Generalized Swept Approximate Message Passing based Kalman Filtering for Dynamic Sparse Bayesian Learning

Christo Kurisummoottil Thomas, Dirk Slock,
EURECOM, Sophia-Antipolis, France, Email: {kurisumm,slock}@eurecom.fr

***Abstract*—Sparse Bayesian Learning (SBL), initially proposed in the Machine Learning (ML) literature, is an efficient and well-studied framework for sparse signal recovery. SBL uses hierarchical Bayes with a decorrelated Gaussian prior in which the variance profile is also to be estimated. This is more sparsity inducing than e.g. a Laplacian prior. However, SBL does not scale with problem dimensions due to the computational complexity associated with the matrix inversion in Linear Mimimum Mean Squared Error (LMMSE) estimation. To address this issue, various low complexity approximate Bayesian inference techniques have been introduced for the LMMSE component, including Variational Bayesian (VB) inference, Space Alternating Variational Estimation (SAVE) or Message Passing (MP) algorithms such as Belief Propagation (BP) or Expectation Propagation (EP) or Approximate MP (AMP). These algorithms may converge to the correct LMMSE estimate. However, in ML we are often also interested in having posterior variance information. We observed that SBL via SAVE provides (largely) underestimated variance estimates. AMP style algorithms may provide more accurate variance information (per component) as we have shown recently. However, one practical issue associated with most AMP versions is that they may diverge even if for a slight deviation from i.i.d Gaussian or right orthogonally invariant measurement matrices. To this end we extend here the more robust Swept AMP (SwAMP) algorithm to Generalized SwAMP (GSwAMP), which handles independent but non-i.i.d. priors and to the case of dynamic SBL. The simulations illustrate the desirable convergence behavior of the proposed GSwAMP-SBL under different scenarios on the measurement matrix.**

I. Introduction

Sparse signal processing has received tremendous attention during the last decade in many fields including massive multi-input multi-output (MIMO) wireless channel estimation, MIMO radar and image or video processing. The signal model for the recovery of a sparse signal vector $\boldsymbol{x}$ can be formulated as, $\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{y}$ are the observations or data, $\mathbf{A}$ is called the measurement or the sensing matrix which is known and is of dimension $M \times N$ with $M < N$. The signal (which can be modeled static or dynamic) $\boldsymbol{x}$ contains only K non-zero (or significant) entries, with $K << N$. In Bayesian inference, the Sparse Bayesian Learning (SBL) algorithm first proposed by [1], [2] is shown to very good reconstruction performance compared to the deterministic schemes including LASSO or FOCUSS. SBL is based on a two or three layer hierarchical prior on the sparse coefficients $\boldsymbol{x}$. The priors for the hyperparameters (precision parameters) should be judiciously chosen such that the marginal prior for $\boldsymbol{x}$ induces sparsity (for e.g. a student-t prior), allowing the majority of the coefficients to tend towards zero. It is worth mentioning that [3] provides a detailed overview of the various sparse signal recovery algorithms which fall under l_1 or l_2 norm minimization approaches such as Basis Pursuit, LASSO etc and SBL methods. The authors justify the superior recovery performance of SBL compared to the above mentioned conventional methods. Nevertheless, the matrix inversion involved in the Linear Minimum Mean Squared Error (LMMSE) step in SBL at each iteration makes it computationally complex even for moderately large data sets. This complexity is the motivation behind approximate inference methods.

Belief Propagation (BP) based SBL algorithms [4] are computationally more efficient. Due to space limitations we refer the reader to a more detailed discussion on the various approximate inference methods for SBL in [5]. However, due to the computational demands associated with the message passing (MP), several approximate message passing methods are proposed in the recent literature, see for e.g. [6]–[8].

It is of great importance to analyze the convergence conditions of approximate message passing based algorithms. State evolution (SE) analysis done on the class of i.i.d matrices ([6], [9]) show that the mean square error converges to the Bayes optimal value in the large system limit. Unfortunately, while AMP performs well for zero-mean i.i.d. projections, performance tends to drastically decline if the measurement matrix deviates even slightly from this case. The authors in [10] have shown even for i.i.d non-zero mean measurement matrix, the AMP algorithm tends to diverge. Hence to overcome these issues several techniques have been proposed in the literature including adaptive damping, mean removal [11] and sequential AMP (called Swept AMP) [12]. However, issues with damping is that it may further slow down the convergence rate, thus making the algorithm highly complex. Also, it is not yet sure how to determine an optimal damping factor.

A. Contributions of this paper

- To handle SBL, we propose an extension of Swept Approximate Message Passing (GSwAMP), which is robust to more general matrix ensembles for $\mathbf{A}$ and achieves better convergence properties compared to the state of the art Generalized Approximate Message Passing based SBL

(GAMP-SBL) in [13] whose convergence is possible only with damping (slowing down the convergence further).

