

# Nonlinear MMSE using Linear MMSE Bricks and Application to Compressed Sensing and Adaptive Kalman Filtering

Dirk Slock

Communication Systems Department, EURECOM, Sophia Antipolis, France



ICASSP 2020, May 4-8,  
ETON Talk

• Thanks to



Anthony Schutz



Bensaid Siouar



Christo Thomas

• This talk is about more than the title = teaser.

# Main Messages

- There are many **Bayesian estimation** problems, many of which are LMMSE (Wiener, Kalman), which contain **hyperparameters** to be tuned, using various approaches.
- Information combining: from **weighted least-squares** to **message passing** in a more general overall Bayesian formulation (e.g. cooperative location estimation)
- **Empirical Bayes (EB)** as principled framework for **bias-variance trade-off**
- but not necessarily using empirical Bayes for hyperparameter estimation: **SURE, Cross Validation**
- **compressive sensing, sparse models, generalization of model order selection to support region, model complexity and structure**
- **Sparse Bayesian Learning (SBL)** is one EB instance, allowing to exploit (approximate) sparsity for compressed sensing
  - can be extended to time-varying scenarios with sparse variations also
  - can be extended to dictionary learning, in particular with Kronecker structured dictionaries
- **message passing** (approximate iterative) inference techniques: **easy to get the mean (estimate) correct** but **more difficult to get correct posterior variances**

## Main Messages (2)

- free energy optimization framework, guided by **mismatched Cramer-Rao Bound (mCRB)** for split in various MP simplification levels (**Belief Propagation (BP)**, **Variational Bayes (VB) - Mean Field (MF)**), allowing performance-complexity trade-off
- **large system analysis (LSA)** yields simplified asymptotic performance analysis for certain measurement matrix models, allowing to show optimality and to justify algorithmic simplifications
- **Approximate Message Passing (AMP)** very similar to approximate large **turbo receivers for CDMA** for which heuristic LSA was performed based on **Replica Method Analysis**.
- AMP can be derived more rigorously from **BP**, using asymptotically justifiable **first-order Taylor series expansions** and **Gaussian approximations**.
- LSA allows tracking of the AMP MSE through the iterations, called **State Evolution (SE)**, showing convergence to MMSE and hence optimality.
- SE requires **statistical models for the measurement matrix  $\mathbf{A}$** . Pushing these model assumptions completely through to the xAMP algorithms may be an unnecessary simplification. The main requirement is **independent rows/columns** as in CDMA random spreading.
- Most xAMP versions require i.i.d.  $x$ , which is not suited for SBL. We present **new LSA for SBL-AMP**.

# Outline

- 1 Introduction
- 2 Static SBL
- 3 Combined BP-MF-EP Framework
- 4 Posterior Variance Prediction: Bayes Optimality
- 5 Performance Analysis of Approximate Inference Techniques (mCRB)
- 6 Dynamic SBL
- 7 Kronecker Structured Dictionary Learning using BP/VB
- 8 Numerical Results and Conclusion

# Kalman Filter

Linear state-space model:

state update equation:

$$\mathbf{x}_{k+1} = \mathbf{F}_k(\theta) \mathbf{x}_k + \mathbf{G}_k(\theta) \mathbf{w}_k$$

measurement equation:

$$\mathbf{y}_k = \mathbf{H}_k(\theta) \mathbf{x}_k + \mathbf{v}_k$$

for  $k = 1, 2, \dots$ , with uncorrelated

- initial state  $\mathbf{x}_0 \sim \mathcal{N}(\hat{\mathbf{x}}_0, \mathbf{P}_0)$ ,
- measurement noise  $\mathbf{v}_k \sim \mathcal{N}(0, \mathbf{R}_k(\theta))$ ,
- state noise  $\mathbf{w}_k \sim \mathcal{N}(0, \mathbf{Q}_k(\theta))$ .

State model known up to some parameters  $\theta$ .

Often  $\mathbf{F}_k(\theta)$ ,  $\mathbf{G}_k(\theta)$ ,  $\mathbf{H}_k(\theta)$  linear in  $\theta$ : bilinear case.

# Numerous Applications

- LMMSE wireless channel estimation:

$x_k$  = FIR filter response,  $\theta$ : Power Delay Profile, AR(1) dynamics in e.g. diagonal  $F$  and  $Q$

- Bayesian adaptive filtering (BAF):

analogous to LMMSE channel estimation, except measurement equation: only one 1D measurement is available per instance. An extremely simplified form of BAF is the so-called Proportionate LMS (P-LMS) algorithm.

- Position tracking (GPS):

$$\mathbf{x}_{t+1} = \begin{bmatrix} 1 & \Delta t & \frac{1}{2}\Delta t^2 \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{bmatrix} \cdot \mathbf{x}_t = \begin{bmatrix} x_t + \Delta t \cdot v_t + \frac{1}{2}\Delta t^2 a \\ v_t + \Delta t \cdot a \\ a \end{bmatrix}$$

$\theta$ : acceleration model parameters (e.g. white noise, AR(1))

- Blind Audio Source Separation (BASS):

$x_k$  = source signals,

$\theta$ : (short+long term) AR parameters, reverb filters

# Static LMMSE (Wiener) Applications

- **Direction of Arrival (DoA) estimation:**  $\mathbf{x}$  = decorrelated sources, apart from the DoA parameters there could also be antenna array calibration parameters or source and noise covariance parameters.
- **Blind and semi-blind channel estimation.** In the techniques that exploit the (white) second-order statistics of  $\mathbf{x}$ , (the unknown part of)  $\mathbf{x}$  gets modeled as Gaussian. Numerous variations: single-carrier, OFDM and CDMA versions, single- and multi-user, single- and multi-stream, with black box FIR channel models or propagation based parameterized channel models.  
Image Deblurring, NMRI Imaging
- **LMMSE receiver (Rx) design:**  $\mathbf{x}$  = Tx symbol sequence to be recovered on the basis of Rx signal, in single- or multi-user settings and other variations as in the channel estimation case. The crosscorrelation between Tx and Rx signals depends on the channel response, which is part of the parameters. The Rx signal covariance matrix on the other hand can be modeled in various ways, non-parametric or parametric. Account for the channel estimation error in the LMMSE Rx design. Another approach: consider the LMMSE filter directly as the parameters.  
LMMSE Tx design, partial CSIR/CSIT.



# Adaptive Kalman Filter solutions

- Extended Kalman Filter (**EKF**)
- other generic nonlinear Kalman Filter extensions:  
Unscented Kalman Filter (**UKF**), Cubature Kalman Filter (**CKF**), Gaussian Sum Filter, Particle Filter (**PF**)
- Recursive Prediction Error Method (**RPEM**) Kalman Filter
- Second-Order Extended Kalman Filter (**SOEKF**)
- Expectation-Maximization (**EM**)/Variational Bayes (**VB**) Kalman Filter

# Time Varying Sparse State Tracking

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

The diagram illustrates the AR(1) process for sparse state tracking. It consists of two main equations:

Top equation:  $\mathbf{y}_t = \mathbf{A}_t \mathbf{x}_t + \mathbf{v}_t$

- $\mathbf{y}_t$  is an  $N \times 1$  vector.
- $\mathbf{A}_t$  is an  $N \times M$  matrix, where  $N \ll M$ .
- $\mathbf{x}_t$  is an  $M \times 1$  vector.
- $\mathbf{v}_t$  is an  $N \times 1$  vector.

Bottom equation:  $\mathbf{x}_t = \mathbf{F} \mathbf{x}_{t-1} + \mathbf{w}_t$

- $\mathbf{x}_t$  is an  $M \times 1$  vector.
- $\mathbf{F}$  is an  $M \times M$  matrix.
- $\mathbf{x}_{t-1}$  is an  $M \times 1$  vector.
- $\mathbf{w}_t$  is an  $M \times 1$  vector.

Define:  $\Xi = \text{diag}(\xi)$ ,  $\mathbf{F} = \text{diag}(f)$ .

$f_i$ : correlation coefficient and  $x_{i,t} \sim \mathcal{CN}(x_{i,t}; 0, \frac{1}{\xi_i})$ . Further,  $\mathbf{w}_t \sim \mathcal{CN}(\mathbf{w}_t; \mathbf{0}, \Gamma^{-1} = \Xi^{-1}(\mathbf{I} - \mathbf{F}\mathbf{F}^H))$

and  $\mathbf{v}_t \sim \mathcal{CN}(\mathbf{v}_t; \mathbf{0}, \gamma^{-1}\mathbf{I})$ . VB leads to Gaussian SAVE-Kalman Filtering (GS-KF).

Applications: Localization, Adaptive Filtering.

# Compressed Sensing Problem

- **Noiseless case:** Given underdetermined  $\mathbf{y}$ ,  $\mathbf{A}$ , the optimization problem is

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}.$$

Can recover  $\mathbf{x}$  and its support for small  $N - \|\mathbf{x}\|_0$   
(small overdetermination if support were known)

- **Noisy case:**

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ subject to } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon.$$

- $l_0$  norm minimization: an NP-hard problem.
- Constrained problem  $\Rightarrow$  **Lagrangian**, **Convex Relaxation** (using  $p$  norm,  $p > 1$ ):

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_p.$$

**Restricted Isometry Property (RIP):**  $\mathbf{A}^T \mathbf{A}$  sufficiently diagonally dominant

- Most identifiability work considered **noiseless** data & **exact sparsity**

# Sparse Signal Recovery Algorithms

## Convex Relaxation based Methods:

- Basis pursuit ( $l_1$  norm)<sup>1</sup>.
- LASSO( $l_1$  norm)<sup>2</sup>
- Dantzig selector<sup>3</sup>
- FOCUSS ( $l_p$  norm, with  $p < 1$ ).

## Greedy Algorithms:

- Matching Pursuit<sup>4</sup>
- Orthogonal Matching Pursuit (OMP)<sup>5</sup>
- CoSaMP<sup>6</sup>

## Iterative Methods:

- Iterative Shrinkage and Thresholding Algorithm (ISTA)<sup>7</sup> or Fast ISTA.
- Approximate Message Passing variants (xAMP)- more details to follow.
- Very recent: **Deep learning based methods** such as Learned ISTA (LISTA)<sup>8</sup>, Learned AMP (LAMP) and Learned Vector AMP (LVAMP)<sup>9</sup>.

<sup>1</sup>Chen, Donoho, Saunders'99, <sup>2</sup>Tibshirani'96, <sup>3</sup>Candes, Tao'07

<sup>4</sup>Mallat, Zhang'93, <sup>5</sup>Tropp, Gilbert'07, <sup>6</sup>Needell, Tropp'09

<sup>7</sup>Daubechies, Defrise, Mol'04, <sup>8</sup>Gregor, Cun'10, <sup>9</sup>Borgerding, Schniter, Rangan'17

# James-Stein Estimator

- Bayesian interpretation of (possibly overdetermined) Compressed Sensing:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 - 2\sigma_v^2 \ln p(\mathbf{x})$$

- Stein and James<sup>10</sup> showed that for i.i.d. linear Gaussian model  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \xi^{-1}\mathbf{I})$ , it is possible to construct a nonlinear estimate of  $\mathbf{x}$  with lower MSE than that of ML for all values of the true unknown  $\mathbf{x}$ .
- A popular design strategy: is to minimize **Stein's unbiased risk estimate (SURE)**, which is an unbiased estimate of the MSE.
- SURE directly approximates the MSE of an estimate from the data, without requiring knowledge of the hyperparameters ( $\xi$ ), it is an instance of **empirical Bayes**.
- Stein's landmark discovery lead to the study of **biased estimators that outperform minimum variance unbiased estimators (MVU)** in terms of MSE, e.g. work by Yonina Eldar<sup>11</sup>.
- Shrinkage estimators** and **penalized maximum likelihood (PML)** estimators.

---

<sup>10</sup>James, Stein'61

<sup>11</sup>Eldar'08

# Kernel Methods in Automatic Control

- Kernel methods in **linear system identification**<sup>12</sup> ( $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}$ ,  $\mathbf{v} \sim \mathcal{N}(\mathbf{v}; \mathbf{0}, \gamma^{-1}\mathbf{I})$ ).
- Traditional methods: maximum likelihood (ML) or prediction error methods (PEM)
- ML/PEM optimal in the large data limit.
- Questions: Model structure design for ML/PEM. **Achieving a good bias-variance trade off.**
- Solution: Parameterized Kernel design and hyperparameter estimation. **Methods for hyperparameter estimation include cross-validation (CV), empirical Bayes (EB),  $C_p$  statistics and Stein's unbiased risk estimate (SURE).**
- Regularized least square estimator ( $\mathbf{P}$  is symmetric and +ve semidefinite kernel matrix):

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathcal{R}^M} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + \frac{1}{\gamma} \mathbf{x}^T \mathbf{P}^{-1} \mathbf{x}.$$

- Parameterized family of matrices,  $\mathbf{P}(\boldsymbol{\eta})$ , where  $\boldsymbol{\eta} \in \mathcal{R}^p$ .  $\boldsymbol{\eta}$  are the hyperparameters.
- Can be interpreted as a **constrained form of SBL**, with a zero-mean Gaussian prior for  $\mathbf{x}$  of which the covariance matrix is a linear combination of some fixed matrices (SBL being a special case with fixed matrices  $\mathbf{e}_i \mathbf{e}_i^T$ ).
- A good overview of Kernel methods, connection with machine learning<sup>13</sup>.

<sup>12</sup>Pillonetto, Nicolao'10

<sup>13</sup>Pillonetto, Dinuzzo, Chen, Nicolao, Ljung'14

# Kernel Hyperparameter Estimation

- Empirical Bayes (EB=Type II ML):

$$\hat{\eta}_{EB} = \arg \min_{\eta} f_{EB}(\mathbf{P}(\eta)),$$

$$f_{EB}(\mathbf{P}(\eta)) = \mathbf{y}^T \mathbf{Q}^{-1} \mathbf{y} + \ln \det(\mathbf{Q}) \text{ with } \mathbf{Q} = \mathbf{A} \mathbf{P} \mathbf{A}^T + \frac{1}{\gamma} \mathbf{I}_N.$$

- Two SURE methods:
- SURE 1:** MSE of signal reconstruction ( $MSE_x(\mathbf{P}) = E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2)$ ):

$$SURE_x : \hat{\eta}_{Sx} = \arg \min_{\eta} f_{Sx}(\mathbf{P}(\eta)), \text{ with}$$

$$f_{Sx}(\mathbf{P}(\eta)) = \frac{1}{\gamma^2} \mathbf{y}^T \mathbf{Q}^{-T} \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-2} \mathbf{A}^T \mathbf{Q}^{-1} \mathbf{y} + \frac{1}{\gamma} \text{tr}\{2\mathbf{R}^{-1} - (\mathbf{A}^T \mathbf{A})^{-1}\},$$

$$\mathbf{R} = \mathbf{A}^T \mathbf{A} + \frac{1}{\gamma} \mathbf{P}^{-1}.$$

- SURE 2:** MSE of output prediction ( $MSE_y(\mathbf{P}) = E(\|\mathbf{A}\hat{\mathbf{x}} + \mathbf{v}^* - \mathbf{y}\|^2)$ ),  $\mathbf{v}^*$  independent from  $\mathbf{v}$ :

$$SURE_y : \hat{\eta}_{Sy} = \arg \min_{\eta} f_{Sy}(\mathbf{P}(\eta)), \text{ with}$$

$$f_{Sy}(\mathbf{P}(\eta)) = \frac{1}{\gamma^2} \mathbf{y}^T \mathbf{Q}^{-T} \mathbf{Q}^{-1} \mathbf{y} + 2 \frac{1}{\gamma} \text{tr}\{\mathbf{A} \mathbf{P} \mathbf{A}^T \mathbf{Q}^{-1}\}$$

# Asymptotic Properties of Hyperparameter Estimators

- Derived **first order optimality conditions**. In general, no closed form expression shown except for special cases for e.g diagonal  $\mathbf{A}$ , ridge regression with  $\mathbf{A}^T \mathbf{A} = N \mathbf{I}_M$ .

## Theorem 1

<sup>14</sup> Assume that  $\mathbf{P}(\boldsymbol{\eta})$  is any +ve definite parameterization of the kernel matrix and  $\frac{\mathbf{A}^T \mathbf{A}}{N} \xrightarrow{N \rightarrow \infty} \boldsymbol{\Sigma}$ , where  $\boldsymbol{\Sigma}$  is positive definite. Then we have the following almost sure convergence.

$$\begin{aligned} \hat{\boldsymbol{\eta}}_{\text{MSE}_x} &\rightarrow \boldsymbol{\eta}_x^*, & \hat{\boldsymbol{\eta}}_{S_x} &\rightarrow \boldsymbol{\eta}_x^* \\ \hat{\boldsymbol{\eta}}_{\text{MSE}_y} &\rightarrow \boldsymbol{\eta}_y^*, & \hat{\boldsymbol{\eta}}_{S_y} &\rightarrow \boldsymbol{\eta}_y^* \\ \hat{\boldsymbol{\eta}}_{\text{EEB}} &\rightarrow \boldsymbol{\eta}_{\text{EB}}^*, & \hat{\boldsymbol{\eta}}_{\text{EB}} &\rightarrow \boldsymbol{\eta}_{\text{EB}}^* \end{aligned}$$

- $\hat{\boldsymbol{\eta}}_{\text{MSE}_x}, \hat{\boldsymbol{\eta}}_{\text{MSE}_y}, \hat{\boldsymbol{\eta}}_{\text{EEB}}$  being the ORACLE estimators which are optimal for any data length  $N$ .
- The two **SURE estimators converge to the best possible hyperparameter in terms of MSE in the asymptotic limit**, “asymptotically consistent”.
- EB estimator converges to another best hyperparameter minimizing the expectation of the EB estimation criterion (**EEB**).
- Convergence of EB is faster** than that of the two SURE estimators.

