Gaussian Variational Bayes Kalman Filtering for Dynamic Sparse Bayesian Learning

Christo Kurisummoottil Thomas* and Dirk Slock

EURECOM, Sophia-Antipolis, France
{kurisumm,slock}@eurecom.fr
http://www.eurecom.fr/cm/

Abstract. Sparse Baysesian Learning (SBL) provides sophisticated (state) model order selection with unknown support distribution. This allows to handle problems with big state dimensions and relatively limited data. The techniques proposed in this paper allow to handle the extension of SBL to time-varying states, modeled as diagonal first-order vector auto-regressive (VAR(1)) processes with unknown parameters. Adding the parameters to the state leads to an augmented state and a non-linear (at least bilinear) state-space model. The proposed approach, which applies also to more general non-linear models, uses Variational Bayes (VB) techniques to approximate the posterior distribution by a factored form, with Gaussian or exponential factors. The granularity of the factorization can take on various levels. In one extreme instance, called Gausian Space Alternating Variational Estimation Kalman Filtering (GSAVE-KF), all state components are treated individually, leading to low complexity filtering. Simulations illustrate the performance of the proposed GVB-KF techniques, which represent an alternative to Linear MMSE (LMMSE) filtering.

Keywords: Sparse Bayesian Learning, Variational Bayes, Kalman Filtering

1 Introduction

Sparse signal reconstruction and compressed sensing (CS) has received significant attraction in the recent years. The signal model for the recovery of a time varying sparse signal can be formulated as,

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

where $\mathbf{y}_t$ is the observations or data at time t, $\mathbf{A}_t$ is called the measurement or the sensing matrix which is known and is of dimension $N \times M$ with $N < M$, $\mathbf{x}_t$ is the M-dimensional sparse signal and $\mathbf{v}_t$ is the additive noise. $\mathbf{x}_t$ contains only K non-zero entries, with $K << M$ and is modeled by a diagonal AR(1) (auto-regressive) process. $\mathbf{v}_t$ is assumed to be a white Gaussian noise, $\mathbf{v}_t \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$.

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

In the time invariant case, to address this problem, there exists a variety of algorithms such as the basis pursuit method [1] and the orthogonal matching pursuit [2]. In Bayesian learning, sparse Bayesian learning (SBL) algorithm was first proposed by [3, 4]. Performance can be further improved by exploiting the temporal correlation across the sparse vectors [5]. However, most of these algorithms do offline or batch processing, whose complexity doesn't scale with the problem size. In order to render low complexity or low latency solutions, online processing algorithms (which processes small set of measurement vectors at any time) will be necessary.

In sparse adaptive estimation [6], a time varying signal $\mathbf{x}_t$ is estimated time-recursively by exploiting the sparsity property of the signal. Conventional adaptive filtering methods such as LMS or recursive least squares (RLS) doesn't exploit the underlying sparseness in the signal $\mathbf{x}_t$ to improve the estimation performance. An approach to combine Kalman filtering and compressed sensing can be found in [7]. Kalman filter focus on estimation of the dynamical state from noisy observations where the dynamic and measurement process are considered to be from linear Gaussian state space model. Compared to the state of the art, we introduce not only sparse filter (state) but also sparse filter variations. We apply SBL now to the prediction error variances of $\mathbf{x}_t$, then trying to sparsify a prediction error variance actually encourage both that the actual variance gets sparse and that the variation gets sparse because a prediction error variance is small if either the quantity variance is small or its variation is small.

In the literature, there exist different KF based methods to handle the joint filtering and parameter estimation problem. One such example is the widely used EM-KF algorithm ([8, 9]) which uses the famous Expectation Maximization technique (EM), and alternating optimization technique for ML estimation. To handle general nonlinear state space models, another variation called as Extended KF (EKF) algorithm exists. In this case, the state is extended with the unknown parameters, rendering the new state update equation nonlinear. A third derivation is the truncated Second-Order EKF (SOEKF) introduced by [10, 11] in which nonlinearities are expanded up to second order, third and higher order statistics being neglected. [12] present a corrected derivation of SOEKF and show that the state of the art contains illogical approximations. In ([11, 13]), the Gaussian SOEKF is derived in which fourth-order terms in the Taylor series expansions are retained and approximated by assuming that the underlying joint probability distribution is Gaussian. In [14], Villares et al. introduced the Quadratic Extended Kalman Filter (QEKF). The authors extend the EKF to deal with quadratic signal models and exploiting the fourth order signal statistics. We proposed a space alternating variational estimation based technique for single measurement vectors in [15].

1.1 Contributions of this paper

– We propose a novel Gaussian approximation Space Alternating Variational Estimation (GSAVE) based SBL technique for LMMSE filtering called GSAVE-KF. The proposed solution is for a multiple measurement case with an AR(1)

process for the temporal correlation of the sparse signal. The update and prediction stages of the proposed algorithm reveals links to the Kalman filter.
- For the static state case, numerical results presented elsewhere [15] suggest that the proposed solution has a faster convergence rate (and hence lower complexity) than (even) the existing fast SBL and performs better than the existing fast SBL algorithms in terms of reconstruction error in the presence of noise.
- For the dynamic state case considered here, simulations suggest that in spite of both significantly reduced computational complexity and the estimation of the unknown (hyper) parameters, the GSAVE-KF algorithm exhibits hardly any MSE degradation in steady-state compared to the standard Kalman filter with known parameters, but at the cost of a significantly increased transient duration.