- We furthermore propose to extend the GSwAMP to dynamic auto-regressive SBL (DAR-SBL) with hyper parameter estimation integrated. Dynamic autoregressive SBL (DAR-SBL) considered here is a case of joint Kalman filtering (KF) with a linear time-invariant diagonal state-space model, and parameter estimation, which can be considered an instance of nonlinear filering.
- Compared to our previous work using BP [5], the proposed GSwAMP-SBL is robust to more general $\mathbf{A}$ (measurement) matrices as is validated through our simulations. Moreover, it provides accurate posterior variance information compared to space alternating variational estimation (SAVE) based SBL scheme proposed in [14].

II. Static SBL - Signal Model

[1] The static compressed sensing problem can be formulated as

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{v}, \tag{1}$$

where $\boldsymbol{y}$ is the observations or data, $\mathbf{A}$ is called the measurement or the sensing matrix which is known and is of dimension $M \times N$ with $M < N$, $\boldsymbol{x}$ is the N-dimensional sparse signal and $\boldsymbol{v}$ is the additive noise. $\boldsymbol{x}$ contains only K non-zero entries, with $K << N$. $\boldsymbol{w}$ is assumed to be a white Gaussian noise, $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{v}; 0, \gamma^{-1}\mathbf{I})$. In Bayesian CS, the sparse signal $\boldsymbol{x}$ (originally it can be deterministic) is assumed to have a two-layer hierarchical prior as in [1]. The hierarchical prior is carefully chosen such that it encourages the sparsity property of $\boldsymbol{x}$. Let $\boldsymbol{\xi}$ be the vector of precision values, with i^{th} element ξ_i. We write the parametrized prior for $\boldsymbol{x}$ as

$$p_{\boldsymbol{x}}(\boldsymbol{x}/\Xi) = \prod_{i=1}^{N} \mathcal{N}(x_i; \mathbf{0}, \Xi^{-1}),\ \Xi = \text{Diag}(\boldsymbol{\xi}). \tag{2}$$

We choose a Gamma prior for $\boldsymbol{\xi}$ (precision parameters), $p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \prod\limits_{i=1}^{N} f_{\xi_i}(\xi_i/a, b) = \prod\limits_{i=1}^{N} \Gamma^{-1}(a) b^a \xi_i^{a-1} e^{-b\xi_i}$. Here $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ represents the 'gamma function'. The resulting marginal pdf of $\boldsymbol{x}$ (student-t distribution) becomes more sparsity inducing than e.g. a Laplacian prior. The inverse of noise variance γ is also assumed to have a Gamma prior, $p_\gamma(\gamma/c, d) = \Gamma^{-1}(c) d^c \gamma_i^{c-1} e^{-d\gamma}$. The advantage of the two-layer prior structure is that the whole machinery of linear MMSE estimation can be exploited, such as e.g., the KF. But this is embedded in other layers making things eventually non-Gaussian. Now the likelihood distribution can be written as

$$p_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{x}, \gamma) = (2\pi)^{-M/2} \gamma^{M/2} e^{\frac{-\gamma||\boldsymbol{y}-\mathbf{A}\boldsymbol{x}||^2}{2}}. \tag{3}$$

To make these priors non-informative [15], we choose them to be small values $a = c = b = d = 10^{-5}$. We denote the unknown parameter vector $\boldsymbol{\theta} = \{\boldsymbol{x}, \Xi, \gamma\}$ and θ_i is each scalar in $\boldsymbol{\theta}$.

A. Fixed Points of Bethe Free Energy and GSwAMP-SBL

When the computation of the posterior distribution becomes intractable, our aim would become to perform probabilistic inference by minimizing the variational free energy (VFE) over an approximate posterior $q(\boldsymbol{\theta})$. The VFE can be written as [16]

$$\mathcal{F}(q) = KLD(q(\boldsymbol{\theta})||P_0(\boldsymbol{\theta})) - <\log p_{\boldsymbol{y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta})>_q \tag{4}$$

where KLD denotes the Kullback-Leibler divergence and $<>_q$ represents the expectation over the approximate distribution q and the prior $P_0(\boldsymbol{\theta}) = p_{\boldsymbol{x}}(\boldsymbol{x}/\Xi) p_{\boldsymbol{\xi}}(\boldsymbol{\xi}) p_\gamma(\gamma/c, d)$. We shall further discuss here briefly the mean field VFE and Bethe free energy (BFE). Under the mean field (MF) approximation, where we consider that the q factorizes over the individual scalar parameters, we can obtain the approximate distribution as $q_{\theta_i}(\theta_i) \propto \exp(<\log p_{\boldsymbol{y}|\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{\theta}) P_0(\boldsymbol{\theta})>_{q_{\boldsymbol{\theta}_{\bar{i}}}(\boldsymbol{\theta}_{\bar{i}})})$. In [16], the authors show that the fixed points of the GAMP MP equations are the stationary points of the cost function termed approximate Bethe Free Energy, which is written below. This simplified form of the Bethe free energy is obtained using the same approximations which lead to GAMP from BP in the large system limit. We denote the MMSE estimate of x_i as $\widehat{x}_i$ and the posterior variance as σ_i^2.

