gamma> .$$

$$\widehat{\mathbf{a}}_{ji} = (\mathbf{b}_j)_{\bar{1}}, \; \mathbf{b}_j = (\mathbf{Y}^{(j)} <\mathbf{X}^{(j)}> <(\bigotimes_{k=N, k\neq j}^{1} \mathbf{A}_k)^T>)_i,$$

$$\boldsymbol{\Upsilon}_{j,i} = \beta_{j,i}\mathbf{I}, \; \beta_{j,i} = \mathrm{tr}\{(\bigotimes_{k=N, k\neq j}^{1} <\mathbf{A}_k^T \mathbf{A}_k^*>) <\mathbf{X}^{(j)H} \mathbf{X}^{(j)}>\},$$

Gamma q for hyper-parameters

$$<\alpha_i> = \frac{a + \frac{1}{2}}{(<|x_i|^2> + b)}, \quad \text{where} \quad <|x_i|^2> = |\widehat{x}_i|^2 + \sigma_i^2.$$

$$<\gamma> = \frac{c + \frac{N}{2}}{(<\left\|\mathbf{y} - (\bigotimes_{j=1}^{N} \mathbf{A}_j)\mathbf{x}\right\|^2> + d)}.$$

Joint VB for KS Matrices

Complex Matrix Normal Distribution

$\mathbf{M}_j = \widehat{\mathbf{A}}_j = <\gamma> \mathbf{B}_j \boldsymbol{\Psi}_j$,

$\boldsymbol{\Psi}_j = (<\gamma> \mathbf{X} \bigotimes_{k=1, k\neq j}^{N} <\mathbf{A}_k^T \mathbf{A}_k^*> \mathbf{X}^H)^{-1}, \mathbf{X} = \mathrm{diag}(\mathbf{x})$

$\mathbf{B}_j$ is with the first row of $(\mathbf{Y}^{(j)} <(\bigotimes_{k=1, k\neq j}^{N} \mathbf{A}_k)^*> <\mathbf{X}^H>)$ removed.

SAVED-KS SBL Algorithm

Given: $\mathbf{y}, \mathbf{A}, M, N$.

Initialization: a, b, c, d are taken to be very low, on the order of 10^{-10}. $\alpha_i^0 = a/b, \forall i, \gamma^0 = c/d$ and $\sigma_i^{2,0} = \frac{1}{\|\mathbf{A}_i\|^2 \gamma^0 + \alpha_i^0}$, $\mathbf{x}^0 = \mathbf{0}$.

At iteration $t+1$,

- Update $\sigma_i^{2,t+1}$, $\widehat{x}_i^{t+1}, \forall i$ from using $\mathbf{x}_{i-}^{t+1}$ and $\mathbf{x}_{i+}^t$.
- Update $\widehat{\mathbf{A}}_{j,i}, \forall i, j$ or $\mathbf{A}_j, \forall j$.
- Compute $<x_i^{2,t+1}>$ and update α_i^t.
- Update the noise variance, γ^{t+1}.
- Continue steps $1-4$ till convergence of the algorithm.

Comparison to SotA

- **Lowering Complexity**: No matrix inversions compared to standard SBL and ALS.
- **Improving Convergence** compared to standard ALS.

Identifiability

- The local identifiability (upto permutation ambiguity) of the KS DL is ensured if the FIM is non-singular.
- $\mathbf{J}(\boldsymbol{\theta}, \mathbf{x}) = [\mathbf{J}(\boldsymbol{\theta})\; \mathbf{J}(\mathbf{x})]$, $\mathbf{J}(\boldsymbol{\theta}) = [\mathbf{J}(\boldsymbol{\theta}_1) \ldots\ldots \mathbf{J}(\boldsymbol{\theta}_N)]$ where, $\mathbf{J}(\boldsymbol{\theta}_j) = \mathbf{F}(\mathbf{x})(\boldsymbol{\theta}_1 \otimes \ldots \mathbf{I}_{I_j P_j} \ldots \otimes \boldsymbol{\theta}_N)$,

