Low Complexity Static and Dynamic Sparse Bayesian Learning Combining BP, VB and EP Message Passing

Christo Kurisummoottil Thomas, Dirk Slock

kurisumm@eurecom.fr, slock@eurecom.fr

State of the Art

Why Combined BP/MF/EP ?

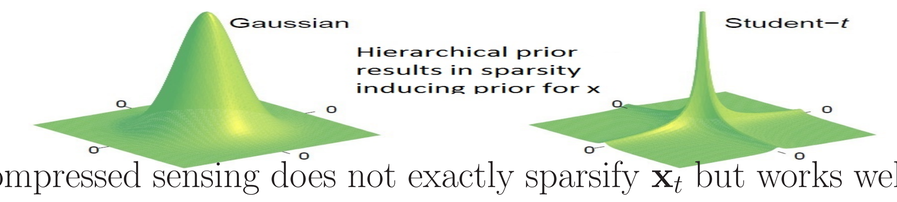
- Dynamic autoregressive SBL (DAR-SBL): a case of joint Kalman filtering (KF) with a linear time-invariant diagonal state-space model, and parameter estimation, which can be considered an instance of nonlinear filering.
- The authors in [RieglerTIT2013] propose a message passing (MP) approach for inferring the posteriors combining belief propagation (BP) and mean field (MF) approximations. As usual, they apply BP to detection (discrete parameters) only and MF to all estimation (continuous parameters).
- The advantages of the MF approach are that it always admits a convergent implementation while BP yields a good approximation of the posterior marginals if the factor graph has no cycles.
- Expectation Propagation (EP)-Gaussian approx., intractable distributions.

Sparse Bayesian Learning (SBL)

- Dynamic SBL (AR(1) state space model):

  State Update: $\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t,\ \mathbf{w}_t \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Lambda}^{-1})$

  Observation: $\mathbf{y}_t = \mathbf{A}^{(t)}\mathbf{x}_t + \mathbf{v}_t,\ \ \mathbf{v}_t \sim \mathcal{CN}(\mathbf{0}, \frac{1}{\gamma}\mathbf{I})$.



- SBL for compressed sensing does not exactly sparsify $\mathbf{x}_t$ but works well in the case of relatively limited date (eg underdetermined) in which case estimation emphasis is given to large unknowns while small unknowns get mostly ignored.
- Gamma prior to the precision of the state $\mathbf{x}_t$ and the innovation sequences $\mathbf{w}_t$ (both with same support), allowing to sparsify the components of $\mathbf{x}_t$.

  $\mathbf{\Lambda}^{-1} = \mathbf{\Gamma}(\mathbf{I} - \mathbf{F}\mathbf{F}^H) = \text{diag}\,(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_M}),\ \ \boldsymbol{\theta} = \{\mathbf{x}, \mathbf{\Lambda}, \gamma, \mathbf{F}\}$

Combined BP and MF Approximation

The fixed points of the BP algorithm are shown to be the stationary points of the Bethe free energy (BFE) [RieglerTIT2013]. However, for the MF approximation in variational Bayes (VB) [BealThesis2003], the approximate posteriors are shown to be converging to a local minimum of the MF free energy which is an approximation of the BFE.

$$F_{BP,MF} = \sum_{a\in\mathcal{A}_{BP}} \sum_{\boldsymbol{\theta}_a} q_a(\boldsymbol{\theta}_a) \ln \tfrac{q_a(\boldsymbol{\theta}_a)}{f_a(\boldsymbol{\theta}_a)} - \sum_{a\in\mathcal{A}_{MF}} \sum_{\mathbf{x}_a} \prod_{i\in\mathcal{N}(a)}$$
$$q_i(\theta_i) \ln f_a(\boldsymbol{\theta}_a) - \sum_{i\in\mathcal{I}} (|\mathcal{N}_{BP}(i)| - 1) \sum_{\theta_i} q_i(\theta_i) \ln q_i(\theta_i).$$
$$\text{Constraints: } \sum_{\theta_i} q_i(\theta_i) = 1,\ \forall i \in \mathcal{I}_{MF} \setminus \mathcal{I}_{BP},\ \sum_{\boldsymbol{\theta}_a} q_a(\boldsymbol{\theta}_a) = 1,$$
$$\forall a \in \mathcal{A}_{BP}, q_i(\theta_i) = \sum_{\boldsymbol{\theta}_a \setminus \theta_i} q_a(\boldsymbol{\theta}_a),\ \forall a \in \mathcal{A}_{BP}, i \in \mathcal{N}(a).$$