$$\begin{aligned}
&\mathcal{F}_{GAMP}^{Bethe}(r_m, \tau_m, w_k, \widehat{x}_m, \sigma_m^2) = -\textstyle\sum_k \log \mathcal{Z}_k \\
&-\textstyle\sum_m \frac{\sigma_m^2 + (\widehat{x}_m - r_m)^2}{2\tau_m} - \textstyle\sum_k \frac{(w_k - \sum_m A_{km}\widehat{x}_m)^2}{2V_k} \\
&-\textstyle\sum_m \log Z(r_m, \tau_m) \text{ with } V_k = \textstyle\sum_m A_{k,m}^2 \sigma_m^2, \\
&\mathcal{Z}_k = \int \frac{e^{-\frac{(w_k - z_k)^2}{2V_k}}}{\sqrt{2\pi V_k}} P_{y_k|z_k}(y_k|z_k) dz_k
\end{aligned} \tag{5}$$

where $\mathbf{z} = \mathbf{A}\boldsymbol{x}$, with z_k being the k^{th} element. $Z(r_m, \tau_m)$ represents the normalization constant, which gets defined as

$$Z(r_m, \tau_m) = \int P_{x_m}(x_m|\xi_m) e^{-\frac{(x_m - r_m)^2}{2\tau_m}} dx_m. \tag{6}$$

By optimizing (5) alternatingly w.r.t $r_m, \tau_m, w_k, \widehat{x}_m, \sigma_m^2$, we reach the Algorithm 1, which is termed as sequential GAMP or Swept GAMP based SBL (GSwAMP-SBL). In Algorithm 1, the functions f_1, f_2 are defined as follows (which represent MMSE estimate in the Gaussian case as in SBL)

$$\begin{aligned}
f_1(r_m, \tau_m) &= r_m \frac{\xi_m^{-1}}{\xi_m^{-1} + \tau_m}, \\
f_2(r_m, \tau_m) &= (\xi_m + \tau_m^{-1})^{-1}.
\end{aligned} \tag{7}$$

[1]Notations: The operator $(\cdot)^H$ represents the conjugate transpose or conjugate for a matrix or a scalar respectively. In the following, the pdf of a Gaussian random variable x with mean μ and variance σ^2 is given by $\mathcal{N}(x; \mu, \nu)$. x_k represents the k^{th} element of any vector $\boldsymbol{x}$. $KL(q||p)$ represents the Kullback-Leibler distance between the two distributions q, p. $\mathbf{A}_{n,:}$ represents the n^{th} row of $\mathbf{A}$. blkdiag$(\cdot)$ represents blockdiagonal part of a matrix. diag$(\mathbf{X})$ or Diag$(\boldsymbol{x})$ represents a vector obtained by the diagonal elements of the matrix $\mathbf{X}$ or the diagonal matrix obtained with the elements of $\boldsymbol{x}$ in the diagonal respectively. $\mathbf{1}_M$ represents a vector of length M with all ones as the elements. For a matrix $\mathbf{A}$, $\mathbf{A} \geq 0$ implies it is non-negative (all the elements of $\mathbf{A}$ are non-negative). $\mathbf{I}$ or $\mathbf{I}_M$ represents the identity matrix. tr$\{\mathbf{A}\}$ represents the trace of $\mathbf{A}$. $\mathbf{A}_{i,j}$ represents the $(i,j)^{th}$ element of matrix $\mathbf{A}$.

Algorithm 1 GSwAMP-SBL

Input: $\boldsymbol{y}, \mathbf{A}$

Initialize: $\widehat{\gamma}_0 = \frac{c}{d}, \widehat{\xi}_0 = \frac{a}{b}, \sigma_m^2 = 1/(\|\mathbf{A}_m\|^2\gamma + \widehat{\xi}_m), \forall m$, $\widehat{\boldsymbol{x}} = \mathbf{A}^T\boldsymbol{y}$, $V_k = 0, \forall k$, $w_k = 0$.

repeat

  for $k = 1$ to M **do**

    $g_k^{(t)} = \frac{y_k - w_k^{(t;N+1)}}{\frac{1}{\widehat{\gamma}} + V_k^{t;N+1}}$