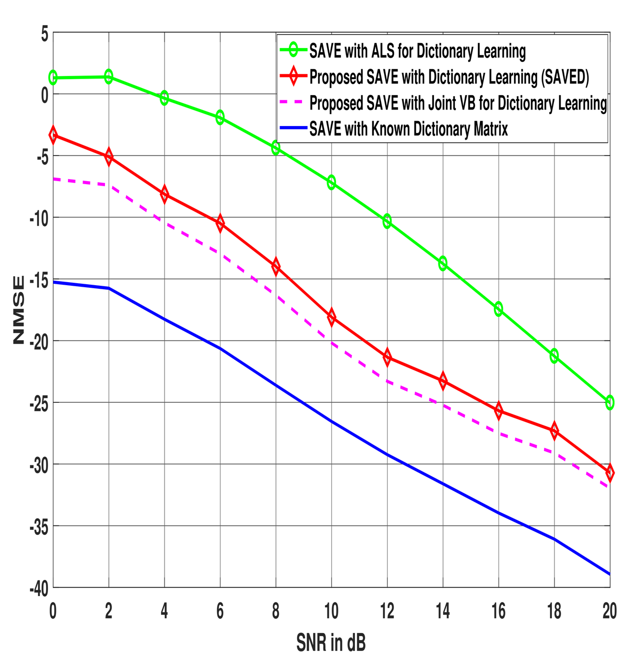
$$\mathbf{J}(\mathbf{x}) = [\mathbf{F}_1(\bigotimes_{j=1}^{N} \boldsymbol{\theta}_j), \ldots, \mathbf{F}_M(\bigotimes_{j=1}^{N} \boldsymbol{\theta}_j))].$$

- $FIM =$

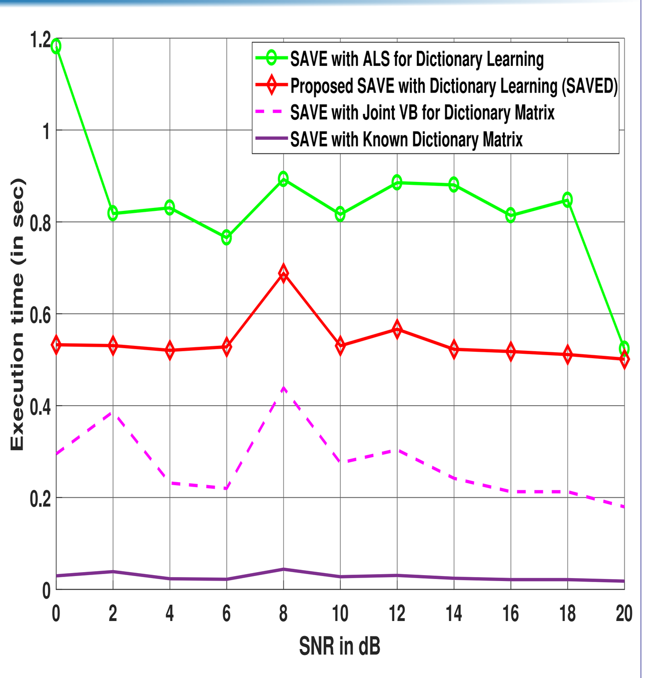
$$\begin{bmatrix} \mathrm{E}(\gamma)\mathbf{J}(\boldsymbol{\theta})^H \mathbf{J}(\boldsymbol{\theta}) & 0 & 0 & 0 \\ 0 & \mathrm{E}(\gamma)\mathbf{J}(\mathbf{x})^H \mathbf{J}(\mathbf{x}) + \mathrm{E}(\boldsymbol{\Gamma}) & 0 & 0 \\ 0 & 0 & a\,\mathrm{E}(\boldsymbol{\Gamma}^{-2}) & 0 \\ 0 & 0 & 0 & N'\,\mathrm{E}(\gamma^{-2}) \end{bmatrix}$$

- For the FIM analysis (with known support of $\mathbf{x}$), then $\mathrm{E}(\gamma)\mathbf{J}(\mathbf{x})^H \mathbf{J}(\mathbf{x}) + \mathrm{E}(\boldsymbol{\Gamma})$ and $a\,\mathrm{E}(\boldsymbol{\Gamma}^{-2})$ becomes invertible if $\prod_{j=1}^{N} I_j > K$. Assuming $\prod_{j=1}^{N} I_j > \sum_{j=1}^{N} (I_j - 1)P_j$, i.e. no. of degrees of freedom in the dictionary $< \prod_{j=1}^{N} I_j$, FIM is non-singular.
- Possibility of FIM singularity even under single measurement vector case.
- Mixture of $P (< N)$ Vandermonde matrix factors and non-parametric KS factors: The identifiability conditions can be restated as, $\prod_{j=1}^{N} I_j > \sum_{j=1}^{P} P_j + \sum_{j=P+1}^{N} (I_j - 1)P_j$.

Numerical Results



$I_1 = 4, I_2 = 10, I_3 = 4$.



Convergence behaviour (20 dB).

Conclusion and Future Work

- Convex combination of structured and unstructured KS factor matrices, For eg, DoA response closeness to the vandermonde.
- Asymptotic performance analysis, mismatched CRBs.