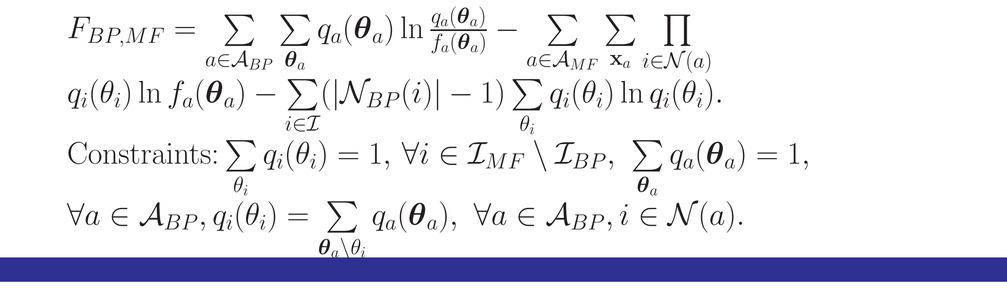
Message Passing Expressions

The fixed point equations corresponding to the constrained optimization of BFE can be written as follows [RieglerTIT2013], (a-factor nodes, i-variable nodes)

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

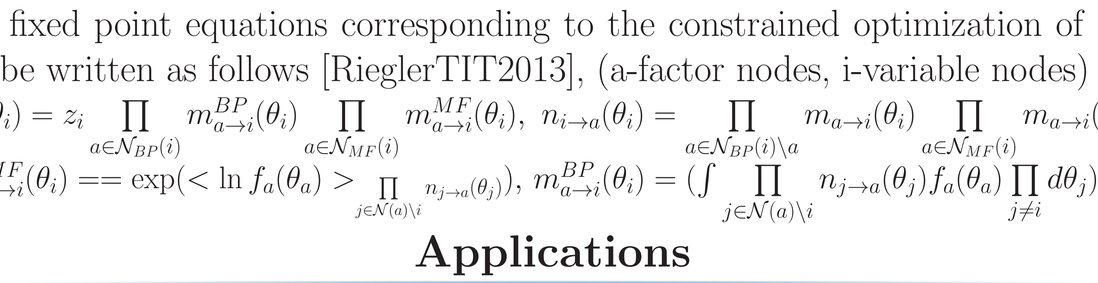
Applications

- **Massive MIMO channel estimation:**



- **Bayesian adaptive filtering.**

Factor Graph for DAR-SBL



All the messages (beliefs or continuous pdfs) passed between the nodes in the above factor graph can be shown to be Gaussian [TanLiTIT2013], parameterized by the mean and variance of the beliefs. With the hard constraints, the equivalent observation model:

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

Novel Contributions

Combined BP-MF-EP DAR SBL

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

Prediction Stage:

- Compute $\widehat{x}_{l,t|t-1}, \sigma^2_{l,t|t-1}$ using EP.

Filtering Stage (BP for $\widehat{x}_{l,t|t}, \sigma^{-2}_{l,t|t}$):

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

Smoothing Stage:

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

- $P_{i,j} = \frac{-B^{(t)2}_{i,j}}{B^{(t)}_{i,i} + \sum_{k\in\mathcal{N}(i)\setminus j} P_{k,i}}$, $\mu_{i,j} = (h_{i,t} + \sum_{k\in\mathcal{N}(i)\setminus j} P_{k,i}\mu_{k,i}), \forall i, j$.
- $\sigma^{-2}_{i,t-1|t} = B^{(t)}_{i,i} + \sum_{k\in\mathcal{N}(i)} P_{k,i}$, $\widehat{x}_{i,t-1|t} = \sigma^2_{i,t-1|t}(h_{i,t} + \sum_{k\in\mathcal{N}(i)} P_{k,i}\mu_{k,i})$
- $\mathbf{\Sigma}^{-(i)}_{t-1|t} = (\widehat{\mathbf{F}}^H_{|t}\mathbf{A}^{(t)\,H}\widetilde{\mathbf{R}}_t^{-1}\mathbf{A}^{(t)}\widehat{\mathbf{F}}_{|t} + \text{diag}\,(\mathbf{A}^{(t)\,H}\widetilde{\mathbf{R}}_t^{-1}\mathbf{A}^{(t)})\mathbf{\Sigma}_{\mathbf{F}|t} + \mathbf{\Sigma}^{-(i-1)}_{t-1|t}),\ \widetilde{\mathbf{R}}_t = \mathbf{A}^{(t)}\mathbf{\Lambda}^{-1}\mathbf{A}^{(t)\,H} + \frac{1}{\gamma}\mathbf{I}$.
- $\widehat{\mathbf{x}}^{(i)}_{t-1|t} = \widehat{\mathbf{\Sigma}}^{(i)}_{t-1|t}(\widehat{\mathbf{\Sigma}}^{-1}_{t-1|t-1}\widehat{\mathbf{x}}^{(i-1)}_{t-1|t} + \widehat{\mathbf{F}}^H\mathbf{A}^{(t)\,H}\widetilde{\mathbf{R}}_t^{-1}\mathbf{y}_t)$.

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

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

Optimal Partitioning of BP/MF nodes

mCRB refers to mismatched CRB (CRB under model misspecification) [RichmondTSP15].

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.

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

Hence: BP may also improve parameter estimation.