    $V_k^{(t+1;1)} = \sum_m A_{km}^2 \sigma_m^{2\,(t)}$

    $w_k^{(t+1;1)} = \sum_m A_{k,m}\widehat{x}_m^{(t)} - V_k^{(t+1;1)} g_k^{(t)}$

  end for

  $S = \text{RandomPermute}([1, 2..., N])$

  for $n = 1$ to N **do**

    $m = S_n$

    $\tau_m^{(t+1)} = \left[\sum_k \frac{A_{km}^2}{\frac{1}{\widehat{\gamma}} + V_k^{(t+1;k)}}\right]^{-1}$

    $r_m^{(t+1)} = \widehat{x}_m^{(t)} + \tau_m^{(t+1)} \sum_k A_{km} \frac{y_k - w_k}{\frac{1}{\widehat{\gamma}} + V_k^{(t+1;k)}}$

    $\widehat{x}_m^{(t+1)} = f_1(r_m^{(t+1)}, \tau_m^{(t+1)})$

    $\sigma_m^{2\,(t+1)} = f_2(r_m^{(t+1)}, \tau_m^{(t+1)})$

  end for

  Hyperparameter Estimation (using MF [14, Section 3]

  for $m = 1$ to N **do**

    $\widehat{\xi}_m^{(t+1)} = \frac{a + 1/2}{|\widehat{\boldsymbol{x}}_m^{(t+1)}|^2 + \sigma_m^{2\,(t)}}$, $\widehat{\gamma}^{(t+1)} = \frac{c + N/2}{\frac{\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2}{2} + d}$.

  end for

until convergence

III. Dynamic AR-SBL

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

$$\begin{array}{ll}\text{State Update: } \boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \\ \text{Observation: } \boldsymbol{y}_t = \mathbf{A}^{(t)}\boldsymbol{x}_t + \boldsymbol{v}_t,\end{array} \tag{8}$$

where $\boldsymbol{x}_t = [x_{1,t}, \;\ldots, \; x_{N,t}]^T$. Diagonal matrices $\boldsymbol{F}$ and Ξ are defined with its elements, $\boldsymbol{F}_{i,i} = f_i, f_i \in (-1, 1)$ and $\Xi = \text{diag}(\boldsymbol{\xi}), \boldsymbol{\xi} = [\xi_1, ...\xi_N]$. Further, $\boldsymbol{w}_t \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Lambda}^{-1})$, where $\boldsymbol{\Lambda}^{-1} = \Xi^{-1}(\mathbf{I} - \boldsymbol{F}\boldsymbol{F}^H) = \text{diag}(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_N})$ and $\boldsymbol{v}_t \sim \mathcal{CN}(\mathbf{0}, \frac{1}{\gamma}\mathbf{I})$. $\boldsymbol{w}_t$ are the complex Gaussian mutually uncorrelated state innovation sequences. Hence we sparsify the prediction error variance $\boldsymbol{w}_t$ also, with the same support as $\boldsymbol{x}_0$ and henceforth enforces the same support set for $\boldsymbol{x}_t, \forall t$. $\boldsymbol{v}_t$ is independent of the $\boldsymbol{w}_t$ process. Although the above signal model seems simple, there are numerous applications such as 1) Bayesian adaptive filtering [17], 2) Wireless channel estimation: multipath parameter estimation as in [18]. In this case, $\boldsymbol{x}_t$ = FIR filter response, and Ξ represents e.g. the power delay profile. We also denote the unknown parameter vector $\boldsymbol{\theta}_t = \{\boldsymbol{x}_t, \boldsymbol{\Lambda}, \gamma, \boldsymbol{F}\}$ and θ_i represents each scalar in $\boldsymbol{\theta}$. Note that we only estimate the reparametrized innovation sequence precision instead of the precision variables ξ_i.

IV. GSwAMP-SBL for Nonlinear Kalman Filtering

The joint distribution $p(\boldsymbol{y}_t, \boldsymbol{\theta}_t / \boldsymbol{y}_{1:t-1})$ can be written as ($\boldsymbol{\Sigma}_{t|t-1}$ represents the diagonal prediction covariance matrix)

$$\begin{array}{l}\ln p(\boldsymbol{y}_t, \boldsymbol{\theta}_t / \boldsymbol{y}_{1:t-1}) = \frac{N}{2}\ln\gamma - \frac{\gamma}{2}\,||\boldsymbol{y}_t - \mathbf{A}_t\boldsymbol{x}_t|\,|^2 + \\ -M\det(\widehat{\boldsymbol{\Sigma}}_{t|t-1}) - \frac{1}{2}\left(\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_{t|t-1}\right)^T \widehat{\boldsymbol{\Sigma}}_{t|t-1}^{-1}\left(\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_{t|t-1}\right) \\ +(c-1)\ln\gamma + c\ln d - d\gamma + \text{constants}.\end{array}$$

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

In the prediction stage, similar as in KF, we compute the posterior, $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1})$, where $\boldsymbol{y}_{1:t-1}$ refers to the observations till time $t-1$. For more detailed derivation, we refer to our previous work [19] due to space limitations. This part gets computed using MF, however, the interation between $x_{m,t}$ and f_m requires Gaussian projection, using expectation propagation (EP) [19]. The resulting Gaussian distribution is parameterized as $x_{l,t} \sim \mathcal{N}(x_{l,t}; \widehat{x}_{l,t|t-1}, \sigma^2_{l,t|t-1})$.