<sup>14</sup>Mu, Chen, Ljung'18



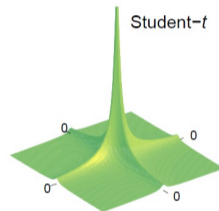
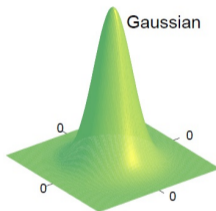
# Sparse Bayesian Learning - SBL

- Bayesian Compressed Sensing: 2-layer hierarchical prior for  $\mathbf{x}$  as in <sup>15</sup>, inducing sparsity for  $\mathbf{x}$  (conjugate priors: posterior pdf of same family as prior pdf) :

$$p_{\mathbf{x}}(x_{i,t}|\xi_i) = \mathcal{N}(x_{i,t}; 0, \xi_i^{-1}), \quad p(\xi_i|a, b) = \Gamma^{-1}(a)b^a \xi_i^{a-1} e^{-b\xi_i}$$

⇒ sparsifying Student-t marginal

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



- Sparsification of the Innovation Sequence: we apply the (Gamma) prior not to the precision of the state  $\mathbf{x}$  but of it's innovation  $\mathbf{w}$ , allowing to sparsify at the same time the components of  $\mathbf{x}$  AND their variation in time (innovation).
- The inverse of the noise variance  $\gamma$  is also assumed to have a Gamma prior,  $p_{\gamma}(\gamma|c, d) = \Gamma^{-1}(c)d^c \gamma^{c-1} e^{-d\gamma}$ .

<sup>15</sup>Tipping'01

# Outline

- 1 Introduction
- 2 Static SBL**
- 3 Combined BP-MF-EP Framework
- 4 Posterior Variance Prediction: Bayes Optimality
- 5 Performance Analysis of Approximate Inference Techniques (mCRB)
- 6 Dynamic SBL
- 7 Kronecker Structured Dictionary Learning using BP/VB
- 8 Numerical Results and Conclusion

## Original SBL Algorithm (Type II ML)

- Original SBL<sup>16</sup>, for a fixed estimate of the hyperparameters  $(\hat{\xi}, \hat{\gamma})$ , the posterior of  $\mathbf{x}$  is Gaussian, i.e.

$$p_{\mathbf{x}}(\mathbf{x}|\mathbf{y}, \hat{\xi}, \hat{\gamma}) = \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \Sigma_L)$$

leading to the (Linear) MMSE estimate for  $\mathbf{x}$

$$\begin{aligned} \hat{\mathbf{x}} &= \hat{\gamma}(\hat{\gamma}\mathbf{A}^T\mathbf{A} + \hat{\Xi})^{-1}\mathbf{A}^T\mathbf{y}, \\ \Sigma_L &= (\hat{\gamma}\mathbf{A}^T\mathbf{A} + \hat{\Xi})^{-1}. \end{aligned} \quad (1)$$

- The hyperparameters are estimated from the likelihood function by marginalizing over the sparse coefficients  $\mathbf{x}$ , the marginalized likelihood being denoted as  $p_{\mathbf{y}}(\mathbf{y}|\xi, \gamma)$ .  $\xi, \gamma$  are estimated by maximizing  $p_{\mathbf{y}}(\mathbf{y}|\xi, \gamma)$  and this procedure is called as Type-II ML. Type-II ML is solved using EM, which leads to the following updates for the hyperparameters.

$$\hat{\xi}_i = \frac{a + \frac{1}{2}}{\left(\frac{\langle x_i^2 \rangle}{2} + b\right)}, \quad \text{where } \langle x_i^2 \rangle = \hat{x}_i^2 + \sigma_i^2. \quad \langle \gamma \rangle = \frac{c + \frac{N}{2}}{\left(\frac{\langle \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 \rangle}{2} + d\right)},$$

$$\begin{aligned} \text{where, } \langle \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 \rangle &= \|\mathbf{y}\|^2 - 2\mathbf{y}^T\mathbf{A}\hat{\mathbf{x}} + \text{tr}(\mathbf{A}^T\mathbf{A}(\hat{\mathbf{x}}\hat{\mathbf{x}}^T + \Sigma)), \\ \Sigma &= \text{diag}(\Sigma_L) = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_M^2), \quad \hat{\mathbf{x}} = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_M]^T. \end{aligned}$$

<sup>16</sup>Tipping'01, Wipf,Rao'04

# Type I vs Type II ML

- Type I  $\implies$  standard MAP estimation (involves integrating out the hyperparameters)

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x}} [\log p_{\mathbf{y}}(\mathbf{y}|\mathbf{x}) + p_{\mathbf{x}}(\mathbf{x})].$$

- Type II  $\implies$  hyperparameters ( $\Psi = \{\xi, \gamma\}$ ) are estimated using an evidence maximization approach

$$\hat{\Psi} = \arg \max_{\Psi} p_{\Psi}(\Psi|\mathbf{y}) = \arg \max_{\Psi} p_{\Psi}(\Psi) \int p_{\mathbf{y}}(\mathbf{y}|\Psi) = \arg \max_{\Psi} p_{\Psi}(\Psi) \int p_{\mathbf{y}}(\mathbf{y}|\mathbf{x}, \gamma) p_{\mathbf{x}}(\mathbf{x}|\xi) d\mathbf{x}.$$

- Why Type II is better than Type I? <sup>17</sup> In the evidence maximization framework instead of looking for the mode of the true posterior  $p_{\mathbf{x}}(\mathbf{x}|\mathbf{y})$ , the true posterior is approximated as  $p_{\mathbf{x}}(\mathbf{x}|\mathbf{y}; \hat{\Psi})$ , where  $\hat{\Psi}$  is obtained by maximizing the true posterior mass over the subspaces spanned by the non zero indexes.
- Type I methods seek the mode of the true posterior and use that as the point estimate of the desired coefficients. Hence, if the true posterior distribution has a skewed peak, then the Type I estimate (Mode) is not a good representative of the whole posterior.

<sup>17</sup>Giri, Rao'16

## Variational Bayesian (VB) Inference

- The computation of the posterior distribution of the parameters is usually intractable. As in SAGE, **SAVE is simply VB with partitioning of the unknowns at the scalar level**. Define  $\theta = \{x, \xi, \gamma\}$ ,  $\theta_i$  represents each scalar and  $\theta_{\bar{i}}$  denotes  $\theta$  excluding  $\theta_i$ .

$$q(\theta) = q_{\gamma}(\gamma) \prod_{i=1}^M q_{x_i}(x_i) \prod_{i=1}^M q_{\xi_i}(\xi_i).$$

- VB compute the factors  $q$  by **minimizing the Kullback-Leibler distance** between the true posterior distribution  $p(\theta|y)$  and the  $q(\theta)$ . From <sup>18</sup>,

$$KLD_{VB} = D_{KL}(q(\theta)||p(\theta|y)) = \int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} d\theta.$$

- The KL divergence minimization is equivalent to **maximizing the evidence lower bound (ELBO)**<sup>19</sup>.

$$\ln p(y) = L(q) + KLD_{VB} = -D_{KL}(q(\theta)||p(\theta, y)) + D_{KL}(q(\theta)||p(\theta|y)), \text{ where,}$$

$\ln p(y)$  is the evidence, and  $\min KLD_{VB}$  becomes equivalent to  $\max L(q)$ , the ELBO.

- We get for the element-wise VB recursions: (Expectation Maximization (EM) = special case:

$$\ln(q_i(\theta_i)) = \langle \ln p(y, \theta) \rangle_{\theta_{\bar{i}}} + c_i,$$

$\theta = \{\theta_s, \theta_d\}$ ,  
 $\theta_s$  random, hidden  
 $\theta_d$  deterministic)

<sup>18</sup>Beal'03, <sup>19</sup>Tzikas, Likas, Galatsanos'08

# Low Complexity-Space Alternating Variational Estimation (SAVE)

- Mean Field (MF) approximation: VB partitioned to scalar level (MF vs VB // SAGE vs EM), results in a SBL algorithm **without any matrix inversions**.
- The resulting **alternating optimization of the posteriors for each scalar in  $\theta$**  leads to

$$\ln(q_i(\theta_i)) = \langle \ln p(\mathbf{y}, \theta) \rangle_{k \neq i} + c_i,$$

$$p(\mathbf{y}, \theta) = p_{\mathbf{y}}(\mathbf{y}|\mathbf{x}, \xi, \gamma) p_{\mathbf{x}}(\mathbf{x}|\xi) p_{\xi}(\xi) p_{\gamma}(\gamma).$$

where  $\theta = \{\mathbf{x}, \xi, \gamma\}$  and  $\theta_i$  represents each scalar in  $\theta$ .

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

- Gaussian approximate posterior for  $x_i$ :**

$$\begin{aligned} \ln q_{x_i}(x_i) = & -\frac{\langle \gamma \rangle}{2} \left\{ \langle \|\mathbf{y} - \mathbf{A}_{\bar{i}} \mathbf{x}_{\bar{i}}\|^2 \rangle - (\mathbf{y} - \mathbf{A}_{\bar{i}} \langle \mathbf{x}_{\bar{i}} \rangle)^T \mathbf{A}_i x_i - \right. \\ & \left. x_i \mathbf{A}_i^T (\mathbf{y} - \mathbf{A}_{\bar{i}} \langle \mathbf{x}_{\bar{i}} \rangle) + \|\mathbf{A}_i\|^2 x_i^2 \right\} - \frac{\langle \xi_i \rangle}{2} x_i^2 + c_{x_i} = -\frac{1}{2\sigma_i^2} (x_i - \hat{x}_i)^2 + c'_{x_i}. \end{aligned}$$

## SAVE Iterations Continued...

- The SAVE iterations for  $\mathbf{x}$  get obtained as

$$\sigma_i^2 = \frac{1}{\langle \gamma \rangle \|\mathbf{A}_i\|^2 + \xi_i}, \quad \hat{\mathbf{x}}_i = \sigma_i^2 \mathbf{A}_i^T (\mathbf{y} - \mathbf{A}_i \hat{\mathbf{x}}_i) \langle \gamma \rangle.$$

- Hyperparameter estimates: same as EM iterations. **Gamma posterior for  $\xi_i$  and  $\gamma$ .**
- No matrix inversions.
- Update of all the variables,  $\mathbf{x}$ ,  $\xi_i$ ,  $\gamma$ , requires simple addition and multiplication operations
- $\mathbf{y}^T \mathbf{A}$ ,  $\mathbf{A}^T \mathbf{A}$  and  $\|\mathbf{y}\|^2$  can be precomputed, so only need to be computed once.
- Remarks:** From LMMSE expression in (1),  $i^{\text{th}}$  row of  $\gamma \mathbf{A}^T \mathbf{A} \hat{\mathbf{x}} + \Xi \hat{\mathbf{x}} = \gamma \mathbf{A}^T \mathbf{y}$ :  
 $\gamma \mathbf{A}_i^T \mathbf{A} \hat{\mathbf{x}} + \xi_i \hat{\mathbf{x}}_i = \gamma \mathbf{A}_i^T \mathbf{y}$  or  $(\gamma \|\mathbf{A}_i\|^2 + \xi_i) \hat{\mathbf{x}}_i = \gamma \mathbf{A}_i^T (\mathbf{y} - \mathbf{A}_i \hat{\mathbf{x}}_i)$   
 Hence **SAVE does linear PIC iterations to compute the LMMSE estimate.**
- However, for the posterior variances :  $\sigma_i^2 = ((\Sigma_L^{-1})_{i,i})^{-1} \leq (\Sigma_L)_{i,i}$  with equality only for diagonal  $\Sigma_L$

# Convergence of SAVE

## Theorem 2

The convergence condition for the sparse coefficients  $x_i$  of the SAVE algorithm<sup>a</sup> can be written as  $\rho(\mathbf{D}^{-1}\mathbf{H}) < 1$ , where  $\mathbf{D} = \text{diag}(\hat{\gamma}\mathbf{A}^T\mathbf{A} + \hat{\Xi})$ ,  $\mathbf{H} = \text{offdiag}(\hat{\gamma}\mathbf{A}^T\mathbf{A})$ .  $\rho(\cdot)$  denotes the spectral radius. Moreover, if SAVE converges, assuming the estimate of hyperparameters are consistent, the posterior mean (point estimate) always converges to the exact value (LMMSE). However, the predicted posterior variance is quite suboptimal.

<sup>a</sup>Thomas,Slock'18

**Remark:** To fix the convergence of SAVE (when  $\rho(\mathbf{D}^{-1}\mathbf{H}) > 1$ ), we can use the diagonal loading method<sup>20</sup>. The modified iterations (with a diagonal loading factor matrix  $\Lambda$ ) can be written as,

$$\begin{aligned} (\mathbf{D} + \tilde{\Xi})\mathbf{x}^{(t+1)} &= -(\mathbf{H} - \tilde{\Xi})\mathbf{x}^{(t)} + \hat{\gamma}\mathbf{A}^T\mathbf{y}, \implies \\ \mathbf{x}^{(t+1)} &= -(\mathbf{D} + \tilde{\Xi})^{-1}(\mathbf{H} - \tilde{\Xi})\mathbf{x}^{(t)} + (\mathbf{D} + \tilde{\Xi})^{-1}\hat{\gamma}\mathbf{A}^T\mathbf{y}. \end{aligned}$$

The convergence condition gets modified as  $\rho((\mathbf{D} + \tilde{\Xi})^{-1}(\mathbf{H} - \tilde{\Xi})) < 1$ . Another point worth noting here is that, if the power delay profile  $\Xi$  is also estimated using MF,  $\hat{\gamma}\text{diag}(\mathbf{A}^T\mathbf{A}) + \hat{\Xi}$ , where  $\hat{\Xi} = \Xi + \tilde{\Xi}$ , with  $\tilde{\Xi} > 0$ . In this case,  $\tilde{\Xi}$  may represent an automatic correction factor (diagonal loading) to force convergence of SAVE for cases where  $\rho(\mathbf{D}^{-1}\mathbf{H}) > 1$ .

<sup>20</sup>Johnson, Bickson, Dolev'09



# NMSE Results

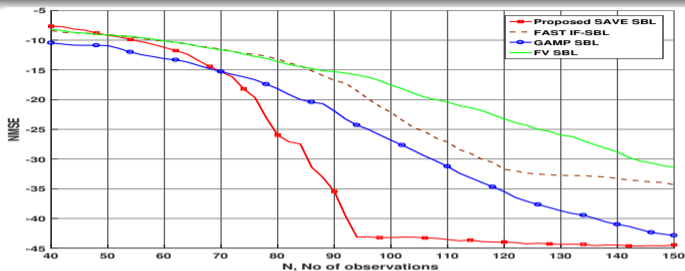


Figure 4: NMSE vs the number of observations ( $M = 200$ ,  $L = 40$ ,  $L$  is the number of non-zero elements).

- For sufficient amount of data, **SAVE has significantly lower MSE than the other fast algorithms.**
- Why? The resulting problem of alternating optimization of  $\mathbf{x}$  and the hyperparameters  $\xi$  and  $\gamma$  appears to be characterized by many local optima. Apparently, **the component-wise VB approach appears to allow to avoid a lot of bad local optima, explaining the better performance, apart from lower complexity.**
- At very low amount of data, suboptimal approaches such as AMP which don't introduce individual hyper parameters per  $\mathbf{x}$  component and assume that the  $x_i$  behave i.i.d, behave better because of the lower number of hyper parameters to be estimated.

# An Overview of Fast SBL Algorithms

- Fast SBL using Type II ML by Tipping<sup>21</sup>: greedy approach handling one  $x_i$  at a time, plus replacing precisions by their convergence values, leading to pruning of the small  $x_i$  components, i.e. explicit sparsity.
- Fast SBL using VB by Shutin et. al.<sup>22</sup>: Shutin uses VB while Tipping is Type II ML as in original SBL. They do both replace precisions by their convergence values. Shutin also added some extra view points in terms of the pruning condition being interpreted as relating between sparsity properties of SBL and a measure of SNR. Main message of the both being faster convergence compared to original SBL, not much reduction in per iteration complexity.
- BP-SBL<sup>23</sup>: In SBL, with fixed hyperparameters, MAP or MMSE estimate (follows from the Gaussian posterior) of  $\mathbf{x}$  can be efficiently computed using belief propagation (BP), since all the messages involved are Gaussian (without any approx.).
- Inverse Free SBL (IF-SBL)<sup>24</sup>: Optimization using a relaxed ELBO.
- Hyperparameter free sparse estimation<sup>25</sup>: Does not require hyperparameter tuning compared to SBL. Uses covariance matching, equivalent to square root LASSO.