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 $\mathbf{J}_t = \mathbf{D} - \mathbf{D}(\mathbf{D} + \mathbf{J}_{t-1})^{-1}\mathbf{D}, if\, \mathbf{J}_{-1} = 0,$ then $\mathbf{J}_t = 0, \forall t \geq 0$. For $\mathbf{x}_t$, $\mathbf{J}_{\mathbf{x},t} = \mathbf{\Lambda} + \gamma\mathbf{A}^{(t)\,H}\mathbf{A}^{(t)} + \mathbf{\Lambda}\mathbf{F}(\mathbf{F}\mathbf{\Lambda}\mathbf{F}^H + \mathbf{J}_{x,t-1})^{-1}\mathbf{\Lambda}\mathbf{F}^H$, diagonal $\mathbf{J}_{\mathbf{x},t}$ $\implies$ MF is sufficient for $\mathbf{x}_t$ at prediction stage.

Smoothing

$\mathbf{y}_t = \mathbf{A}^{(t)}\mathbf{F}\mathbf{x}_{t-1} + \widetilde{\mathbf{v}}_t$, where $\widetilde{\mathbf{v}}_t = \mathbf{A}^{(t)}\mathbf{w}_{t-1} + \mathbf{v}_t$, Define $\mathbf{D} = (\mathbf{I} - \mathbf{F}\mathbf{F}^H)^{-1}$,

$$\mathbf{J}_{\mathbf{F},t} = \mathbf{\Gamma}\,\text{diag}\,(\mathbf{A}^{(t)\,H}\widetilde{\mathbf{R}}_t^{-1}\mathbf{A}^{(t)}) + \mathbf{D} - \mathbf{D}(\mathbf{D} + \mathbf{J}_{\mathbf{F},t})^{-1}\mathbf{D},\ \ \mathbf{J}_{p,t} = \begin{bmatrix} \mathbf{J}_{\mathbf{\Lambda},t} & \mathbf{J}_{\mathbf{\Lambda}\gamma,t} \\ \mathbf{J}_{\mathbf{\Lambda}\gamma,t} & \mathbf{J}_{\gamma\gamma} \end{bmatrix}$$

We obtain $\mathbf{J}_{\mathbf{x},t} = \mathbf{F}^T\mathbf{A}^{(t)\,H}\widetilde{\mathbf{R}}_t^{-1}\mathbf{A}^{(t)}\mathbf{F} + \mathbf{\Lambda} - \mathbf{\Lambda}\mathbf{F}(\mathbf{F}\mathbf{\Lambda}\mathbf{F}^H + \mathbf{J}_{\mathbf{x},t-1})^{-1}\mathbf{\Lambda}\mathbf{F}^H$, which is a full matrix.

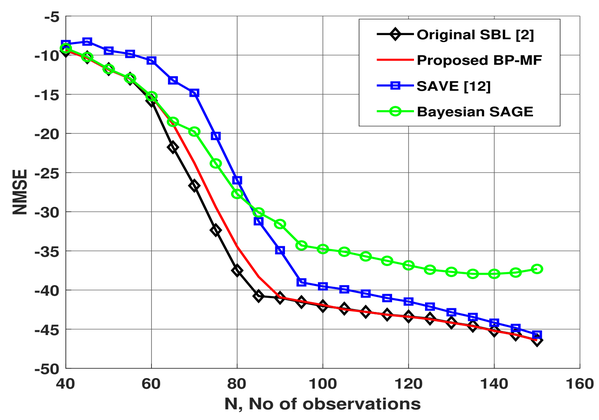
Conclusions:

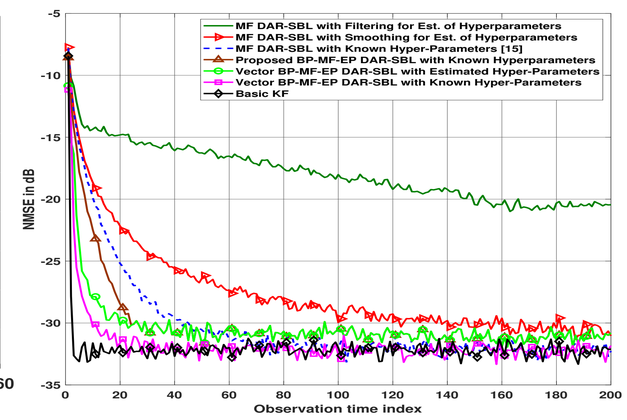
- BP required for smoothing of $\mathbf{x}_t$, $\mathbf{x}_{t-1|t}$.
- Lag 1 smoothing is sufficient for AR(1) parameters.

Corollary

- For the smoothing stage, an optimal partitioning is to apply BP for estimation of the state vector $\mathbf{x}$, $\widehat{\mathbf{x}}_{t-1|t}$ and MF for $\mathbf{F}$.

Numerical Results





Static SBL DAR-SBL

Conclusion and Future Work

- Introduced a fast SBL algorithm called BP-MF-EP DAR-SBL, which uses a combination of BP, VB and EP techniques to better approximate the posteriors of the sparse state $\mathbf{x}$ and parameters and track a time varying $\mathbf{x}$.
- Extension to dictionary learning ongoing.