B. Measurement Update (Filtering) Stage

For the measurement update stage, the posterior for $\boldsymbol{x}_t$ is inferred using GSwAMP-SBL in Algorithm 1. The posterior mean and diagonal covariance matrix of the estimate computed at $\boldsymbol{x}_t$ are denoted by $\widehat{\boldsymbol{x}}_{t|t}, \boldsymbol{\Sigma}_{t|t}$. We denote each entries in $\widehat{\boldsymbol{x}}_{t|t}$ as $\widehat{\boldsymbol{x}}_{l,t|t}$ respectively. In the measurement stage, the prior for $\boldsymbol{x}_t$ gets replaced by the posterior estimate from the prediction stage. We refer to our previous work [5] for detailed discussions on the fitering stage. One remark here is that compared to our previous work using BP in [5], using GSwAMP gives a more computationally feasible implementation and accurate posterior variances, where $\sigma^2_{l,t|t}$ incorporates the effect of all $\sigma^2_{l',t|t}, l' \neq l$. $\sigma^2_{l,t|t}$ represents the diagonal elements of the posterior covariance matrice $\boldsymbol{\Sigma}_{t|t}$.

C. Lag-1 Smoothing Stage

We obtain the system model for the smoothing stage (by combining the AR(1) stage in the measurement model) as follows

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

where $\widetilde{\boldsymbol{v}}_t \sim \mathcal{CN}(0, \widetilde{\boldsymbol{R}}_t)$ with $\widetilde{\boldsymbol{R}}_t = \mathbf{A}^{(t)}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{(t)\,H} + \frac{1}{\gamma}\mathbf{I}$. We show in [5, Lemma 1] that KF is not enough to adapt the hyperparameters, instead we need at least a lag 1 smoothing (i.e. the computation of $\widehat{\boldsymbol{x}}_{t-1|t}, \boldsymbol{\Sigma}_{t-1|t}$ through GSwAMP-SBL). Here, we first do a noise whitening by multiplying $\boldsymbol{y}_t$ with $\widetilde{\boldsymbol{R}}_t^{-1/2}$. Hence, we can rewrite the observation model as

$$\begin{array}{l}\widehat{\boldsymbol{y}}_t = \widehat{\mathbf{A}}^{(t)}\boldsymbol{x}_{t-1} + \widehat{\boldsymbol{v}}_t, \text{ where } \widehat{\boldsymbol{v}}_t = \widetilde{\boldsymbol{R}}_t^{-1/2}\widetilde{\boldsymbol{v}}_t, \\ \widehat{\mathbf{A}}^{(t)} = \widetilde{\boldsymbol{R}}_t^{-1/2}\mathbf{A}^{(t)}\boldsymbol{F}.\end{array} \tag{10}$$

The joint distribution can be factorized as, $p(\boldsymbol{y}_t, \boldsymbol{\theta} / \boldsymbol{y}_{1:t-1}) = p(\boldsymbol{y}_t / \boldsymbol{\theta}_t) p(\boldsymbol{x}_{t-1} / \boldsymbol{y}_{1:t-1}) p(\boldsymbol{F}, \boldsymbol{\Lambda}, \gamma \mid \boldsymbol{y}_{1:t-1})$.

$$\begin{array}{l}\ln p(\boldsymbol{y}_t, \boldsymbol{\theta}_{t-1} / \boldsymbol{y}_{1:t-1}) = \frac{-1}{2}\ln\det\widetilde{\boldsymbol{R}}_t \\ -|f_m|^2|x_m|^2\mathbf{A}_m^{(t)\,T}\widetilde{\boldsymbol{R}}_t^{-1}\mathbf{A}_m^{(t)} \\ +2\Re(f_m^H x_m^H \mathbf{A}_m^{(t)\,H}\widetilde{\boldsymbol{R}}_t^{-1}(\boldsymbol{y}_t - \mathbf{A}_{\bar{m}}^{(t)}\boldsymbol{F}_{\bar{m}}\boldsymbol{x}_{\bar{m},t})) + c_f,\end{array} \tag{11}$$