<sup>21</sup>Tipping, Faul'03, <sup>22</sup>Shutin, Buchgraber, Kulkarni, Poor'11

<sup>23</sup>Tan, Li'10, <sup>24</sup>Duan, Yang, Fang, Li'17

<sup>25</sup>Zachariah, Stoica'15

# Complexity Comparisons-SBL Algorithms

| Algorithm  | Complexity per Iteration   | Convergence (No of iterations) | Sparsity  | Optimization function  | Local Optimum   |
|--|--|--------------------------------|---|--|---|
| Type I   | $O(M^3)$   |                                | Exact sparsity  | Type I ML (Depending upon the prior used, type 1 ML corresponds to LASSO/Re-weighted l1/l2 min. problems)  |   |
| Type II SBL  | $O(M^3)$   |                                | Exact sparsity ( $\alpha_i$ converges to $\infty$ )   | Type II ML solved using EM   |   |
| Fast SBL using Type II ML (Tipping, Faul'03) (Focus more on Convergence speed) | $O(L^3), L \leq M$   | $\ll L$                        | Exact sparsity (Using an entry dependent thresholding condition which follows from the computation of stationary point of $\alpha_i$ )                    | Type II ML (stationary points of $\alpha_i$ are computed to accelerate convergence)  | Convergence to a local optimum.                                 |
| Fast SBL (using VB) by Shutin (Focus more on Convergence speed)                | $O(L^3), L \leq M$   | $\ll L$                        | Exact sparsity (Using a pruning condition similar as in Tipping's)  | Maximization of ELBO in VB   | Convergence to a local optimum of ELBO (Mean field free energy) |
| Hyperparameter free SBL (Zachariah, Stoica'15)                                 | $O(M^2)$   | $\ll M$                        | The final objective function is a weighted square root LASSO. So the sum of l2 norm of (y and Ax) and weighted l1 norm of x which promotes sparsity here. | LMMSE estimator for x with Covariance matching for PDP, finally giving rise to an objective function which can be interpreted as weighted square root LASSO. | Convergence to a local optimum                                  |
| BP-SBL (Tan, Li'10)  | $O(MN)$ (Similar complexity as xAMP, see matrix form of the BP-SBL in the upcoming slides) | $\log(MN)$                     | Does not give exact sparsity  | Posterior of x computed using BP and EM for hyper-parameters   | Convergence to local optimum of Bethe Free Energy (BFE)         |
| GAMP-SBL (Shoukairi, Schniter, Rao'18)   | $O(MN)$  | $\ll M$                        | Does not give exact sparsity  | Using GAMP for posterior of x, EM for hyperparameters  | Convergence to local optimum of LSL-BFE                         |
| SAVE   | $O(MN)$  | $\ll M$                        | Does not give exact sparsity  | Maximization of ELBO in VB   | Convergence to a local optimum of ELBO                          |
| Inverse Free SBL (Duan, Yang, Fang, Li'17)                                     | $O(MN)$  | $\ll M$ (similar to GAMP SBL)  | Does not give Exact sparsity  | Maximization of an approximate ELBO in VB  | Convergence to a local optimum of the approximate ELBO          |

# Outline

- 1 Introduction
- 2 Static SBL
- 3 Combined BP-MF-EP Framework**
- 4 Posterior Variance Prediction: Bayes Optimality
- 5 Performance Analysis of Approximate Inference Techniques (mCRB)
- 6 Dynamic SBL
- 7 Kronecker Structured Dictionary Learning using BP/VB
- 8 Numerical Results and Conclusion

# Approximate Inference Cost Functions: An Overview

- ML min. KLD of  $p_{\mathbf{y}}(\mathbf{y}|\theta)$  to empirical distribution of  $\mathbf{y}$  ( $p_{\mathbf{y}}(\mathbf{y}) = \delta(\mathbf{y} - \mathbf{y})$ ):

$$\theta_{min,KL} = \arg \min_{\theta} D_{KL}(p_{\mathbf{y}}(\mathbf{y}) || p_{\mathbf{y}}(\mathbf{y}|\theta)) = \arg \max_{\theta} \ln(p_{\mathbf{y}}(\mathbf{y}|\theta)) = \theta_{MLE}.$$

- VB minimizes KLD of factored approximate posterior ( $q(\theta) = \prod_i q_{\theta_i}(\theta_i)$ ):

$$KLD_{VB} = D_{KL}(q(\theta) || p(\theta|\mathbf{y})).$$

- Variational Free Energy (VFE) ( $U(q) =$  Average System Energy,  $H(q) =$  Entropy). Assume actual posterior  $p(\theta|\mathbf{y}) = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})} = \frac{\prod_a p_a(\theta_a)}{Z}$  and  $F_H = -\ln Z$  (Helmholtz Free Energy or log-partition function).

$$F(q(\theta)) = D_{KL}(q(\theta) || p(\theta|\mathbf{y})) + F_H = - \underbrace{\sum_{\theta} q(\theta) \sum_a \ln p_a(\theta_a)}_{U(q)} + \underbrace{\sum_{\theta} q(\theta) \ln q(\theta)}_{-H(q)} = D_{KL}(q(\theta) || \prod_a p_a(\theta_a)).$$

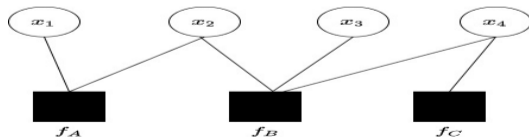


Figure 5: A small factor graph representing the posterior  $p(x_1, x_2, x_3, x_4) = \frac{1}{Z} f_A(x_1, x_2) f_B(x_2, x_3, x_4) f_C(x_4)$ <sup>26</sup>.

<sup>26</sup>Yedidia, Freeman, Weiss'05

## Approximate Inference Cost Functions: An Overview (2)

- $F(q) \geq F_H$ , equality only if  $q(\theta) = p(\theta|y)$ . Practical approach: upper bound  $F_H$  by minimizing  $F(q)$  over a restricted class of probability distributions leading to Kikuchi, BP or MF approximations.
- Belief Propagation (BP) minimizes Bethe Free Energy (BFE), Mean Field (MF) minimizes MFFE (MF Free Energy). BP converges to exact posterior when the factor graph is a tree. For MF (VB pushed to scalar level),  $q(\theta) = \prod_{i=1}^M q_{\theta_i}(\theta_i)$ .
- $$MFFE \geq BFE \geq VFE.$$
- Region based Free Energy approximations (RFE) (more details in the next slide): The intuitive idea behind a RFE approximation is to break up the factor graph into a set of large regions that include every factor and variable node, and say that the overall free energy is the sum of the free energies of all the regions. BP is a special case of this.
- Expectation Propagation (EP): derived using BFE under moment matching constraints.

# Region Based Free Energy

- A region  $R$  of a factor graph to be a set  $\mathcal{V}_R$  of variable nodes and set  $\mathcal{A}_R$  of factor nodes, such that  $a \in \mathcal{A}_R \implies$  all variable nodes connected to  $a$  are in  $\mathcal{V}_R$ .  $\theta_R$  is defined as the set of all variable nodes belonging to the region  $R$ .

- Region energy is defined as  $E_R(\theta_R) = - \sum_{a \in \mathcal{A}_R} \ln p_a(\theta_a)$ .

- Region free energy using region entropy and region average energy:

$$U_R(q_R) = \sum_{\theta_R} q_R(\theta_R) E_R(\theta_R), \quad H_R(q_R) = \sum_{\theta_R} q_R(\theta_R) \ln q_R(\theta_R).$$

and  $F_R(q_R) = U_R(q_R) - H_R(q_R)$ .

- Region-based free energy using region-based entropy and region-based average energy:

$$U_{\mathcal{R}}(\{q_R\}) = \sum_{R \in \mathcal{R}} c_R U_R(q_R), \quad H_{\mathcal{R}}(\{q_R\}) = \sum_{R \in \mathcal{R}} c_R H_R(q_R).$$

and  $F_{\mathcal{R}}(\{q_R\}) = U_{\mathcal{R}}(\{q_R\}) - H_{\mathcal{R}}(\{q_R\})$ .

- The intuitive idea: break up the factor graph into a set of large regions that include every factor and variable node, and say that **the overall VFE is the sum of the VFEs of all the regions**. If some of the large regions overlap, then we will have erred by counting the free energy contributed by some nodes two or more times, so we then need to **subtract out the free energies of these overlap regions in such a way that each factor and variable node is counted exactly once** (weight  $c_R$  takes care of this).
- BP: **Each factor node (and it's neighbouring variable nodes) form one set of regions**. Another set of regions which contain **only one variable node**.

# Variational Free Energy (VFE) Framework

- Intractable joint posterior distribution of the parameters  $\theta = \{\mathbf{x}, \mathbf{A}, \mathbf{f}, \mathbf{\Gamma}, \gamma\}$ .
- Actual posterior:  $p(\theta) = \frac{1}{Z} \underbrace{\prod_{a \in \mathcal{A}_{BP}} p_a(\theta_a) \prod_{b \in \mathcal{A}_{MF}} p_b(\theta_b)}_{\text{factor nodes}}$ , where  $\mathcal{A}_{BP}, \mathcal{A}_{MF}$  = set of factor nodes belonging to the BP/MF part with  $\mathcal{A}_{BP} \cap \mathcal{A}_{MF} = \emptyset$ .
- The whole  $\theta$  is partitioned into the set  $\theta_i$  (variable nodes), and we want to approximate the true posterior  $p(\theta)$  by an approximate posterior  $q(\theta) = \prod_i q_i(\theta_i)$ .
- $\mathcal{N}_{BP}(i), \mathcal{N}_{MF}(i)$  – 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<sup>27</sup> are written as below (let  $q_i(\theta_i)$  represents the belief about  $\theta_i$  (the approximate posterior))

$$F_{BP, MF} = \sum_{a \in \mathcal{A}_{BP}} [D_{KL}(q_a(\theta_a) \| p_a(\theta_a)) + D_{KL}(q_a(\theta_a) \| \prod_{i \in \mathcal{N}(a)} q_i(\theta_i))] + \sum_{b \in \mathcal{A}_{MF}} D_{KL}(\prod_{i \in \mathcal{N}(b)} q_i(\theta_i) \| p_b(\theta_b)),$$

<sup>27</sup>Riegler, Kirkelund, Manchón, Fleury'13



# Message Passing (MP) Expressions

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

$$\sum_{\theta_i} q_i(\theta_i) = 1, \forall i \in \mathcal{I}_{MF} \setminus \mathcal{I}_{BP}, \quad \sum_{\theta_a} q_a(\theta_a) = 1, \forall a \in \mathcal{A}_{BP},$$

$$q_i(\theta_i) = \sum_{\theta_a \setminus \theta_i} q_a(\theta_a), \quad \forall a \in \mathcal{A}_{BP}, i \in \mathcal{N}(a).$$

- The **fixed point equations of the constrained optimization of the approximate VFE:**

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

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

$$m_{a \rightarrow i}^{MF}(\theta_i) = \exp(\langle \ln p_a(\theta_a) \rangle_{\prod_{j \in \mathcal{N}(a) \setminus i} n_{j \rightarrow a}(\theta_j)}), \quad \langle \cdot \rangle_q \text{ is the expectation w.r.t } q \quad (2)$$

$$m_{a \rightarrow i}^{BP}(\theta_i) = \langle p_a(\theta_a) \rangle_{\prod_{j \in \mathcal{N}(a) \setminus i} n_{j \rightarrow a}(\theta_j)} \quad \text{factor to variable nodes}$$

- Expectation Propagation (EP):** The constraints in BFE can often be too complex to yield computationally tractable messages, the following constraint relaxation leads to EP<sup>28</sup>.

$$E_{q_a}(t(\theta_i)) = E_{q_i}(t(\theta_i)) \implies m_{a \rightarrow i}^{BP}(\theta_i) = \frac{\text{Proj}_{\phi}(\int q_a(\theta_a) \prod_{j \in \mathcal{N}(a), j \neq i} d\theta_j)}{n_{i \rightarrow a}(\theta_i)}, \quad q_a(\theta_a) = \frac{1}{z_a} p_a(\theta_a) \prod_{j \in \mathcal{N}(a)} n_{j \rightarrow a}(\theta_j)$$

where  $\phi$  represents the family of distributions characterized by the sufficient statistics  $t(\theta_i)$ .

<sup>28</sup>Minka'01

# What do the MP Expressions Indicate?

- BP-MF combo = alternating optimization of Lagrangian<sup>29</sup>:

$$\mathcal{L} = F_{BP, MF} + \sum_a \gamma_a [\sum_{\theta_a} q_a(\theta_a) - 1] + \sum_i \gamma_i [\sum_{\theta_i} q_i(\theta_i) - 1] + \sum_i \sum_{a \in \mathcal{N}(i)} \sum_{\theta_i} \lambda_{ai}(\theta_i) [q_i(\theta_i) - \sum_{\theta_a \setminus \theta_i} q_a(\theta_a)].$$

- At any iteration or convergence:

$$q_a(\theta_a) = p_a(\theta_a) \left( \prod_{i \in \mathcal{N}(a)} q_i(\theta_i) \exp[-\lambda_{ai}(\theta_i)] \right) \exp[\gamma_a - 1] = \frac{1}{z_a} p_a(\theta_a) \prod_{i \in \mathcal{N}(a)} \underbrace{\frac{q_i(\theta_i)}{m_{a \rightarrow i}(\theta_i)}}_{n_{i \rightarrow a}(\theta_i)}, \quad a \in \mathcal{A}_{BP}$$

$$q_i(\theta_i) = \underbrace{\exp[|\mathcal{N}_{BP}(i)| - 1 + \mathbb{I}_{\mathcal{I}_{MF} \setminus \mathcal{I}_{BP}}(i) \gamma_i]}_{1/z_i} \prod_{a \in \mathcal{N}_{MF}(i)} \underbrace{\exp(\langle \ln p_a(\theta_a) \rangle_{q_j(\theta_j), j \in \mathcal{N}(a) \setminus i})}_{m_{a \rightarrow i}^{MF}(\theta_i)} \prod_{a \in \mathcal{N}_{BP}(i)} \underbrace{\exp(\lambda_{ai}(\theta_i))}_{m_{a \rightarrow i}^{BP}(\theta_i)}.$$

where  $\mathbb{I}_{\mathcal{A}}(i) =$  indicator function for  $i \in \mathcal{A}$ .

- Applying the marginalization constraint  $q_i(\theta_i) = \sum_{\theta_a \setminus \theta_i} q_a(\theta_a)$ ,  $\forall a \in \mathcal{A}_{BP}$  leads to the expression for

$m_{a \rightarrow i}^{BP}(\theta_i)$  as in (2).

- The Lagrange multipliers  $\lambda_{ai}$  are indeed the log of the BP messages and  $\gamma_a, \gamma_i$  lead to the normalization constants  $z_a, z_i$  for the beliefs  $q_a(\theta_a), q_i(\theta_i)$ , respectively.

$$\lambda_{ai}(\theta_i) = \ln m_{a \rightarrow i}^{BP}(\theta_i).$$

<sup>29</sup>Yedidia, Freeman, Weiss'05

# Outline

- 1 Introduction
- 2 Static SBL
- 3 Combined BP-MF-EP Framework
- 4 Posterior Variance Prediction: Bayes Optimality**
- 5 Performance Analysis of Approximate Inference Techniques (mCRB)
- 6 Dynamic SBL
- 7 Kronecker Structured Dictionary Learning using BP/VB
- 8 Numerical Results and Conclusion

## SBL using BP: Predictive Posterior Variance Bayes Optimality

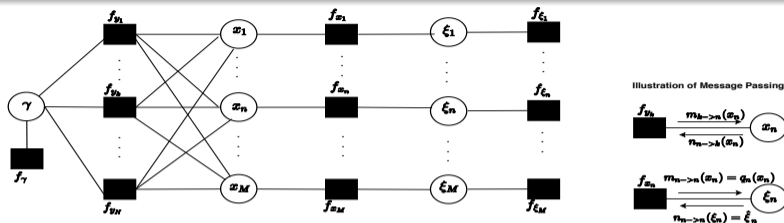


Figure 6: Factor Graph for the static SBL. Dark square nodes are the factor nodes and circle nodes represent the variable nodes.

- All the messages (beliefs or continuous pdfs) passed between them are all Gaussian<sup>30</sup>. So in message passing (MP), it suffices to represent them by two parameters, which are the mean and variance of the beliefs.
- We represent  $\sigma_{n,k}^{-2}$  as the inverse variance (precision) of the message passed from variable node  $n$  (corresponding to  $x_n$ ) to factor node  $k$  (corresponds to  $y_k$ ) and  $\hat{x}_{n,k}$  be the mean of the message passed from  $n$  to  $k$ , total  $NM$  of them.
- Similarly  $\sigma_{k,n}^{-2}, \hat{x}_{k,n}$  for messages from  $k$  to  $n$ .

<sup>35</sup>Tan, Li'10

## SBL using BP: Message Passing Expressions

- We start with the MP expressions derived in<sup>36</sup>. Define the matrix  $\mathbf{S}$  with entries  $\sigma_{k,n}^{-2}$ . The Gaussian beliefs are parameterized as  $m_{k \rightarrow n}(x_n) = \mathcal{N}(x_n; \hat{x}_{k,n}, \sigma_{k,n}^2)$  and  $n_{n \rightarrow k}(x_n) = \mathcal{N}(x_n; \hat{x}_{n,k}, \sigma_{n,k}^2)$ .
- Interpretation of  $m_{n \rightarrow k}(x_n)$ : Bayesian information combining:** At variable node  $n$ , we have

$$\hat{\mathbf{x}}_n = \begin{bmatrix} \hat{x}_{1,n} \\ \vdots \\ \hat{x}_{N,n} \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} x_n + \mathcal{N}(\tilde{\mathbf{x}}_n; \mathbf{0}, \text{diag}(\mathbf{S}_{:,n})^{-1}) \text{ with prior } \mathcal{N}(x_n; 0, \xi_n^{-1}).$$

- $x_n, \hat{\mathbf{x}}_{\bar{k},n}$  ( $\hat{\mathbf{x}}_n$  excluding  $\hat{x}_{k,n}$ ) are jointly Gaussian and hence lead to "extrinsic" "posterior" message for node  $k$ :

$$\hat{x}_{n,k} = \sigma_{n,k}^2 \sum_{i \neq k} \sigma_{i,n}^{-2} \hat{x}_{i,n}, \quad \sigma_{n,k}^{-2} = \xi_n + \sum_{i \neq k} \sigma_{i,n}^{-2}.$$

- Interpretation of  $m_{k \rightarrow n}(x_n)$ : Interference Cancellation:** Substituting  $x_m = \hat{x}_{m,k} + \tilde{x}_{m,k}$  ("extrinsic" information from variables  $m \neq n$  for measurement  $k$ ) in  $y_k = \sum_m A_{k,m} x_m + v_k$  leads to the 1-1 measurement

$$(y_k - \sum_{m \neq n} A_{k,m} \hat{x}_{m,k}) = A_{k,n} x_n + (v_n + \sum_{m \neq n} A_{k,m} \tilde{x}_{m,k}),$$

$$\text{with total "noise" } v_n + \sum_{m \neq n} A_{k,m} \tilde{x}_{m,k} \text{ of variance } \gamma^{-1} + \sum_{m \neq n} A_{k,m}^2 \sigma_{m,k}^2.$$