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

$$\begin{aligned}&\sigma^{-2}_{f_m|t} = (|\widehat{x}_{m,t-1|t}|^2 + \sigma^2_{m,t-1|t})\mathbf{A}^{(t)\,T}_m \widetilde{\boldsymbol{R}}_t^{-1} \mathbf{A}^{(t)}_m,\\ &\widehat{f}_{m|t} = \sigma^2_{f_m|t}\widehat{x}^H_{m,t-1|t}\mathbf{A}^{(t)\,H}_m \widetilde{\boldsymbol{R}}_t^{-1}(\boldsymbol{y}_t - \mathbf{A}^{(t)}_{\bar{m}} \widehat{\boldsymbol{F}}_{\bar{m}|t}\widehat{\boldsymbol{x}}_{\bar{m},t-1|t}).\end{aligned} \tag{12}$$

Algorithm 2 GSwAMP based DAR-SBL

Initialization $\widehat{f}_{l|0}, \widehat{\lambda}_{l|0} = \frac{a}{b}, \widehat{\gamma}_0 = \frac{c}{d}, \widehat{x}_{l,0|0} = 0, \sigma^2_{l,0|0} = 0, \forall l$.
Define $\boldsymbol{\Sigma}_{t-1|t-1} = \text{diag}(\sigma^2_{l,t|t-1})$.
for $t = 1 : T$ do
Prediction Stage:
1) Compute $\widehat{x}_{l,t|t-1}, \sigma^2_{l,t|t-1}$ using EP and MF [5], $\widehat{x}_{l,t|t-1} = \widehat{f}_{l|t-1}\widehat{x}_{l,t-1|t-1}$, $\sigma^2_{l,t|t-1} = |\widehat{f}_{l|t-1}|^2\sigma^2_{l,t-1|t-1} + \sigma^2_{f_l|t-1}(|\widehat{x}_{l,t-1|t-1}|^2 + \sigma^2_{l,t-1|t-1}) + \widehat{\lambda}^{-1}_{l|t-1}$.

Filtering Stage:
1) Compute $\widehat{x}_{l,t|t}, \sigma^{-2}_{l,t|t}$ using GSwAMP (iterated convergence).

Smoothing Stage:
Initialization: $\boldsymbol{\Sigma}^{(0)}_{t-1|t} = \boldsymbol{\Sigma}_{t-1|t-1}, \widehat{\boldsymbol{x}}^{(0)}_{t-1|t} = \widehat{\boldsymbol{x}}_{t-1|t-1}$. Update $\widetilde{\boldsymbol{R}}_t, \widehat{\mathbf{A}}^{(t)}$.
1) Compute $\widehat{\boldsymbol{x}}_{t-1|t}, \boldsymbol{\Sigma}_{t-1|t}$ using GSwAMP (iterated until convergence)

Estimation of hyperparameters (Define: $x'_{k,t} = x_{k,t} - f_k x_{k,t-1}, \zeta_t = \beta\zeta_{t-1} + (1-\beta) < \left\|\boldsymbol{y}_t - \mathbf{A}^{(t)}\boldsymbol{x}_t\right\|^2 >$**) :**
1) Compute $\widehat{f}_{l|t}, \sigma^2_{f_{l|t}}$ from (12), $\widehat{\gamma}_t = \frac{c+N}{(\zeta_t+d)}$ and $\lambda_{l|t} = \frac{(a+1)}{(<\left|x'_{k,t}\right|^2>_{|t}+b)}$.

V. Simulation Results

To elucidate further the excellent convergence proporties of the GSwAMP-SBL algorithm from other state of the art AMP-SBL versions, we evaluate the normalized MSE (NMSE) performance under different scenarios of $\mathbf{A}$ matrices such as ill-conditioned, non-zero mean matrices for static SBL. We also illustrate the performance of the BP based DAR-SBL compared to our suboptimal methods which are based on MF. Note that simulations are performed with dimensions of $\mathbf{A}, M = 150, N = 250$. The power delay profile (variances of x_i) for the SBL model in Section III is chosen as d^{i-1}, with $d = 0.93$ and starting with index $i = 1$. Further we analysis the following scenarios in the simulations. In Figure 1 and Figure 2, we also assume that the hyperparameters are unknown and get estimated as proposed in our Algorithm 1.

A. *ill-conditioned* $\mathbf{A}$ *case:*

We construct the matrix $\mathbf{A}$ with condition number $\kappa > 1$. Let $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\mathbf{U}, \mathbf{V}^T$ are the left and right singular vectors of an i.i.d-Gaussian matrix. Further, we select the singular values such that $\frac{\Sigma_{i,i}}{\Sigma_{i,i+1}} = \kappa^{1/(M-1)}$, for $i = 1, 2..., M-1$ and $\Sigma_{i,i}$ is the i^{th} diagonal element of $\boldsymbol{\Sigma}$. The more the condition number, the more $\mathbf{A}$ deviates from the i.i.d-Gaussian case. In Figure 1, we plot the NMSE values as a function of the condition number for different algorithms such as original SBL (LMMSE-SBL), SAVE-SBL [14], proposed GSwAMP-SBL and damped GAMP-SBL [13]. In fact, in the simulations we observed that GAMP-SBL does not converge without using damping and there does not exist any closed form solution for the optimal damping value. Hence depending on the particular scenario being considered and also on the dimensions, the damping value may change. However, the proposed GSwAMP-SBL is more robust in the sense that it does not require any damping and convergence to a local optimum is guaranteed.