So the (deterministic) estimate and variance from this measurement by itself are

$$\hat{x}_{k,n} = A_{k,n}^{-1} (y_k - \sum_{m \neq n} A_{k,m} \hat{x}_{m,k}) \text{ and } \sigma_{k,n}^{-2} = A_{k,n}^2 \left( \frac{1}{\gamma} + \sum_{m \neq n} A_{k,m}^2 \sigma_{m,k}^2 \right)^{-1}.$$

<sup>36</sup>Tan, Li'10

# SBL using BP: MP Expressions in Matrix Form

- **Posterior marginals:**  $x_n, \hat{x}_f$  are jointly Gaussian and hence MMSE estimate leads to the messages  $\mathcal{CN}(x_n; \hat{x}_n, \sigma_n^2)$ :  $\sigma_n^2 = (\xi_n + \sum_i \sigma_{i,n}^{-2})^{-1}$ ,  $\hat{x}_n = \sigma_n^2 (\sum_i \sigma_{i,n}^{-2} \hat{x}_{i,n})$ .
- **In matrix form** ( $\mathbf{S}', \mathbf{M}'$  of dimension  $M \times N$ ,  $\mathbf{S}, \mathbf{M}$  of dimension  $N \times M$  with entries  $\sigma_{n,k}^{-2}, \hat{x}_{n,k}, \sigma_{k,n}^{-2}, \hat{x}_{k,n}$ , respectively):

$$\begin{aligned} \mathbf{S}' &= \Xi \mathbf{1}_M \mathbf{1}_N^T + \mathbf{S}^T (\mathbf{1}_N \mathbf{1}_N^T - \mathbf{I}_N), \\ \mathbf{L} &= \text{diag}(\mathbf{S}^T \mathbf{M}) \mathbf{1}_N \mathbf{1}_N^T - (\mathbf{S} \circ \mathbf{M})^T, \quad \mathbf{M}' = \mathbf{S}'_{inv} \circ \mathbf{L}, \quad \mathbf{M}'_{n,k} = \hat{x}_{n,k}. \end{aligned}$$

Similarly, for the messages at the factor nodes, define  $\mathbf{C}$  to be the matrix with entries  $A_{k,n}^2 \sigma_{k,n}^2$  ( $\circ$  represents Hadamard (element-wise) product,  $\mathbf{A}_{inv}$  denotes element-wise inverse.)

$$\begin{aligned} \mathbf{C} &= \left( \frac{1}{\gamma} \mathbf{I}_N + \text{diag}(\mathbf{B} \mathbf{S}'_{inv}) \right) (\mathbf{1}_N \mathbf{1}_M^T) - \mathbf{B} \circ \mathbf{S}'_{inv}^T, \quad \mathbf{S} = \mathbf{C}_{inv} \circ \mathbf{B}_{inv}, \quad \mathbf{B} = \mathbf{A} \circ \mathbf{A}, \\ \mathbf{V} &= (\mathbf{y} - \text{diag}(\mathbf{A} \mathbf{M}') \mathbf{1}_N) \mathbf{1}_M^T + \mathbf{A} \circ \mathbf{M}'^T, \quad \mathbf{M} = \mathbf{A}_{inv} \circ \mathbf{V}, \end{aligned}$$

- Computational complexity  $\mathcal{O}(dMN)$ ,  $d \ll M, N$ .

## Existing Convergence Conditions of Gaussian BP

- In loopy GaBP, if the mean of the posterior belief converges, it converges to the true posterior<sup>37</sup>. Independently analyzed in<sup>38</sup>.
- Posterior variances (if initialized with values  $> 0$ ) always converge to a unique stationary point, but need not to the true posterior variance.
- Further in<sup>39</sup> show that the convergence condition of GaBP can be shown to be related to the spectral radius of a matrix  $|R|$  (element-wise absolute values), where  $J = I - R$ , with  $J = \gamma A^T A + \Xi$ , which is indeed the posterior precision matrix.
- Diagonally dominant  $J$  is one such example which satisfies this condition.

---

<sup>37</sup>Rusmevichientong, Van Roy'01

<sup>38</sup>Weiss, Freeman'01

<sup>39</sup>Malioutov, Johnson, Willsky'06

## Existing Convergence Conditions of Gaussian BP (Cont'd)

- In<sup>40</sup> shows that depending on the underlying graphical structure (Gaussian Markov Random Field (GMRF) or factor graph based factorization) Gaussian BP (GaBP) may exhibit different convergence properties.
- They prove that the convergence condition for the mean provided based on the factor graph representation encompasses much larger class of models than those given by the GMRF based **walk-sumnable condition**<sup>41</sup>.
- GaBP always converges if the **factor graph is a union of a single loop and a forest (a forest is a disjoint union of trees)**.
- Moreover, they also analyze the convergence of the inverse of the message variances (message information matrix) and analytically show that with arbitrary positive semidefinite matrix initialization, **the message information matrix converges to a unique positive definite matrix**.
- So we can conclude that for BP there is a **decoupling between the dynamics of the variance updates and that of the mean updates**.
- (Generalized) approximate message passing (GAMP or AMP)** or their variant **vector approximate message passing (VAMP)** exhibit convergence to **Bayes Optimal MMSE** for **i.i.d.** or **right orthogonally invariant matrices A**.

<sup>40</sup>Du, Ma, Wu, Kar, Moura'18, <sup>41</sup>Malioutov, Johnson, Willsky'06



# Large System Analysis: Useful Results

## Theorem 3 (Theorem 1<sup>42</sup>)

Let  $\mathbf{Q}_M \in \mathbb{C}^{M \times M}$  be a Hermitian deterministic matrix and  $\mathbf{A}_M = \mathbf{X}_M \mathbf{D} \mathbf{X}_M^H = \sum_{i=1}^N d_i \mathbf{x}_i \mathbf{x}_i^H$ , with diagonal  $\mathbf{D}$  and  $\mathbf{X}_M$  containing  $N$  independent columns  $\mathbf{x}_i$  with covariance matrix  $\Theta_i$ . Also, assume that  $\mathbf{Q}_M, \Theta_i$  have uniformly bounded spectral norms. Then, for any  $z > 0$

$$\frac{1}{M} \text{tr}\{\mathbf{Q}_M (\mathbf{A}_M + z \mathbf{I}_M)^{-1}\} - \frac{1}{M} \text{tr}\{\mathbf{Q}_M \mathbf{T}(z)\} \xrightarrow[M \rightarrow \infty]{a.s.} 0, \text{ with,}$$

$$\mathbf{T}(z) = \left( \sum_{i=1}^N \frac{d_i \Theta_i}{1 + e_i(z)} + z \mathbf{I}_M \right)^{-1}, \text{ where,}$$

$e_i(z) = e_i^{(\infty)}(z)$  is defined as the unique positive solution of

$$e_i(z) = \frac{1}{M} \text{tr}\{d_i \Theta_i \left( \sum_{i=1}^N \frac{d_i \Theta_i}{1 + e_i(z)} + z \mathbf{I}_M \right)^{-1}\}.$$

## Lemma 4 (Lemma 4, Appendix VI, WagnerTIT2012)

$\mathbf{x}_M^H \mathbf{A}_M \mathbf{x}_M - \frac{1}{M} \text{tr}\{\mathbf{A}_M\} \xrightarrow[M \rightarrow \infty]{} 0$  when the elements of  $\mathbf{x}_M$  are iid with zero mean and variance  $1/M$  and independent of  $\mathbf{A}_M$ , and similarly when  $\mathbf{y}_M$  is independent of  $\mathbf{x}_M$ , that  $\mathbf{x}_M^H \mathbf{A}_M \mathbf{y}_M \xrightarrow[M \rightarrow \infty]{a.s.} 0$ .

<sup>42</sup>Wagner, Couillet, Debbah, Slock'12

# Bayes Optimality of Per Component MSE of Gaussian BP

## Theorem 5

In the large system limit (LSL), under i.i.d  $\mathbf{A}$ , the predicted (by BP or xAMP algorithms) per component MSE (or the posterior variance  $\sigma_n^2$ ) converges exactly to the Bayes optimal values (i.e. the diagonal elements of the posterior covariance matrix for LMMSE). This result being applicable for AMP (Generalized AMP (GAMP) also under i.i.d  $\mathbf{A}$ ), since the derivation of AMP follows from BP under the LSL.

### Outline of the derivation:

- In the large system limit, we can approximate (neglecting terms of  $\mathcal{O}(A_{i,j}^2)$ )  $\sigma_{n,k}^{-2} = \xi_n + \sum_i \sigma_{i,n}^{-2} = \sigma_n^{-2}$ , independent of  $k$ . Further we define  $\mathbf{S} = \text{diag}(\sigma_n^{-2})$ .
- Considering the term  $\sigma_{k,n}^{-2} = A_{k,n}^2 (\frac{1}{\gamma} + \sum_{m \neq n} A_{k,m}^2 \sigma_{m,k}^{-2})^{-1}$ , in the LSL it can be approximated by  $\sigma_{k,n}^{-2} = A_{k,n}^2 (\frac{1}{\gamma} + \mathbf{A}_{k,:} \mathbf{S}^{-1} \mathbf{A}_{k,:}^T)^{-1}$ .  $\mathbf{A}_{k,:} \mathbf{S}^{-1} \mathbf{A}_{k,:}^T \xrightarrow[a.s]{M \rightarrow \infty} \frac{1}{N} \text{tr}\{\mathbf{S}^{-1}\} = \tau_{BP}$ .  $\mathbf{A}_{k,:}$  represents the  $k^{\text{th}}$  row of  $\mathbf{A}$ . From posterior belief variances, it follows that  $MSE = \text{tr}\{\mathbf{S}^{-1}\}$ . Further we obtain,  $\sigma_n^{-2} = \xi_n + (\frac{1}{\gamma} + \tau_{BP})^{-1} \sum_i A_{i,n}^2$ ,  $\sum_i A_{i,n}^2 \xrightarrow[a.s]{M \rightarrow \infty} 1$ , thus  $\sigma_n^{-2} = \xi_n + (\frac{1}{\gamma} + \tau_{BP})^{-1}$ .
- Define:  $\mathbf{A}_{\bar{i}}$  represents the matrix obtained by removing the  $i^{\text{th}}$  column of  $\mathbf{A}$ . Similarly, we define  $\Xi_{\bar{i}}$ .

## Outline of Derivation

$$\tau_{BP} = \frac{1}{N} \sum_{n=1}^M (\xi_n + (\frac{1}{\gamma} + \tau_{BP})^{-1})^{-1}. \quad (3)$$

Next step is to simplify the expression for LMMSE posterior covariance in the LSL using similar techniques as above.

$$\begin{aligned} \Sigma_L &= \Xi^{-1} - \Xi^{-1} \mathbf{A}^T (\mathbf{A} \Xi^{-1} \mathbf{A}^T + \frac{1}{\gamma})^{-1} \mathbf{A} \Xi^{-1}, \\ \mathbf{A}_i^T (\mathbf{A} \Gamma^{-1} \mathbf{A}^T + \frac{1}{\gamma})^{-1} \mathbf{A}_i &\xrightarrow{(a)} \mathbf{D}_{i,i}, \quad \mathbf{D}_{i,i} = \frac{e}{1 + \frac{e}{\xi_i}} \end{aligned}$$

where (a) follows from first applying matrix inversion lemma and then Theorem 1 in <sup>43</sup> to the term  $\mathbf{A}_i^T (\mathbf{A}_i \Gamma_i^{-1} \mathbf{A}_i^T + \frac{1}{\gamma})^{-1} \mathbf{A}_i$  in the denominator and  $e$  is defined as the unique positive solution of the following fixed point equation ( $\frac{1}{N} \text{tr}\{\Sigma_L\} = \tau$ ),

$$e = \left( \frac{1}{N} \sum_{i=1}^M \frac{\xi_i^{-1}}{1 + \frac{e}{\xi_i}} + \frac{1}{\gamma} \right)^{-1}, \quad \tau = \frac{1}{N} \sum_{i=1}^M \frac{\xi_i^{-1}}{1 + \frac{e}{\xi_i}},$$

$$\text{From } e, \quad \frac{1}{e} - \frac{1}{\gamma} = \frac{1}{N} \sum_{i=1}^M \frac{\xi_i^{-1}}{1 + \frac{e}{\xi_i}}, \quad \frac{1}{e} = \frac{1}{\gamma} + \tau, \quad (4)$$

$$\tau = \frac{1}{N} \sum_{i=1}^M \frac{\xi_i^{-1}}{(\frac{1}{\gamma} + \tau)^{-1} \xi_i^{-1} + 1} = \frac{1}{N} \sum_{i=1}^M \frac{1}{\xi_i + (\frac{1}{\gamma} + \tau)^{-1}}.$$

**Conclusion:** From (3), (4),  $\tau, \tau_{BP}$  can be obtained as the solution of same fixed point equation, which also proves that per component MSE is Bayes optimal (comparing expressions of  $\sigma_n^2$  and  $(\Sigma_L)_{n,n}$ ).

<sup>43</sup>Wagner, Couillet, Debbah, Slock'12

# State of the Art: Approximate MP (xAMP)

- AMP<sup>44</sup> is originally derived from Gaussian approximations of loopy BP and first order Taylor series approximations.
- AMP is proven to be asymptotically Bayes optimal in MMSE (only for i.i.d.  $\mathbf{A}$ ).
- Generalized AMP (GAMP)<sup>45</sup> - AMP generalized to arbitrary input and output product distributions. Applications in nonlinear (e.g. amplitude only) compressed sensing, 1-bit ADC communication systems, etc. However, state evolution (SE) only for i.i.d.  $\mathbf{A}$ .
- S-AMP<sup>46</sup> extends AMP to more general matrix ensembles (similar to VAMP). The fixed points of S-AMP are stationary points of (EP-)VFE under a set of moment consistency constraints in the large system limit (LSL).
- Vector AMP (VAMP)<sup>47</sup> - rigorous scalar SE that holds for the much broader class of right-orthogonally invariant random matrices  $\mathbf{A}$ .
- ADMM-GAMP<sup>48</sup> - GAMP algorithm based on direct minimization of a LSL approximation of the BFE (LSL-BFE), convergent for much wider class of  $\mathbf{A}$  compared to GAMP.

---

<sup>44</sup>Bayati, Montanari'11

<sup>45</sup>Rangan'11

<sup>46</sup>Çakmak, Winther, Fleury'14

<sup>47</sup>Rangan, Schniter, Fletcher'19

<sup>48</sup>Rangan, Fletcher, Schniter, Kamilov'17

# Posterior Mean in the Large System Limit (LSL)

Further defining the following terms,

$$z_{k,n} = y_k - \sum_{m \neq n} A_{k,m} \hat{x}_{m,k}, \text{ So } \hat{x}_{k,n} = A_{k,n}^{-1} z_{k,n}. \quad (5)$$

Also, in the LSL,  $\hat{x}_{n,k}$  can be written as,  $\hat{x}_{n,k} = \hat{x}_n + \delta_{n \rightarrow k}$ , where  $\delta_{n \rightarrow k}$  is of the  $O(\frac{1}{\sqrt{N}})$ . This approximation follows from writing  $\hat{x}_{n,k} = \sigma_{n,k}^2 (\sum_i \sigma_{i,n}^{-2} \hat{x}_{i,n} - \sigma_{k,n}^{-2} \hat{x}_{k,n}) = \hat{x}_n + \delta_{n \rightarrow k}$ , with

$\delta_{n \rightarrow k} = \sigma_{n,k}^2 \sigma_{k,n}^{-2} \hat{x}_{k,n}$ , where  $\sigma_{k,n}^{-2} \hat{x}_{k,n} \propto A_{k,n} \propto \frac{1}{\sqrt{N}}$ . Substituting  $\hat{x}_{n,k}$  in  $z_{k,n}$ ,

$z_{k,n} = y_k - \sum_m A_{k,m} \hat{x}_m - \sum_m A_{k,m} \delta_{m \rightarrow k} + A_{k,n} \hat{x}_n + O(\frac{1}{N}) = z_k + \delta_{k \rightarrow n}$ , all the terms containing  $A_{i,j}^2$  or

$A_{i,j} \delta_{j \rightarrow i}$  becomes  $O(\frac{1}{N})$  and  $\delta_{k \rightarrow n} = A_{k,n} \hat{x}_n$ , also here

$$z_k = y_k - \sum_m A_{k,m} \hat{x}_m - \sum_m A_{k,m} \delta_{m \rightarrow k}. \quad (6)$$

$$\hat{x}_{n,k} \approx \sigma_n^2 (\frac{1}{\gamma} + \tau_{BP})^{-1} \sum_{i \neq k} A_{i,n} z_{i,n}. \quad (7)$$

We can write  $\hat{x}_{n,k} = f_n(\sum_{i \neq k} A_{i,n} z_{i,n})$ . Here  $f_n$  is a linear function for the Gaussian case (i.e.

$f_n(x) = \sigma_n^2 (\frac{1}{\gamma} + \tau_{BP})^{-1} x$ ).

## Posterior Mean in the LSL (Onsager Correction)

Performing a **first order Taylor series approximation of  $f$**  around

$$\sum_i A_{i,n} z_{i,n}, \hat{x}_{n,k} = f_n(\sum_i A_{i,n} z_{i,n}) - A_{k,n} z_{k,n} f'_n(\sum_i A_{i,n} z_{i,n}), f'_n \text{ being derivative evaluated at } \sum_i A_{i,n} z_{i,n}.$$

Further substituting for  $z_{i,n}$  from (5),

$$\begin{aligned} \hat{x}_{n,k} &= \hat{x}_n + \delta_{n \rightarrow k}, \quad \hat{x}_n = f_n(\sum_i A_{i,n} z_i + \sum_i A_{i,n} \delta_{i \rightarrow n}) \\ \text{and } \delta_{n \rightarrow k} &= -A_{k,n} z_k f'_n(\sum_i A_{i,n} z_i). \end{aligned} \quad (8)$$

Substituting for  $\delta_{i \rightarrow n} = A_{i,n} \hat{x}_n$  and with the large system approximation  $\sum_i A_{i,n}^2 \xrightarrow{M \rightarrow \infty} 1$ ,

$$\hat{x}_n = f_n(\sum_i A_{i,n} z_i + \sum_i A_{i,n}^2 \hat{x}_n) = f_n(\sum_i A_{i,n} z_i + \hat{x}_n).$$

In vector form:  $\hat{\mathbf{x}} = \mathbf{f}(\mathbf{A}^T \mathbf{z} + \hat{\mathbf{x}})$ , which is the **AMP recursion for the mean**, where  $(\mathbf{f}(\mathbf{x}))_n = f_n(x_n)$ .

Also from (6), substituting  $\delta_{n \rightarrow k}$  from (8) and defining  $\mathbf{z}_t = [z_1, \dots, z_N]^T$  at iteration  $t$  :

$$\mathbf{z}_t = (\mathbf{y} - \mathbf{A} \hat{\mathbf{x}}_t) + \frac{1}{\delta} \mathbf{z}_{t-1} \langle \mathbf{f}'(\mathbf{A}^T \mathbf{z}_{t-1}) \rangle, \quad (9)$$

where  $\delta = \frac{N}{M}$  is a constant,  $\langle \mathbf{f}'(\mathbf{x}) \rangle = \frac{1}{M} \sum_{m=1}^M f'_m(x_m)$ , and  $\frac{1}{\delta} \mathbf{z}_{t-1} \langle \mathbf{f}'(\mathbf{A}^T \mathbf{z}_{t-1}) \rangle$  is the **Onsager term**.

# Original AMP Iterations and SBL-AMP

- The difference in AMP vs SBL-AMP is that in AMP  $f_m(x) = f(x)$ : same function for every component.
- The AMP iterations (for any Lipschitz-continuous component-wise shrinkage function  $\mathbf{f}$  and i.i.d  $\mathbf{x}$ ) can be written as

$$\mathbf{z}_t = \mathbf{y} - \mathbf{A}\hat{\mathbf{x}}_t + \frac{1}{\delta}\mathbf{z}_{t-1} < \mathbf{f}'(\hat{\mathbf{x}}_{t-1} + \mathbf{A}^T\mathbf{z}_{t-1}) >,$$

$$\hat{\mathbf{x}}_{t+1} = \mathbf{f}(\hat{\mathbf{x}}_t + \mathbf{A}^T\mathbf{z}_t).$$

- Onsager correction decouples the input to AMP<sup>49</sup>  $\mathbf{r}_t = \hat{\mathbf{x}}_t + \mathbf{A}^T\mathbf{z}_t = \mathbf{x} + \mathcal{N}(\mathbf{n}_t; \mathbf{0}, \tau_t\mathbf{I}_M)$

in case of  $\mathcal{N}(\mathbf{x}; \mathbf{0}, \frac{1}{\xi}\mathbf{I})$ , we get LMMSE  $\hat{\mathbf{x}}_{t+1} = \mathbf{f}(\mathbf{r}_t) = b_t\mathbf{r}_t$ ,  $b_t = \frac{\frac{1}{\xi}}{\frac{1}{\xi} + \tau_t}$

and State Evolution (SE)  $\tau_{t+1} = \frac{1}{\gamma} + \frac{1}{\delta}(1 - b_t)^2\frac{1}{\xi} + \frac{1}{\delta}b_t^2\tau_t = \frac{1}{\gamma} + \frac{1}{\delta}(\xi + \tau_t^{-1})^{-1}$ .

- SBL-AMP (for SBL  $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \Xi^{-1})$ ) - Iterations decouple  $\mathbf{r}_t$ :  $\mathbf{r}_t = \mathbf{x} + \mathcal{N}(\mathbf{n}_t; \mathbf{0}, \tau_t\mathbf{I})$  leading to  $\hat{\mathbf{x}}_{t+1} = \mathbf{f}(\mathbf{r}_t) = \mathbf{F}_t\mathbf{r}_t$ , with diagonal  $\mathbf{F}_t = (\mathbf{I}_M + \tau_t\Xi)^{-1}$ .  
Define  $\mathbf{A}_m$  as the  $m^{\text{th}}$  column of  $\mathbf{A}$  and  $\mathbf{A}_{\bar{m}}$  as the matrix excluding column  $m$ , vector  $\delta_{\bar{m} \rightarrow k}$  contains as entries  $\delta_{n \rightarrow k}$ ,  $n \neq m$ :

Consider  $m^{\text{th}}$  noise element  $n_{m,t} = \mathbf{A}_m^T\mathbf{A}_{\bar{m}}\tilde{\mathbf{x}}_{\bar{m},t} - \mathbf{A}_m^T\Delta_m + \mathbf{A}_m^T\mathbf{v}$ ,  $\Delta_{m,k} = \mathbf{A}_{k,\bar{m}}\delta_{\bar{m} \rightarrow k}$ ,

leading to  $\tau_{t+1} = \frac{1}{\gamma} + \frac{1}{\delta} \frac{1}{M} \sum_{n=1}^M (\xi_n + \tau_t^{-1})^{-1}$ .

<sup>49</sup>Bayati, Montanari'11

# Factor Graph for Vector BP-SBL

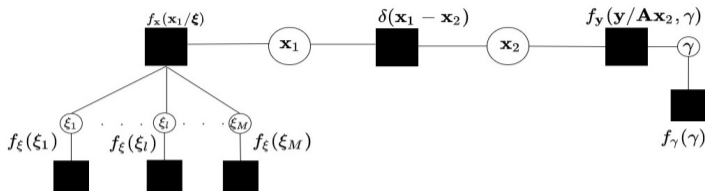


Figure 7: Factor Graph for the Vector BP from which GVAMP-SBL is derived.

Treating all measurements  $\mathbf{y}$  jointly leads to a tree structured factored graph, but no more extrinsic information between measurements, which motivates duplicating  $\mathbf{x}$ .

For the factor graph, we use the factorization of the posterior as follows

$$p(\mathbf{x}, \boldsymbol{\xi}, \gamma) \propto p_{\mathbf{y}}(\mathbf{y}/\mathbf{A}\mathbf{x}_2, \gamma^{-1}\mathbf{I})p_{\mathbf{x}}(\mathbf{x}_1/\boldsymbol{\xi})\delta(\mathbf{x}_1 - \mathbf{x}_2)[\prod_i p_{\xi_i}(\xi_i)]p_{\gamma}(\gamma),$$

where we created two identical variables  $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{x}$  similar to <sup>50</sup>.

<sup>50</sup>Rangan, Schniter, Fletcher'19



## Unitarily Invariant SBL using Vector AMP (VAMP)

- **Generalized AMP (GAMP)** and in particular **GAMP-SBL**<sup>51</sup> extends AMP to a non-i.i.d. prior but is limited to i.i.d.  $\mathbf{A}$ , leading to the introduction of damping to increase chances of convergence.
- Consider the economy SVD  $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ ,  $\mathbf{U}^T\mathbf{U} = \mathbf{I}_d$ ,  $\mathbf{V}^T\mathbf{V} = \mathbf{I}_d$ ,  $d = \text{rank}(\mathbf{A})$ .
- The class of **Right-Orthogonally Invariant (ROI)**  $\mathbf{A}$  considers a uniformly distributed random orthogonal factor  $\mathbf{V}$  (Haar distributed). ROI: the distribution of  $\mathbf{A}\mathbf{W}$  or  $\mathbf{V}^T\mathbf{W}$  is the same as that of  $\mathbf{A}$  or  $\mathbf{V}^T$  for any square orthogonal  $\mathbf{W}$ .
- **VAMP**<sup>52</sup> exploits ROI  $\mathbf{A}$  and its convergence is robust for a much large set of matrices  $\mathbf{A}$  than AMP. But VAMP does not apply directly to SBL since it is derived for i.i.d.  $\mathbf{x}$ .
- **Orthogonal AMP (OAMP)**<sup>53</sup> unitarily invariant AMP, using decorrelated linear estimation and divergence free nonlinear estimator (Onsager term vanishes).
- We propose **Generalized VAMP-SBL (GVAMP-SBL)** which combines ROI  $\mathbf{A}$  with non i.i.d.  $\mathbf{x}$  as needed for SBL.
- We also propose **SVD-GAMP-SBL** which is SBL-AMP applied to  $\mathbf{y}$ ,  $\mathbf{A}$  replaced by  $\mathbf{U}^T\mathbf{y}$ ,  $\mathbf{\Lambda}\mathbf{V}^T$  (SBL-AMP is GAMP-SBL for the case of i.i.d. Gaussian measurements).
- SBL using **UTAMP (AMP with unitary transformation)**<sup>54</sup>, derived from GAMP (using heuristics), is quite approximate due to the scalar EP (averaging of the different variance parameters in GAMP).

<sup>46</sup>Shoukairi, Rao'18, <sup>47</sup>Rangan, Schniter, Fletcher'19, <sup>48</sup>Ma, Ping'17, <sup>49</sup>Luo, Guo, Huang, Xi'19

## Useful Results for MMSE Estimation with Non-Gaussian Distributions

We use the following two results from Lemma 2<sup>50</sup>, which we restate here. For any random variable whose posterior distribution is of the form

$$f_x(x|r, \tau) = \frac{1}{Z(r)} \exp(\ln f(x) + \tau xr),$$

where  $Z(r)$  is the normalization constant. Then, the following relation between the mean and variance of the posterior for  $x$  holds

$$\begin{aligned} \frac{\partial}{\partial r} \ln Z(r) &= E(x|r) = g(r, \tau), \quad \text{"denoising function"} \\ \frac{\partial^2}{\partial r^2} \ln Z(r) &= g'(r, \tau) = \tau \text{Var}(x|r, \tau). \end{aligned}$$

Here,  $g'(r, \tau)$  represents the derivative w.r.t. the first argument  $r$  and  $\text{Var}(x|r, \tau)$  represents the variance of  $x$  w.r.t. the posterior distribution  $f_x(x|r, \tau)$ .

---

<sup>50</sup>Rangan'11

## GVAMP-SBL Derivation

- We start by initializing the MP with the Gaussian approximation as,  $m_{\delta \rightarrow x_1}(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1; \mathbf{r}_1, \text{Diag}(\boldsymbol{\tau}_1)^{-1})$ , so a diagonal EP instead of the scalar EP as in VAMP<sup>51</sup>. Diagonal EP being motivated by the diagonal prior covariance in SBL. Using the MP rules discussed, we can write the belief at the node  $\mathbf{x}_1$  as (we omit the iteration index),

$$q(\mathbf{x}_1) \propto p_{\mathbf{x}}(\mathbf{x}_1) \mathcal{N}(\mathbf{x}_1; \mathbf{r}_1, \text{Diag}(\boldsymbol{\tau}_1)^{-1}).$$

For a given estimate of the hyperparameter, we obtain the value of the mean of the belief as,  $\hat{\mathbf{x}}_{1,n} = \mathbf{g}_{1,n}(r_{1,n}, \tau_{1,n}^{-1})$ , where the expectation is w.r.t the density function

$$p(x_{1,n} | r_{1,n}, \tau_{1,n}^{-1}) = \exp \left[ -\frac{\tau_{1,n}}{2} |r_{1,n} - x_{1,n}|^2 + \ln p(x_{1,n}) \right].$$

The corresponding posterior variance can be obtained as,  $\eta_{1,n}^{-1} = \tau_{1,n}^{-1} \mathbf{g}'_{1,n}(r_{1,n}, \tau_{1,n}^{-1})$ .

- Diagonal EP:**  $\text{Proj}_{\phi}(q(\mathbf{x}_1)) = \mathcal{N}(\mathbf{x}_1; \hat{\mathbf{x}}_1, \text{Diag}(\boldsymbol{\eta}_1)^{-1})$ ,  $\phi$  represents the set of multivariate Gaussian with diagonal covariance.
- According to EP rule,  $n_{\mathbf{x}_1 \rightarrow \delta}(\mathbf{x}_1) (= m_{\delta \rightarrow \mathbf{x}_2}(\mathbf{x}_2))$ , which is the "extrinsic info" becomes

$$n_{\mathbf{x}_1 \rightarrow \delta}(\mathbf{x}_1) = \frac{\mathcal{N}(\mathbf{x}_1; \hat{\mathbf{x}}_1, \text{Diag}(\boldsymbol{\eta}_1)^{-1})}{\mathcal{N}(\mathbf{x}_1; \mathbf{r}_1, \text{Diag}(\boldsymbol{\tau}_1)^{-1})} = \mathcal{N}(\mathbf{x}_1; (\boldsymbol{\eta}_1 \cdot * \hat{\mathbf{x}}_1 - \boldsymbol{\tau}_1 \cdot * \mathbf{r}_1) ./ (\boldsymbol{\eta}_1 - \boldsymbol{\tau}_1), \text{Diag}(\boldsymbol{\eta}_1 - \boldsymbol{\tau}_1)^{-1}).$$

<sup>51</sup>Rangan, Schniter, Fletcher'19

## GVAMP-SBL (EP-BP, LSL justified // AMP)

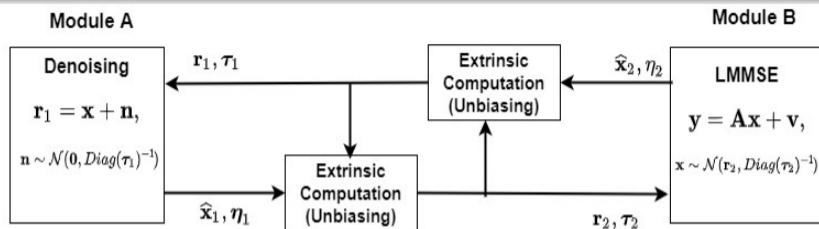


Figure 8: Illustration of the GVAMP-SBL.

- Further, we can obtain the belief at node  $\mathbf{x}_2$  as  $q(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2; \mathbf{r}_2, \text{Diag}(\boldsymbol{\tau}_2)^{-1})p_{\mathbf{y}}(\mathbf{y}/\mathbf{A}\mathbf{x}_2)$ . The point estimate (or the LMMSE mean) becomes after diagonal EP

$$\hat{\mathbf{x}}_2 = \mathbf{g}_2(\mathbf{r}_2, \boldsymbol{\tau}_2) = \underbrace{\text{diag}((\hat{\boldsymbol{\gamma}}\mathbf{A}^T\mathbf{A} + \text{Diag}(\boldsymbol{\tau}_2)^{-1})^{-1})}_{\text{Diag}(\boldsymbol{\eta}_2)} (\hat{\boldsymbol{\gamma}}\mathbf{A}^T\mathbf{y} + \text{Diag}(\boldsymbol{\tau}_2)^{-1}\mathbf{r}_2),$$

and after using SVD of  $\mathbf{A}$ ,  $\mathbf{g}'_2(\mathbf{r}_2, \boldsymbol{\tau}_2) = \text{diag}((\hat{\boldsymbol{\gamma}}\mathbf{V}\boldsymbol{\Lambda}^2\mathbf{V}^T + \text{Diag}(\boldsymbol{\tau}_2))^{-1})\text{Diag}(\boldsymbol{\tau}_2)^{-1}$ . Further, we obtain the "extrinsic update"  $n_{\mathbf{x}_2 \rightarrow \delta}(\mathbf{x}_2) (= m_{\delta \rightarrow \mathbf{x}_1}(\mathbf{x}_1))$  as follows

$$n_{\mathbf{x}_2 \rightarrow \delta}(\mathbf{x}_2) \propto \frac{\mathcal{N}(\mathbf{x}_2; \hat{\mathbf{x}}_2, \text{Diag}(\boldsymbol{\eta}_2)^{-1})}{\mathcal{N}(\mathbf{x}_2; \mathbf{r}_2, \text{Diag}(\boldsymbol{\tau}_2)^{-1})} = \mathcal{N}(\mathbf{x}_2; (\boldsymbol{\eta}_2 \cdot * \hat{\mathbf{x}}_2 - \boldsymbol{\tau}_2 \cdot * \mathbf{r}_2) ./ (\boldsymbol{\eta}_2 - \boldsymbol{\tau}_2), \text{Diag}(\boldsymbol{\eta}_2 - \boldsymbol{\tau}_2)^{-1}).$$

# Proposed GVAMP-SBL

**Initialization:**  $\tau_1^{(0)} \geq 0, \mathbf{r}_1^{(0)} = 0$ .

## Denoising

$$\hat{\mathbf{x}}_1^{(t)} = \mathbf{g}_1(\mathbf{r}_1^{(t)}, \tau_1^{(t)})$$

$$\boldsymbol{\eta}_1^{(t)} = \tau_1^{(t)} ./ \mathbf{g}'_1(\mathbf{r}_1^{(t)}, \tau_1^{(t)})$$

$$\tau_2^{(t)} = \boldsymbol{\eta}_1^{(t)} - \tau_1^{(t)}$$

$$\mathbf{r}_2^{(t)} = (\boldsymbol{\eta}_1^{(t)} .* \hat{\mathbf{x}}_1^{(t)} - \tau_1^{(t)} .* \mathbf{r}_1^{(t)}) ./ \tau_2^{(t)}$$

## LMMSE Estimation

$$\hat{\mathbf{x}}_2^{(t)} = \mathbf{g}_2(\mathbf{r}_2^{(t)}, \tau_2^{(t)})$$

$$\boldsymbol{\eta}_2^{(t)} = \tau_2^{(t)} ./ \mathbf{g}'_2(\mathbf{r}_2^{(t)}, \tau_2^{(t)})$$

$$\tau_1^{(t+1)} = \boldsymbol{\eta}_2^{(t)} - \tau_2^{(t)}$$

$$\mathbf{r}_1^{(t+1)} = (\boldsymbol{\eta}_2^{(t)} .* \hat{\mathbf{x}}_2^{(t)} - \tau_2^{(t)} .* \mathbf{r}_2^{(t)}) ./ \tau_1^{(t+1)}$$

## Hyperparameter Estimation (using MF [Section 3<sup>52</sup>])

$$\hat{\xi}_i^{(t)} = \frac{a+1/2}{|\hat{\mathbf{x}}_{1,i}^{(t)}|^2 + \boldsymbol{\eta}_{1,i}^{(t)} - 1}, \hat{\gamma}^{(t)} = \frac{c+N/2}{\langle \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 \rangle / 2 + d}$$

./ or .\* represent elementwise multiplication or division as in Matlab.