B. *Non-zero mean* $\mathbf{A}$ *case:*

In this case, we generate each entries of $\mathbf{A}$ as i.i.d Gaussian with a non-zero mean, $A_{i,j} \sim \mathcal{N}(\mu, \frac{1}{M})$. We plot the NMSE performance for different algorithms in Figure 2 as a function of the mean of $\mathbf{A}$. We observe that GAMP-SBL does not converge, in this case apart from damping we may require mean removal procedure also as in noted in [11]. However, SAVE-SBL and the proposed GSwAMP-SBL converges without any mean removal procedure. Hence, GSwAMP-SBL would be preferred from an implementation complexity perspective. SAVE-SBL has the incorrect posterior variance issue which we have observed in our previous papers.

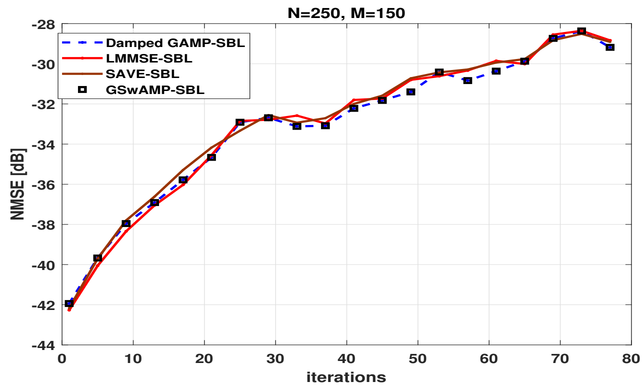

Fig. 1. NMSE vs Condition number of the measurement matrix $\mathbf{A}$.

C. *DAR-SBL*

For the observation model, the parameters chosen are $N = 256, M = 200$. All signals are considered to be real in the simulation. All the elements of the factor matrix $\mathbf{A}^{(t)}_j$ (time varying) are generated i.i.d. from a Gaussian distribution with mean 0 and variance 1. The rows of $\mathbf{A}^{(t)}$ are scaled by $\sqrt{16}$ so that the signal part of any scalar observation has unit variance. Taking the SNR to be 20dB, the variance of each element of $\boldsymbol{v}_t$ (Gaussian with mean 0) is computed as 0.01.
Consider the state update, $\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{w}_t$. To generate $\boldsymbol{x}_0$, the first 16 elements are chosen as Gaussian (mean 0 and variance 1) and then the remaining elements of the vector $\boldsymbol{x}_0$ are

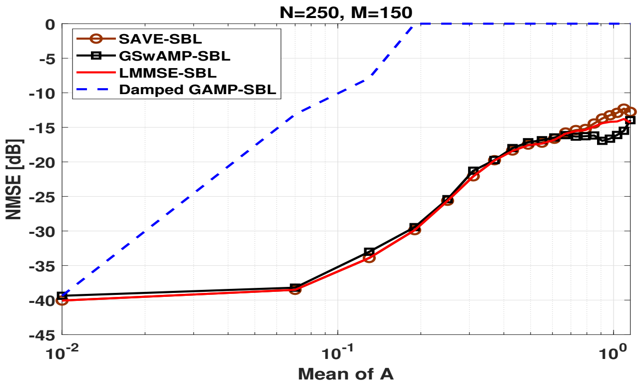

Fig. 2. NMSE vs the mean of $\mathbf{A}$.

put to zero. Then the elements of $\boldsymbol{x}_0$ are randomly permuted to distribute the 30 non-zero elements across the whole vector. The diagonal elements of $\boldsymbol{F}$ are chosen uniformly in $[0.9, 1)$. Then the covariance of $\boldsymbol{w}_t$ can be computed as $\Xi^{-1}(\mathbf{I}-\boldsymbol{F}\boldsymbol{F}^T)$. Note that Ξ^{-1} contains the variances of the elements of $\boldsymbol{x}_t$ (including $t = 0$), where for the non-zero elements of $\boldsymbol{x}_0$ the variance is 1. Following observations can be made from the simulations.Our proposed low complexity algorithm using BP has similar performance as that of joint VB which has higher complexity. In Figure 3, we evaluate the performance of the BP-MF-EP DAR SBL and show that the parameter estimation benefits from BP. Indeed, note that BP can also be implemented by GSwAMP-SBL for better computational feasibility and under i.i.d $\mathbf{A}$ BP and GSwAMP-SBL converge to the same solution. "MF DAR-SBL" refers to the sub-optimal version with no BP and only MF for filtering or smoothing of $\boldsymbol{x}_t$. Also we show the drastic improvement in performance with lag-1 smoothing for hyperparameter estimation compared to just using filtering.

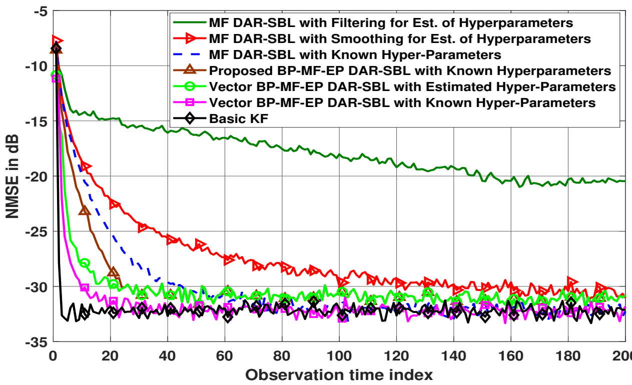

Fig. 3. DAR-SBL: NMSE as a function of time.

VI. Conclusions

In this paper, we look at the robustness of the SBL algorithm under deviations from i.i.d Gaussian assumptions of measurement matrix. Towards this direction, we propose a GSwAMP-SBL algorithm which implements the GAMP sequentially rather than parallel as in the original GAMP version by Rangan [7]. Among the many techniques proposed for improving the convergence properties of AMP algorithms, GSwAMP stands out due to the low cost per iteration, compared to the highly complex nature of damping or mean removal based algorithms in the literature. We also integrate hyper parameter estimation (by MF) and an extension of the GSwAMP-SBL for a time varying sparse signal is also proposed.

VII. Acknowledgments

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

References

[1] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learn. Res.*, vol. 1, 2001.

[2] D. P. Wipf and B. D. Rao, "Sparse Bayesian Learning for Basis Selection ," *IEEE Trans. on Sig. Proc.*, vol. 52, no. 8, Aug. 2004.

[3] R. Giri and Bhaskar D. Rao, "Type I and type II bayesian methods for sparse signal recovery using scale mixtures," *IEEE Trans. on Sig Process.*, vol. 64, no. 13, 2018.

[4] X. Tan and J. Li, "Computationally Efficient Sparse Bayesian Learning via Belief Propagation ," *IEEE Trans. on Sig. Proc.*, vol. 58, no. 4, Apr. 2013.

[5] C. K. Thomas and D. Slock, "Low Complexity Static and Dynamic Sparse Bayesian Learning Combining BP, VB and EP Message Passing," in *52nd IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2019.

[6] Mohsen Bayati and Andrea Montanari, "The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing ," *IEEE Trans. on Inf. Theory*, vol. 57, no. 2, February 2011.

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

[8] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector Approximate Message Passing," *IEEE Trans. On Info. Theo.*, vol. 65, no. 10, Oct. 2019.

[9] D. L. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," in *IEEE Intl. Symp. on Info. Theo.*, 2012.

[10] F. Caltagirone, F. Krzakala, and L. Zdeborová, "On the Convergence of Approximate Message Passing," in *IEEE Intl. Symp. on Info. Theo.*, 2014.

[11] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, "Adaptive damping and mean removal for the generalized approximate message passing algorithm," in *In IEEE Intl. Conf. on Acous., Spee. and Sig. Process. (ICASSP)*, April 2015.

[12] A. Manoel, F. Krzakala, E. W. Tramel, and L. Zdeborová, "Swept Approximate Message Passing for Sparse Estimation," in *Proc. of the 32nd Intl. Conf. on Mach. Learn.*, 2015.

[13] Maher Al-Shoukairi, Philip Schniter, and Bhaskar D. Rao, "GAMP-based low complexity sparse bayesian learning algorithm," *IEEE Trans. on Sig. Process.*, vol. 66, no. 2, January 2018.

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

[15] J. O. Berger, "Statistical Decision Theory and Bayesian Analysis," *Second ed. New York: Springer-Verlag*, 1985.

[16] F. Krzakala, A. Manoel, E. W. Tramel, and L. Zdeborova, "Variational Free Energies for Compressed Sensing," in *IEEE Intl. Sympo. Info. Theo.*, Honolulu, HI, USA, Jun. 2014.

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

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

[19] C. K. Thomas and D. Slock, "Gaussian Variational Bayes Kalman Filtering for Dynamic Sparse Bayesian Learning," in *ITISE*, 2018.