$x_{k,i}$  represents the  $i^{\text{th}}$  element in any vector  $x_k$  above.

<sup>52</sup>Thomas, Slock'18

# Proposed (LSL-)SVD-GAMP-SBL

**Initialization:** Initialize  $\tau_x^{(0)} > 0$  and  $x^{(0)}$ . Set  $s^{(-1)} = 0, t = 0, y' = \mathbf{U}^T y$ . Below,  $|\mathbf{A}|^2$  represents the componentwise magnitude square of  $\mathbf{A}$ .

**Repeat Until Converged**

$$\tau_p^{(t)} = |\Lambda \mathbf{V}^T|^2 \tau_x^{(t)} \left( \xrightarrow{M, N \rightarrow \infty} \frac{\mathbf{1}^T \tau_x^{(t)}}{M} \Lambda^2 \mathbf{1}_M : \text{LSL-SVD-GAMP-SBL} \right)$$

$$\mathbf{p} = \Lambda \mathbf{V}^T \mathbf{x}^{(t)} - \tau_p \cdot * \mathbf{s}^{(t-1)}$$

$$\tau_s^{(t)} = \mathbf{1} ./ (\tau_p^{(t)} + \hat{\gamma}^{(t)-1} \mathbf{1})$$

$$\mathbf{s}^{(t)} = \tau_s^{(t)} \cdot * (\mathbf{y}' - \mathbf{p})$$

$$\mathbf{1} ./ \tau_q^{(t)} = |\mathbf{V} \Lambda|^2 \tau_s^{(t)} = \text{diag}(\mathbf{V} \Lambda^T \text{diag}(\tau_s^{(t)}) \Lambda \mathbf{V}^T) \mathbf{1} \left( \xrightarrow{M, N \rightarrow \infty} (1/\tau_q^{(t)}) \mathbf{1} : \text{LSL-SVD-GAMP-SBL} \right)$$

$$\mathbf{r}^{(t)} = \mathbf{x}^{(t)} + \tau_q^{(t)} \cdot * \mathbf{V} \Lambda^T \mathbf{s}^{(t)}$$

$$\tau_x^{(t+1)} = \tau_q^{(t)} \cdot * \mathbf{g}'_1(\mathbf{r}^{(t)}, \tau_q^{(t)})$$

$$\mathbf{x}^{(t+1)} = \mathbf{g}_1(\mathbf{r}^{(t)}, \tau_q^{(t)})$$

Hyperparameter estimation remains the same as in GVAMP-SBL.

- Intuition: After the unitary transformation with  $\mathbf{U}^T$ ,  $\mathbf{y}'$  is the observation and  $\Lambda \mathbf{V}^T$  plays the role of the measurement matrix in which  $\mathbf{V}$  can be treated as i.i.d. (approximately in the LSL).

Further, it can be verified that  $|\Lambda \mathbf{V}^T|^2 \tau_x^{(t)} = \text{diag}(\Lambda \mathbf{V}^T \text{diag}(\tau_x^{(t)}) \mathbf{V} \Lambda^T) \mathbf{1}$ . Using Corollary 1 in

<sup>53</sup>  $\mathbf{V}^T \text{diag}(\tau_x^{(t)}) \mathbf{V}$  converges a.s. to  $\frac{\mathbf{1}_M^T \tau_x^{(t)}}{M} \mathbf{I}_M$  for Haar  $\mathbf{V}$ .

<sup>53</sup>Takeuchi'20

## Complexity Issues with GVAMP-SBL

- In original VAMP, performing economy SVD of  $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$  and applying matrix inversion lemma,  $(\gamma\mathbf{A}^T\mathbf{A} + \tau_2^{-1}\mathbf{I})^{-1}\gamma\mathbf{A}^T\mathbf{y}$  reduces to  $\gamma\mathbf{V}(\gamma\mathbf{\Lambda}^2 + \tau_2^{-1}\mathbf{I})^{-1}\mathbf{\Lambda}\mathbf{U}^T\mathbf{y}$ . Hence this does not require any big matrix inversion. All the substantial computations reduce to **matrix-vector multiplications**. Note that all the precision matrices are multiples of identity in VAMP, so  $\tau_1\mathbf{I}$ ,  $\tau_2\mathbf{I}$  (due to scalar EP whereas it is diagonal EP in GVAMP-SBL).
- Matrix inverse operation in the computation of  $\mathbf{g}_2(\cdot, \cdot)$  or  $\mathbf{g}'_2(\cdot, \cdot)$  does not simplify w.r.t. LMMSE.
- We can use deterministic equivalents for  $((\hat{\gamma}\mathbf{V}^T\mathbf{\Lambda}^2\mathbf{V} + \text{Diag}(\tau_2^{(t)}))^{-1})_{i,i}$  resulting from the large system analysis for Haar (random unitary) matrices as in <sup>54</sup> (see next slide).
- Based on some different but related work in <sup>55</sup> that exploits the asymptotic freeness concept from free probability, to justify the approximation  $((\hat{\gamma}\mathbf{V}^T\mathbf{\Lambda}^2\mathbf{V} + \text{Diag}(\tau_2^{(t)}))^{-1})_{i,i} = (\delta + \tau_{2,i}^{(t)})^{-1}$ , for some scalar  $\delta$ , we can guess the expression for the desired deterministic equivalents.
- The same work appears to consider that the whole matrix  $(\hat{\gamma}\mathbf{V}^T\mathbf{\Lambda}^2\mathbf{V} + \text{Diag}(\tau_2^{(t)}))^{-1}$  can be considered as diagonal, which may be OK for the resulting error covariance but not for the LMMSE estimation operation.

<sup>54</sup>Couillet, Hoydis, Debbah'12

<sup>55</sup>Çakmak, Opper'18

# Large System Simplifications

## Lemma 6

Let  $\mathbf{P}$  be any Hermitian matrix with bounded spectral norm and let  $\mathbf{V} \in \mathcal{C}^{M \times N}$  be  $N < M$  columns of a Haar distributed (unitary) random matrix. Let  $\mathbf{A}$  be a nonnegative definite matrix with  $\|\mathbf{A}\| < \infty$  ( $\|\mathbf{A}\|$  represents the spectral norm) and  $\mathbf{D}$  be any diagonal matrix with positive entries. Then the following convergence result holds almost surely,

$$\frac{1}{M} \text{tr}\{\mathbf{A}(\mathbf{V}\mathbf{P}\mathbf{V}^T + \mathbf{D})^{-1}\} - \frac{1}{M} \text{tr}\{\mathbf{A}(\bar{e}\mathbf{I} + \mathbf{D})^{-1}\} \xrightarrow{\text{a.s.}} 0.$$

The scalar  $\bar{e}$  can be obtained as the unique solution (fixed point) of the following system of equations,

$$\begin{aligned} \bar{e} &= \frac{1}{M} \text{tr}\{\mathbf{P}(e\mathbf{P} + (1 - e\bar{e})\mathbf{I}_N)^{-1}\}, \\ e &= \frac{1}{M} \text{tr}\{\mathbf{A}(\bar{e}\mathbf{I}_M + \mathbf{D})^{-1}\}. \end{aligned}$$



# Simulations Setup

- To motivate further the posterior variance prediction analysis detailed in Theorem 5, we compare the posterior variances of each  $x_i$  for different approximate inference methods based on BP or MF in Figure 9.
- We compare SAVE and various AMP based algorithms which are robust to measurement matrices which are beyond i.i.d. UTAMP-SBL is the algorithm derived in <sup>56</sup> based on the SVD transformation of  $\mathbf{A}$  from GAMP.
- Legend "GAMP-SBL" corresponds to the algorithm in <sup>57</sup>.
- Dimensions of  $\mathbf{A}$ ,  $M = 1000$ ,  $N = 500$ . The power delay profile (variances of  $x_i$ ) for the SBL model is chosen as  $d^{i-1}$ , with  $d = 0.995$  and starting with index  $i = 1$ .

---

<sup>56</sup>Luo, Guo, Huang, Xi'19

<sup>57</sup>Shoukairi, Schniter, Rao'18

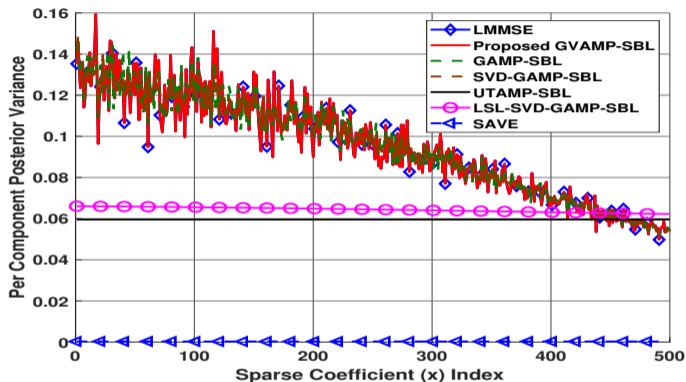
Per Component MSE under i.i.d.  $\mathbf{A}$  case

Figure 9: Per component MSE (posterior variance), i.i.d.  $\mathbf{A}$ .

### Key Points

- SAVE has such ridiculously low posterior variance, which clearly exhibits the MF suboptimality.

# Convergence Behaviour

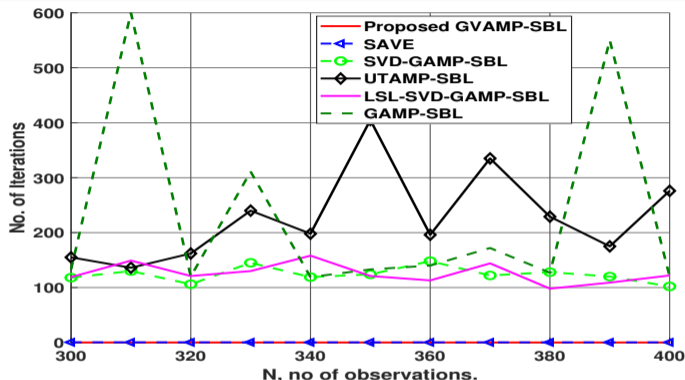


Figure 10: No of iterations to converge as a function of  $N$ .

## Key Points:

- The complexity of GVAMP-SBL is  $\mathcal{O}(4MN) \times T_{it1}$  and that of Algo 2 being  $\mathcal{O}(MN) \times T_{it2}$ , where  $T_{it1}, T_{it2}$  represents the number of iterations.
- It is clearly evident from Figure that GVAMP-SBL converges in very few iterations (less than 10) compared to Algo 2 which takes more than 100 iterations to converge.

# Partial Fourier Matrix Case

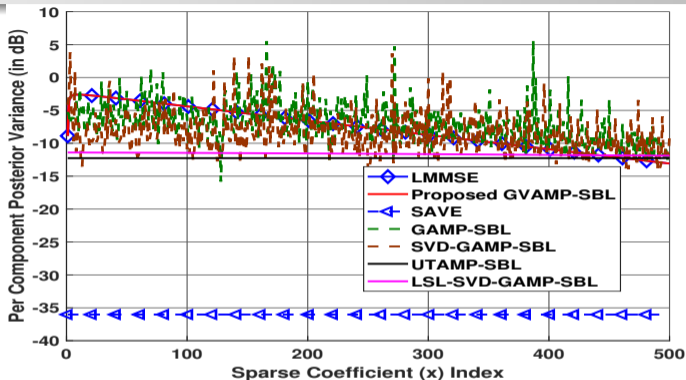


Figure 11: Per component MSE (posterior variance), Partial Fourier **A**.

## Key Points:

- In the case of partial Fourier measurements, we select  $\mathbf{A} = \mathbf{S}\mathbf{U}$ , where  $\mathbf{U}$  represents the  $M \times M$  discrete Fourier transform matrix (DFT) and  $\mathbf{S} \in 0, 1^{N \times M}$  is a subsampling matrix.
- In this case, we observe that the posterior variance of the proposed GVAMP-SBL converges exactly to the LMMSE estimator values. However, the SVD based GAMP-SBL versions are having convergence issues which lead to suboptimal performance.

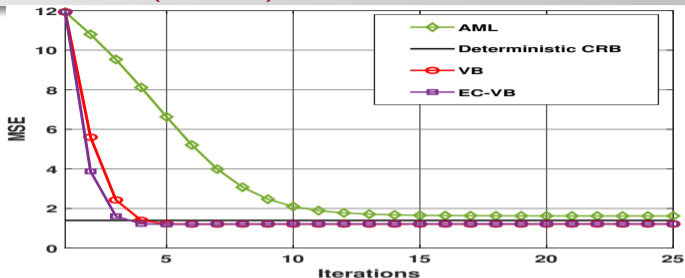
# Open Issues?

- Robustness of GVAMP-SBL to nonlinear measurement case, with applications:
  - Quantized compressed sensing (Finite Resolution ADCs in MaMIMO)
  - Binary linear classification
  - Phase retrieval
  - Robust regression (Case of outliers or non-AWGN noise)
- State evolution analysis for GVAMP-SBL

# Outline

- 1 Introduction
- 2 Static SBL
- 3 Combined BP-MF-EP Framework
- 4 Posterior Variance Prediction: Bayes Optimality
- 5 Performance Analysis of Approximate Inference Techniques (mCRB)**
- 6 Dynamic SBL
- 7 Kronecker Structured Dictionary Learning using BP/VB
- 8 Numerical Results and Conclusion

# Why Mismatched CRB (mCRB)?



- VB allows to attain lower MSE than the (deterministic) CRB. So, consider the Bayesian CRB?
- A Bayesian CRB is valid only if the (Gaussian) prior for  $x$  is the correct prior. VB converge to the most appropriate priors even if in fact the parameter  $x$  are deterministic! This requires mCRBs.
- mCRB corresponds to Laplace approximation<sup>58</sup> of MAP or VB.  
Evaluation of marginal likelihood or free energy using Laplace's method - a Gaussian approximation of the posterior  $q$  around a maximum a posteriori (MAP) estimate, motivated by the fact that in the asymptotic regime (large amount of data or high SNR), the posterior approaches a Gaussian around the MAP point<sup>59</sup>.

<sup>58</sup>Šmídl, Quinn'05

<sup>59</sup>Fortunati, Gini, Greco, Richmond'17

## Convergence Point for mCRB (for Laplace Approximation)

- Main message: with approximate posteriors in all variants of MP, the CRB needs to be replaced by mCRB. Bayesian mCRB in principle, for which the SotA is not yet fully developed.
- Mismatches: fictitious prior (for empirical Bayes, e.g. James-Stein) + replace the actual posterior  $p(\cdot)$  by  $q(\cdot)$ .
- Under a mismatched distribution model, the convergence point  $\bar{\theta}$  (also called as pseudo true parameter) is used to evaluate the effectiveness of the estimator, since no true parameter vector may exist under the assumed distribution  $q$ .
- The VB convergence point (of complete  $\theta$ ) is the MAP of  $E_p(\sum_i \ln(q_{\theta_i}(\theta_i)))$  (assuming large amount of data), so  $\ln$  of product of  $q$ 's = sum of  $\ln$  of  $q$ 's and converges to its expected value according to actual pdf  $p$  (LLN).

$$\begin{aligned}\bar{\theta}_i &= \arg \max_{\theta_i} E_{p(\mathbf{y}, \theta^0)} \ln q(\theta_i) \\ &= \arg \max_{\theta_i} E_{p(\mathbf{y}|\theta^0)} \ln \langle p(\mathbf{y}, \theta) \rangle_i.\end{aligned}$$

- The expectation over  $p(\theta)$  (being deterministic) disappears above.



## Outline of the mCRB Derivation

- We define  $\zeta = \hat{\theta} - \bar{\theta}$ ,  $\tilde{\theta} = \bar{\theta} - \theta^0$ ,  $\tilde{\theta} = \hat{\theta} - \theta^0 = \zeta + \tilde{\theta}$ . For any choice of score function  $\eta$  using a matrix generalization of the Cauchy Schwartz inequality [RichmondTSP2015, Kantor:15], the error correlation matrix can be written as,

$$\mathbf{mCRB} = \mathbf{R}_{\tilde{\theta}\tilde{\theta}} = E_p \tilde{\theta} \tilde{\theta}^H \geq \mathbf{R}_{\zeta\eta} \mathbf{R}_{\eta\eta}^{-1} \mathbf{R}_{\eta\zeta} + \tilde{\theta} \tilde{\theta}^H,$$

where  $\mathbf{R}_{\zeta\eta} = E(\zeta \eta^H)$  and  $\mathbf{R}_{\zeta\zeta} = E(\zeta \zeta^H)$ .

- The choice of the **score function**: it should be **zero mean and depends on the sufficient statistic for estimating  $\theta$** .
- We chose the score function:  $\eta = \frac{\partial}{\partial \theta^*} \ln q(\theta) |_{\bar{\theta}}$ .
- The Taylor series expansion of the data likelihood around  $\bar{\theta}$  is given by,

$$\log q(\mathbf{y}, \bar{\theta} + \Delta\theta) = \log q(\mathbf{y}, \bar{\theta}) + \Delta\theta^H \frac{\partial \log q(\mathbf{y}, \theta)}{\partial \theta^*} |_{\bar{\theta}} + \Delta\theta^H \frac{\partial^2 \log q(\mathbf{y}, \theta)}{\partial \theta^* \theta^T} |_{\bar{\theta}} \Delta\theta + o(\|\Delta\theta\|^2).$$

## Approximations in the Asymptotic Limit

- Equating the derivative w.r.t  $\Delta\theta^*$  to be zero yields an approximation of the error term  $\zeta$  as,

$$\zeta = -\left(\frac{\partial^2 \log q(\mathbf{y}, \theta)}{\partial \theta^* \theta^{*T}} \Big|_{\bar{\theta}}\right)^{-1} \frac{\partial \log q(\mathbf{y}, \theta)}{\partial \theta^*} \Big|_{\bar{\theta}}.$$

- We can replace the Hessian and  $\frac{\partial \log q(\mathbf{y}, \theta)}{\partial \theta^*}$  above by  $E_{p(\mathbf{y}|\theta)}\left(\frac{\partial^2 \log q(\mathbf{y}, \theta)}{\partial \theta^* \theta^{*T}}\right)$  and  $E_{p(\mathbf{y}|\theta)}\left(\frac{\partial \log q(\mathbf{y}, \theta)}{\partial \theta^*}\right)$ , respectively in the asymptotic limit.
- We arrive at,

$$\begin{aligned} E_{p(\mathbf{y}|\theta)} \frac{\partial^2 \log q(\mathbf{y}, \theta)}{\partial \theta^* \theta^{*T}} &= -\mathcal{V}^H \mathbf{Q} \mathcal{V}, \\ \mathbf{Q} &= \frac{1}{\sigma^2} \text{blkdiag}(0, \mathcal{V}_f^H \langle \mathcal{H}^H \mathcal{H} \rangle \mathcal{V}_f + \langle \alpha \rangle \mathbf{I}, \langle \mathcal{F}^H \mathcal{F} \rangle + \langle \beta \rangle \mathbf{I}), \\ E_{p(\mathbf{y}|\theta)} \frac{\partial \log q(\mathbf{y}, \theta)}{\partial \theta^* \theta^{*T}} \Big|_{\bar{\theta}} &= -\mathcal{V}^H \bar{\mathbf{Q}} \mathcal{V}. \end{aligned}$$

## Optimal Partitioning of BP/MF nodes

- mCRB refers to mismatched CRB (CRB under model misspecification)<sup>60</sup>.
- In the **combined BP/VB framework**, applied to joint **detection** and **parameter estimation**, traditionally **BP** is applied to the **detection** part, whereas the simpler **VB** is applied to the **parameter estimation** part.

**Theorem:** If the parameter partitioning in VB is such that the different parameter blocks are decoupled at the level of Fisher Information Matrix (FIM), 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.

$$mCRB_{BP} = \text{blkdiag}(CRB) = \text{blkdiag}(FIM^{-1}),$$

$$mCRB_{VB} = (\text{blkdiag}(FIM))^{-1},$$

So,

$$mCRB_{BP} = mCRB_{VB} \text{ if } FIM = \text{blkdiag}(FIM).$$

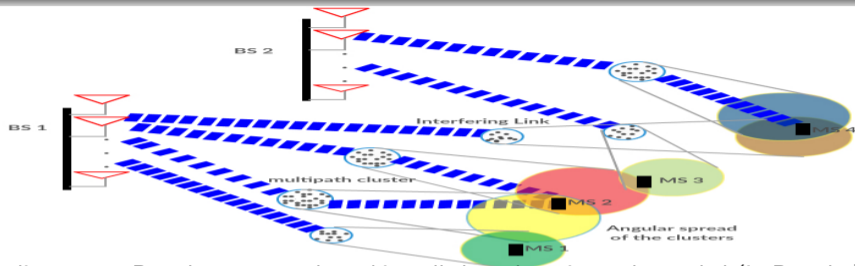
- **Hence:** BP may also improve parameter estimation.  
But loopy BP may not reach it's mCRB.

<sup>60</sup>Richmond,Horowitz'15

# Outline

- 1 Introduction
- 2 Static SBL
- 3 Combined BP-MF-EP Framework
- 4 Posterior Variance Prediction: Bayes Optimality
- 5 Performance Analysis of Approximate Inference Techniques (mCRB)
- 6 Dynamic SBL**
- 7 Kronecker Structured Dictionary Learning using BP/VB
- 8 Numerical Results and Conclusion

# Tensor Representation (Channel Tracking in MaMIMO OFDM)



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

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

- 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.

# Time Varying Sparse State Tracking

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

The diagram illustrates the AR(1) process for sparse state tracking. It consists of two main equations:

Top equation:  $\mathbf{y}_t = \mathbf{A}_t \mathbf{x}_t + \mathbf{v}_t$

- $\mathbf{y}_t$  is an  $N \times 1$  vector.
- $\mathbf{A}_t$  is an  $N \times M$  matrix, where  $N \ll M$ .
- $\mathbf{x}_t$  is an  $M \times 1$  vector.
- $\mathbf{v}_t$  is an  $N \times 1$  vector.

Bottom equation:  $\mathbf{x}_t = \mathbf{F} \mathbf{x}_{t-1} + \mathbf{w}_t$

- $\mathbf{x}_t$  is an  $M \times 1$  vector.
- $\mathbf{F}$  is an  $M \times M$  matrix.
- $\mathbf{x}_{t-1}$  is an  $M \times 1$  vector.
- $\mathbf{w}_t$  is an  $M \times 1$  vector.

Define:  $\Xi = \text{diag}(\xi)$ ,  $\mathbf{F} = \text{diag}(f)$ .

$f_i$ : correlation coefficient and  $x_{i,t} \sim \mathcal{CN}(x_{i,t}; 0, \frac{1}{\xi_i})$ . Further,  $\mathbf{w}_t \sim \mathcal{CN}(\mathbf{w}_t; \mathbf{0}, \Gamma^{-1} = \Xi^{-1}(\mathbf{I} - \mathbf{F}\mathbf{F}^H))$

and  $\mathbf{v}_t \sim \mathcal{CN}(\mathbf{v}_t; \mathbf{0}, \gamma^{-1}\mathbf{I})$ . VB leads to Gaussian SAVE-Kalman Filtering (GS-KF).

Joint Dictionary Learning and Sparse Excitation Tracking.

## 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 <sup>61</sup>.
- $y_n$  – 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  $\hat{x}_{n,l}^{(t)}$  and variance  $\nu_{n,l}^{(t)}$ .
- The prediction about  $x_t$  can be computed from the time update equation of the standard Kalman filter, Here we denote  $\hat{f}_{k|t-1}$  as the estimate of  $f_k$  given the observations till  $t - 1$  and  $\tilde{f}_{k|t-1}$  represents the error in the estimation. Similary we can represent  $x_{k,t-1} = \hat{x}_{k,t-1|t-1} + \tilde{x}_{k,t-1|t-1}$ ,  $\tilde{x}_{k,t-1|t-1}$  being the estimation error.

$$\begin{aligned} \hat{x}_{k,t|t-1} &= \hat{f}_{k|t-1} \hat{x}_{k,t-1|t-1}, \quad \tilde{x}_{k,t|t-1} = \hat{f}_{k|t-1} \tilde{x}_{k,t-1|t-1} + \tilde{f}_{k|t-1} x_{k,t-1} + w_{k,t}, \\ \implies \sigma_{k,t|t-1}^2 &\stackrel{(a)}{=} |\hat{f}_{k|t-1}|^2 \sigma_{k,t-1|t-1}^2 + \sigma_{\tilde{f}_k}^2 (|\hat{x}_{k,t-1|t-1}|^2 + \sigma_{k,t-1|t-1}^2) + \frac{1}{\hat{\gamma}_{k|t-1}}, \end{aligned}$$

<sup>61</sup>Thomas,Slock'asilo19DynamicSBL

# Filtering Stage

- Measurement Update (Filtering) Stage:** For the measurement update stage, the posterior for  $\mathbf{x}_t$  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  $d_{l,t} = \left( \sum_{n=1}^N \nu_{n,l}^{(t)-1} \right)^{-1}$ ,  $r_{l,t} = d_{l,t} \left( \sum_{n=1}^N \frac{\hat{x}_{n,l}^{(t)}}{\nu_{n,l}^{(t)}} + \frac{\hat{x}_{l,t|t-1}}{\sigma_{l,t|t-1}^2} \right)$ .

$$\sigma_{l,t|t}^{-2} = \xi_{l,t} + d_{l,t}^{-1}, \quad \hat{x}_{l,t|t} = \frac{r_{l,t}}{1 + d_{l,t} \sigma_{l,t|t}^{-2}}.$$

- With the hard constraints, the equivalent observation model can be written as,

$$y_n - \sum_{l' \neq l}^M A_{n,l'} \hat{x}_{l',n} = A_{n,l} x_l + \sum_{l' \neq l}^M A_{n,l'} \tilde{x}_{l',n} + v_n, \text{ where,} \\ \tilde{x}_{l',n} \sim \mathcal{CN}(\tilde{x}_l; \mathbf{0}, \nu_{l',n}), \text{ and } m_{f_{\delta_n \rightarrow x_l}} \propto \mathcal{CN}(x_l; \hat{x}_{n,l}, \nu_{n,l}),$$



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



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

- We show in Lemma 1<sup>62</sup> that **KF is not enough to adapt the hyperparameters, instead we need at least a lag 1 smoothing** (i.e. the computation of  $\hat{\mathbf{x}}_{k,t-1|t}, \sigma_{k,t-1|t}^2$  through BP). For the smoothing stage, we use BP.
- Gaussian Posterior for  $\mathbf{x}_t$ :

$$\begin{aligned} \sigma_{k,t-1|t}^{-2,(i)} &= (\hat{f}_{k|t}^2 + \sigma_{f_{k|t}}^2) \mathbf{A}_k^H(t) \tilde{\mathbf{R}}_t^{-1} \mathbf{A}_k(t) + \sigma_{k,t-1|t-1}^{-2}, \\ \langle \mathbf{x}_{k,t-1|t}^{(i)} \rangle &= \sigma_{k,t-1|t}^{2,(i)} (\hat{f}_{k|t}^H \mathbf{A}_k^H(t) \tilde{\mathbf{R}}_t^{-1} (\mathbf{y}_t - \mathbf{A}_k(t) \mathbf{F}_{k|t}) \langle \mathbf{x}_{k,t-1|t}^{(i-1)} \rangle + \frac{\hat{\mathbf{x}}_{k,t-1|t-1}}{\sigma_{k,t-1|t-1}^2}). \end{aligned}$$

- Applying the MF rule, the resulting Gaussian distribution for  $\mathbf{f}$  has mean,  $\sigma_{f_i|t}^{-2}$  and variance,  $\hat{\mathbf{f}}_{i|t}$ .

$$\begin{aligned} \sigma_{f_i|t}^{-2} &= (|\hat{\mathbf{x}}_{i,t-1|t}|^2 + \sigma_{i,t-1|t}^2) \mathbf{A}_i^T(t) \tilde{\mathbf{R}}_t^{-1} \mathbf{A}_i(t), \\ \hat{\mathbf{f}}_{i|t} &= \sigma_{f_i|t}^2 \hat{\mathbf{x}}_{i,t-1|t}^H \mathbf{A}_i^H(t) \tilde{\mathbf{R}}_t^{-1} (\mathbf{y}_t - \mathbf{A}_i(t) \hat{\mathbf{F}}_{i|t} \hat{\mathbf{x}}_{i,t-1|t}). \end{aligned}$$

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

<sup>62</sup>Thomas,Slock'asilo19DynamicSBL

# Combined BP-MF-EP DAR SBL

Initialization  $\hat{\mathbf{f}}_{l|0}, \hat{\gamma}_{l|0} = \frac{a}{b}, \hat{\gamma}_0 = \frac{c}{d}, \hat{\mathbf{x}}_{l,0|0} = 0, \sigma_{l,0|0}^2 = 0, \forall l$ . Define  $\Sigma_{t-1|t-1} = \text{diag}(\sigma_{l,t|t-1}^2)$ .  
for  $t = 1 : T$  do

**Prediction Stage (Estimation of  $\mathbf{x}_t$  from  $\mathbf{Y}_{t-1}$ ):**

- Compute  $\hat{\mathbf{x}}_{l,t|t-1}, \sigma_{l,t|t-1}^2$  using EP.

**Filtering Stage (BP for  $\hat{\mathbf{x}}_{l,t|t}, \sigma_{l,t|t}^2$ ):** Repeat until convergence

- Compute  $\hat{\mathbf{x}}_{n,l}^{(t)}, \nu_{n,l}^{(t)}$  and update  $\hat{\mathbf{x}}_{l,t|t}, \sigma_{l,t|t}^2$ . • Compute  $\nu_{l,n}^{(t)}, \hat{\mathbf{x}}_{l,n}^{(t)}$ .

**Smoothing Stage (Estimation of  $\mathbf{x}_t$  from  $\mathbf{Y}_{t+1}$ ):**

**Initialization:**  $\Sigma_{t-1|t}^{(0)} = \Sigma_{t-1|t-1}, \hat{\mathbf{x}}_{t-1|t}^{(0)} = \hat{\mathbf{x}}_{t-1|t-1}$ . Define

$$\mathbf{B}^{(t)} = \mathbf{F}^T \mathbf{A}^{(t)T} \tilde{\mathbf{R}}_t^{-1} \mathbf{A}^{(t)} \mathbf{F} + \Sigma_{t-1|t-1}, \mathbf{h}_t = \mathbf{F}^T \mathbf{A}^{(t)T} \tilde{\mathbf{R}}_t^{-1} \mathbf{y}_t, \tilde{\mathbf{R}}_t = \mathbf{A}^{(t)} \mathbf{\Gamma}^{-1} \mathbf{A}^{(t)H} + \frac{1}{\gamma} \mathbf{I}.$$

- $P_{i,j} = \frac{-B_{i,j}^{(t)}}{B_{i,j}^{(t)} + \sum_{k \in \mathcal{N}(i) \setminus j} P_{k,i}}, \mu_{i,j} = (h_{i,t} + \sum_{k \in \mathcal{N}(i) \setminus j} P_{k,i} \mu_{k,i}), \forall i, j$ .
- $\sigma_{i,t-1|t}^2 = B_{i,j}^{(t)} + \sum_{k \in \mathcal{N}(i)} P_{k,i}, \hat{\mathbf{x}}_{i,t-1|t} = \sigma_{i,t-1|t}^2 (h_{i,t} + \sum_{k \in \mathcal{N}(i)} P_{k,i} \mu_{k,i})$
- $\Sigma_{t-1|t}^{-(i)} = (\hat{\mathbf{F}}_t^H \mathbf{A}^{(t)H} \tilde{\mathbf{R}}_t^{-1} \mathbf{A}^{(t)} \hat{\mathbf{F}}_t + \text{diag}(\mathbf{A}^{(t)H} \tilde{\mathbf{R}}_t^{-1} \mathbf{A}^{(t)}) \Sigma_{F|t} + \Sigma_{t-1|t}^{-(i-1)})$ .
- $\hat{\mathbf{x}}_{t-1|t}^{(i)} = \hat{\Sigma}_{t-1|t}^{(i)} (\hat{\Sigma}_{t-1|t-1}^{-1} \hat{\mathbf{x}}_{t-1|t}^{(i-1)} + \hat{\mathbf{F}}^H \mathbf{A}^{(t)H} \tilde{\mathbf{R}}_t^{-1} \mathbf{y}_t)$ .

**Estimation of hyperparameters (Define:  $x'_{k,t} = x_{k,t} - f_k x_{k,t-1}, \zeta_t = \beta \zeta_{t-1} + (1 - \beta) \langle \|\mathbf{y}_t - \mathbf{A}^{(t)} \mathbf{x}_t\|^2 \rangle$ ):**

- Compute  $\hat{f}_{l|t}, \sigma_{\hat{f}_{l|t}}^2$  using MF rule,  $\hat{\gamma}_t = \frac{c+N}{(\zeta_t+d)}$  and  $\gamma_{l|t} = \frac{(a+1)}{\langle |x'_{k,t}|^2 \rangle_{|t+b}}$ .

# Identifiability

- Non-singularity of FIM  $\implies$  local identifiability.
- Lemma:** The AR(1) model parameters require (at least lag 1) smoothing for identifiability.

For the AR(1) parameters, we obtain the FIM submatrix

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

FIM recursions show that filtering may be enough for the estimation of AR(1) parameters. However, estimation of  $\mathbf{f}$  by MF shows that we need the true value  $\mathbf{f}$  to get  $\hat{\mathbf{f}}$ .

- $p(\mathbf{f}/\mathbf{x}_t, \mathbf{y}_t) = p(\mathbf{f}/\mathbf{x}_t)$ . This suggests that posterior of  $\mathbf{f}$  given  $\mathbf{x}_t$  does not depend on  $\mathbf{y}_t$  or in other words the observations doesn't provide any extra information about  $\mathbf{f}$  other than the prior  $p(\mathbf{f}/\mathbf{x}_t)$  and hence  $\mathbf{f}$  is **globally not identifiable** <sup>63</sup>.

<sup>63</sup>Gelfand,Sahu'99

# Numerical Result-DAR SBL

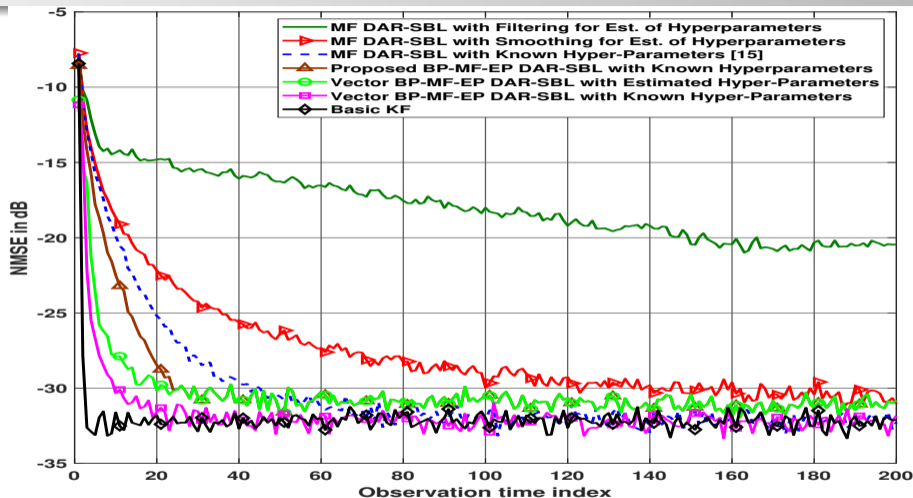


Figure 12: DAR-SBL: NMSE as a function of time.

# Mysteries Remaining

- The  $mCRB$  analysis indicates that the  $\mathbf{x}$  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 be reflected in the observation that the hyperparameters could be estimated (in what corresponds to filtering) by Type-II Maximum Likelihood (ML)<sup>64</sup> (ie ML for hyperparameters, with the random parameters  $\mathbf{x}$  integrated out).
- Characterization of local and global identifiability for a mix of Bayesian and Deterministic parameters.
- Fast version of type-II ML for dynamic AR-SBL.

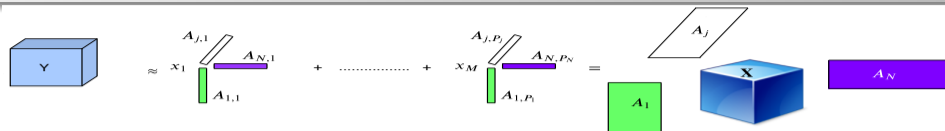
---

<sup>64</sup>Giri, Rao'16

# Outline

- 1 Introduction
- 2 Static SBL
- 3 Combined BP-MF-EP Framework
- 4 Posterior Variance Prediction: Bayes Optimality
- 5 Performance Analysis of Approximate Inference Techniques (mCRB)
- 6 Dynamic SBL
- 7 Kronecker Structured Dictionary Learning using BP/VB**
- 8 Numerical Results and Conclusion

# Kronecker Structured Tensor Models



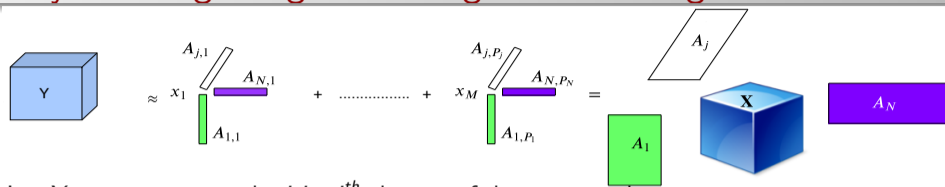
- Tensor signals appear in many applications: massive multi-input multi-output (MIMO) radar, massive MIMO (MaMIMO) channel estimation, speech processing, image and video processing.
- Exploiting tensorial structure beneficial compared to estimating unstructured dictionary. From Canonical Polyadic Decompositions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

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

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

$\mathbf{Y}_t \in \mathcal{C}^{l_1 \times l_2 \times \dots \times l_N}$  is the observations or data at time  $t$ ,  $\mathbf{A}_{j,i}(t) \in \mathcal{C}^{l_j}$ , the factor matrix  $\mathbf{A}_j(t) = [\mathbf{A}_{j,1}(t), \dots, \mathbf{A}_{j,P_j}(t)]$  and the overall unknown parameters are  $\llbracket \mathbf{A}_1(t), \dots, \mathbf{A}_N(t); \mathbf{x}_t \rrbracket$ ,  $\mathbf{x}_t$  is  $M (= \prod_{j=1}^N P_j)$ -dimensional sparse center tensor and  $\mathbf{w}_t, \mathbf{v}_t$  are the state or measurement noise.

# Dictionary Learning using Tensor Signal Processing



- Let  $Y_{i_1, \dots, i_N}$  represents the  $i_1 i_2 \dots i_N^{\text{th}}$  element of the tensor and  $\mathbf{y} = [y_{1,1, \dots, 1}, y_{1,1, \dots, 2, \dots}, y_{1,1, 2, \dots, i_N}]^T$ , then it can be verified that [Sidiropoulos:TSP17],  $\mathbf{y}_t = (\mathbf{A}_1(t) \otimes \mathbf{A}_2(t) \dots \otimes \mathbf{A}_N(t)) \mathbf{x}_t + \mathbf{w}_t$ ,  $\mathbf{w} \sim \mathcal{N}(\mathbf{w}; 0, \gamma^{-1} \mathbf{I})$ ,  
Matrix Unfolding:  $\mathbf{Y}^{(n)} = \mathbf{A}_n(t) \mathbf{X}^{(n)} (\mathbf{A}_N(t) \otimes \dots \otimes \mathbf{A}_{n+1}(t) \otimes \mathbf{A}_{n-1}(t) \dots \otimes \mathbf{A}_1(t))^T$ .
- $\mathbf{A}_j(t)$  is of dimension,  $I_j \times P_j$  and the resulting Tensor is  $\mathcal{C}^{I_1 \times \dots \times I_N}$ .
- 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

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

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



## 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 <sup>65</sup>) ( $\text{tr}\{(\bigotimes_{k=N, k \neq j}^1 \langle \mathbf{A}_k^T(t) \mathbf{A}_k^*(t) \rangle \langle \mathbf{X}^{(j)T} \mathbf{X}^{(j)} \rangle)\mathbf{I}\}$ ),  $\implies$  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  $\mathbf{v}_f(\tau_i(t))$  and  $\mathbf{v}_f(\tau_j(t))$  may be correlated.
- The joint VB estimates (mean and covariance) can be obtained as

$$\begin{aligned} \mathbf{M}_j^T &= \widehat{\mathbf{A}}_{1,j}^T(t) = \langle \gamma \rangle \Psi_j^{-1} \mathbf{B}_j^T, \\ \Psi_j &= (\langle \gamma \rangle \langle \mathbf{X}^{(j)} (\bigotimes_{k=N, k \neq j}^1 \langle \mathbf{A}_k^T(t) \mathbf{A}_k^*(t) \rangle) \mathbf{X}^{(j)T} \rangle), \end{aligned} \quad (10)$$

where  $\mathbf{V}_j = \langle \mathbf{X}^{(j)} \rangle \langle (\bigotimes_{k=N, k \neq j}^1 \mathbf{A}_k(t))^T \rangle$  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 (10),  $\widehat{\mathbf{A}}_{j,i}^T(t) = \Psi_j^{-1} \mathbf{b}_{j,i}$ .  $\mathbf{b}_{j,i}$  represents the  $i^{\text{th}}$  column of  $\mathbf{B}_j^T$ .

<sup>65</sup>Thomas, Slock, ICASSP'19

# Optimal Partitioning of the Measurement Stage and KS DL

## Lemma 7

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  $\mathbf{A}_{j,i}(t)$  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 Lemma 1<sup>66</sup>, 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).

$$\mathbf{y}_t = \underbrace{\left(\sum_{r=1}^M x_{r,t} \mathbf{F}_r\right)}_{\mathbf{F}(\mathbf{x}_t)} \underbrace{\left(\otimes_{j=1}^N \Phi_{j,t}\right)}_{\mathbf{f}(\Phi_t)} + \mathbf{w}_t. \quad \mathbf{J}(\Phi_t) = [\mathbf{J}(\Phi_{1,t}) \dots \mathbf{J}(\Phi_{N,t})]$$

where,  $\mathbf{J}(\Phi_{j,t}) = \mathbf{F}(\mathbf{x}_t)(\Phi_{1,t} \otimes \dots \otimes \mathbf{I}_{l_j p_j} \otimes \dots \otimes \Phi_{N,t})$ ,

$$FIM = \begin{bmatrix} E(\gamma) \mathbf{J}(\Phi_t)^T \mathbf{J}(\Phi_t) & 0 & 0 & 0 \\ 0 & E(\gamma) \mathbf{J}(\mathbf{x}_t)^T \mathbf{J}(\mathbf{x}_t) + E(\Xi) & 0 & 0 \\ 0 & 0 & aE(\Xi) & 0 \\ 0 & 0 & 0 & (N+c-1)E(\gamma^{-2}) \end{bmatrix}$$

<sup>66</sup>Kalyan, Thomas, Slock'19

# Outline

- 1 Introduction
- 2 Static SBL
- 3 Combined BP-MF-EP Framework
- 4 Posterior Variance Prediction: Bayes Optimality
- 5 Performance Analysis of Approximate Inference Techniques (mCRB)
- 6 Dynamic SBL
- 7 Kronecker Structured Dictionary Learning using BP/VB
- 8 Numerical Results and Conclusion**

# BP-MF-EP Outperforms SAVED-KS DL

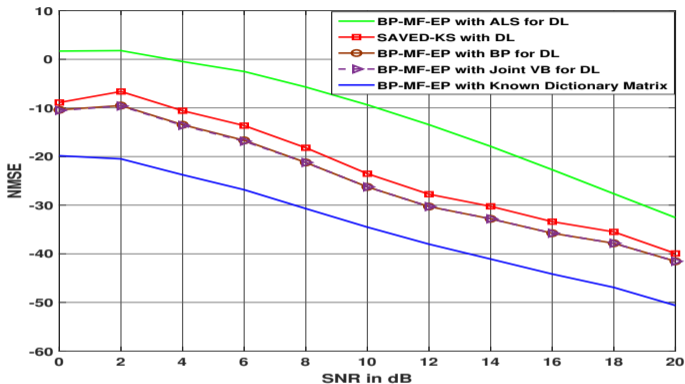


Figure 13: Static SBL: NMSE as a function of  $N$ .

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













# Conclusions and Thank You!

- Further advancements from <sup>67</sup>: 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.














---

<sup>67</sup>Thomas,Slock'asilo19ConvergenceAnalysisSBL














# References I

-  M. Bayati, A. Montanari, "The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing," in *IEEE Trans. on Info. Theo.*, Feb. 2011.
-  M. J. Beal, "Variational Algorithms for Approximate Bayesian Inference," in *Thesis, Univeristy of Cambridge, UK*, May 2003.
-  M. Borgerding, P. Schniter, S. Rangan, "AMP-Inspired Deep Networks for Sparse Linear Inverse Problems," *IEEE Trans. on Sig. Process.*, Aug. 2017.
-  B. Çakmak, M. Opper, "Expectation Propagation for Approximate Inference: Free Probability Framework," in *IEEE Inter. Sympo. On Info. Theo. (ISIT)*, 2018.
-  B. Çakmak, O. Winther, B. H. Fleury, "S-AMP: Approximate Message Passing for General Matrix Ensembles," in *IEEE Intl. Sympo. Info. Theo.*, 2014.
-  E. Candes, T. Tao, "The Dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ ," *The annals of Statistics*, 2007.
-  S. S. Chen, D. L. Donoho, M. A. Saunders, "Atomic decomposition by Basis Pursuit," *SIAM J. Sci. Comput.*, 1999.
-  R. Couillet, J. Hoydis, M. Debbah", "Random Beamforming over Quasi-Static and Fading Channels: A Deterministic Equivalent Approach," in *IEEE Trans. On Info. Theo.*, 2012.
-  I. Daubechies, M. Defrise, C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint.," *Comm. on Pure and Applied Mathematics*, 2004.
-  J. Du, S. Ma, Y-C. Wu, S. Kar, J. M. F. Moura, "Convergence Analysis of Distributed Inference with Vector-Valued Gaussian Belief Propagation," in *Jrnl. of Mach. Learn. Res.*, Apr. 2018.
-  H. Duan, L. Yang, J. Fang, H. Li, "Fast Inverse-Free Sparse Bayesian Learning via Relaxed Evidence Lower Bound Maximization," in *IEEE Sig. Process. Lett.*, Jun. 2017.
-  Y. C. Eldar "Rethinking Biased Estimation: Improving Maximum Likelihood and the CramérRao Bound," *Found. and Tren. in Sig. Process.*, 2008.

# References II
















-  S. Fortunati, F. Gini, M. S. Greco, C. D. Richmond, "Performance Bounds for Parameter Estimation under Misspecified Models: Fundamental Findings and Applications," in *IEEE Sig. Proc. Mag.*, Nov. 2017.
-  A. E. Gelfand, S. K. Sahu, "Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models," in *Journ. of the Americ. Stat. Assoc.*, Mar. 1999.
-  R. Giri, B. Rao, "Type I and Type II Bayesian Methods for Sparse Signal Recovery Using Scale Mixtures," in *IEEE Trans. Sig. Process.*, July 2016.
-  K. Gregor, Y. LeCun, "Learning Fast Approximations of Sparse Coding," *Intl. Conf. on Mach. Learn.*, 2010.
-  W. James, C. Stein "Estimation with quadratic loss," *Proc. 4th Berkeley Symp. Mathematical Statistics Probability*, 1961.
-  J. K. Johnson, D. Bickson, D. Dolev, "Fixing Convergence of Gaussian Belief Propagation," in *IEEE Intl. Symp. on Info. Theo.*, 2009.
-  M. Luo, Q. Guo, D. Huang, J. Xi, "Sparse Bayesian Learning using Approximate Message Passing with Unitary Transformation," in *IEEE VTS Asia Pac. Wire. Commun. Symp., APWCS*, Aug. 2019.
-  D. M. Malioutov, J. K. Johnson, A. S. Willsky, "Walk-Sums and Belief Propagation in Gaussian Graphical Models," in *Jrnl. of Mach. Learn. Res.*, Oct. 2006.
-  S. Mallat, Z. Zhang, "Matching Pursuits with time-frequency dictionaries.," *IEEE Trans. Sig. Process.*, 1993.
-  J. Ma, L. Ping "Orthogonal AMP," in *IEEE Access*, Mar. 2017.
-  T. P. Minka, "Expectation propagation for approximate Bayesian inference ," in *Proc. of Conf. on Uncert. in Art. Intell. (UAI)*, 2001.
-  B. Mu, T. Chen, L. Ljung, "On asymptotic properties of hyperparameter estimators for kernel-based regularization methods," in *Automatica*, May. 2018.
-  D. Needell, J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples.," *Applied and computational harmonic analysis* , 2009.

# References III

-  G. Pillonetto, F. Dinuzzo, T. Chen, G. D Nicolao, L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, Feb. 2014.
-  G. Pillonetto, G. D. Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, 2010.
-  S. Rangan, "Generalized Approximate Message Passing for Estimation with Random Linear Mixing," in *IEEE Intl. Sympo. Info. Theo.*, 2011.
-  S. Rangan, A. K. Fletcher, P. Schniter, U. S. Kamilov, "Inference for Generalized Linear Models via Alternating Directions and Bethe Free Energy Minimization," in *IEEE Trans. Inf. Theory*, Jan. 2017.
-  S. Rangan, P. Schniter, A. K. Fletcher, "Vector Approximate Message Passing," in *IEEE Trans. Inf. Theory*, Oct. 2019.
-  C. D. Richmond and L. L. Horowitz, "Parameter bounds on estimation accuracy under model misspecification," in *IEEE Trans. on Sig. Process.*, May. 2015.
-  E. Riegler, G. E. Kirkelund, C. N. Manchón, B. H. Fleury, "Merging Belief Propagation and the Mean Field Approximation: a Free Energy Approach," in *IEEE Trans. on Info. Theo.*, Jan. 2013.
-  P. Rusmevichientong, B. Van Roy, "An analysis of belief propagation on the turbo decoding graph with Gaussian densities," *IEEE Trans. Inf. Theory*, 2001.
-  M. Al-Shoukairi, P. Schniter, B. D. Rao, "GAMP-Based Low Complexity Sparse Bayesian Learning Algorithm," in *IEEE Trans. on Sig. Process.*, Jan. 2018.
-  D. Shutin, T. Buchgraber, S. R. Kulkarni, H. V. Poor, "Fast Variational Sparse Bayesian Learning With Automatic Relevance Determination for Superimposed Signals," in *IEEE Trans. Sig. Process.*, Dec. 2011.
-  V. Šmídl, A. Quinn "The Variational Bayes Method in Signal Processing," in *Springer Series on Sig. and Comm. Tech.*, 2005.
-  K. Takeuchi, "Rigorous Dynamics of Expectation- Propagation-Based Signal Recovery from Unitarily Invariant Measurements," in *IEEE Trans. Inf. Theory*, Jan. 2020.
-  X. Tan, J. Li, "Computationally Efficient Sparse Bayesian Learning via Belief Propagation," in *IEEE Trans. on Sig. Proc.*, Apr. 2010.



## References IV

-  C. K. Thomas, K. Gopala, D. Slock, "Sparse Bayesian learning for a bilinear calibration model and mismatched CRB," in *EUSIPCO*, 2019.
-  C. K. Thomas, D. Slock, "SAVE - space alternating variational estimation for sparse Bayesian learning," in *IEEE Data Sci. Wkshp.*, Jun. 2018.
-  C. K. Thomas, D. Slock, "Space Alternating Variational Estimation and Kronecker Structured Dictionary Learning," in *IEEE ICASSP*, 2019.
-  C. K. Thomas and D. Slock, "Convergence Analysis Of Sparse Bayesian Learning under Approximate Inference Techniques," in *Asilomar Conf. on Sig., Sys., and Comp.*, Nov. 2019.
-  C. K. Thomas and D. Slock, "Low Complexity Static and Dynamic Sparse Bayesian Learning Combining BP, VB and EP Message Passing," in *IEEE Asilomar*, Nov. 2019.
-  R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Roy. Statist. Soc. Series B (Methodol.)*, 1996.
-  M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learn. Res.*, 2001.
-  M. E. Tipping, A. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *AISTATS*, 2003.
-  J. A. Tropp, A. C. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Trans. Info. Theo.*, 2007.
-  D. G. Tzikas, A. C. Likas, N. P. Galatsanos, "The variational approximation for Bayesian inference," in *IEEE Sig. Process. Mag.*, Nov. 2008.
-  S. Wagner, R. Couillet, M. Debbah, D. Slock, "Large System Analysis of Linear Precoding in MISO Broadcast Channels with Limited Feedback," in *IEEE Trans. Inf. Theory*, July. 2012.
-  Y. Weiss, W. T. Freeman, "Correctness of belief propagation in Gaussian graphical models of arbitrary topology," *NIPS*, 2000.
-  D. P. Wipf, B. D. Rao, "Sparse Bayesian Learning for Basis Selection," *IEEE Trans. Sig. Process.*, Aug 2004.
-  J. S. Yedidia, W. T. Freeman, Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," in *IEEE Trans. on Info. Theo.*, Jun. 2005.
-  D. Zachariah, P. Stoica, "Online Hyperparameter-Free Sparse Estimation Method," in *IEEE Trans. on Sig. Proc.*, July. 2015.