In the following, boldface lower-case and upper-case characters denote vectors and matrices respectively. the operators $tr(\cdot)$, $(\cdot)^T$ represents trace,and transpose respectively. The operator $(\cdot)^H$ represents the conjugate transpose or conjugate for a matrix or a scalar respectively. A complex Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Theta}$ is distributed as $\mathbf{x} \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Theta})$. $diag(\cdot)$ represents the diagonal matrix created by elements of a row or column vector. The operator $< x >$ or $\mathrm{E}(\cdot)$ represents the expectation of x. $||\cdot||$ represents the Frobenius norm. $\Re\{(\cdot)\}$ represents the real part of $(\cdot)$. All the variables are complex here unless specified otherwise.

2 State Space Model

Sparse signal $\mathbf{x}_t$ is modeled using an AR(1) process with correlation coefficient matrix $\mathbf{F}$, with $\mathbf{F}$ diagonal. The state space model can be written as follows,

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t, \quad \text{State Update,} \\ \mathbf{y}_t &= \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t, \qquad \text{Observation,} \end{aligned} \tag{2}$$

where $\mathbf{x}_t = [x_{1,t}, \ ..., \ x_{M,t}]^T$. Matrices $\mathbf{F}$ and $\boldsymbol{\Gamma}$ are defined as,

$$\mathbf{F} = \begin{bmatrix} f_1 & 0 & \dots & 0 \\ 0 & f_2 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_M \end{bmatrix}, \boldsymbol{\Gamma} = \begin{bmatrix} \frac{1}{\sqrt{\alpha_1}} & \dots & 0 \\ 0 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{\sqrt{\alpha_M}} \end{bmatrix}, \tag{3}$$

Here f_i represents the correlation coefficient and α_i represents the inverse variance of $x_{i,t} \sim \mathcal{CN}(0, \frac{1}{\alpha_i})$. Further, $\mathbf{w}_t \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Gamma}(\mathbf{I}-\mathbf{F}\mathbf{F}^H))$ and $\mathbf{v}_t \sim \mathcal{CN}(\mathbf{0}, \frac{1}{\gamma}\mathbf{I})$. $\mathbf{w}_t$ are the complex Gaussian mutually uncorrelated state innovation sequences. $\mathbf{v}_t$ is independent of the $\mathbf{w}_t$ process. Further we define, $\boldsymbol{\Lambda} = \boldsymbol{\Gamma}(\mathbf{I} - \mathbf{F}\mathbf{F}^H) = \mathrm{diag}\,(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_M})$.

Although the above signal model seems simple, there are numerous applications such as 1) Bayesian adaptive filtering [16, 17], 2)Wireless channel estimation: multi-path parameter estimation as in [18, 19]. In this case, $\mathbf{x}_t$ = FIR filter response, and $\boldsymbol{\Gamma}$ represents e.g. the power delay profile.

3 VB-SBL

In Bayesian compressive sensing, a two-layer hierarchical prior is assumed for the $\mathbf{x}$ as in [3]. The hierarchical prior is such that it encourages the sparsity property of $\mathbf{x}_t$ or of the innovation sequences $\mathbf{v}_t$.

$$\begin{aligned} p(\mathbf{x}_t/\boldsymbol{\Gamma}) &= \prod_{i=1}^{M} p(x_{i,t}/\alpha_i) = \prod_{i=1}^{M} \mathcal{N}(0, \alpha_i^{-1}), \\ p(\mathbf{x}_t/\mathbf{x}_{t-1}, \mathbf{F}, \boldsymbol{\Gamma}) &= \prod_{i=1}^{M} p(x_{i,t}/x_{i,t-1}, \alpha_i, f_i) = \prod_{i=1}^{M} \mathcal{N}(f_i x_{i,t-1}, \frac{1}{\alpha_i}). \end{aligned} \tag{4}$$

For the convenience of analysis, we reparameterize α_i in terms of λ_i and assume a Gamma prior for $\boldsymbol{\Lambda}$,

$$p(\boldsymbol{\Lambda}) = \prod_{i=1}^{M} p(\lambda_i/a, b) = \prod_{i=1}^{M} \Gamma^{-1}(a) b^a \lambda_i^{a-1} e^{-b\lambda_i}. \tag{5}$$

The inverse of noise variance γ is also assumed to have a Gamma prior,

$$p(\gamma/c, d) = \Gamma^{-1}(c) d^c \gamma_i^{c-1} e^{-d\gamma}. \tag{6}$$

Now the likelihood distribution can be written as,

$$p(\mathbf{y}_t/\mathbf{x}_t, \gamma) = (2\pi)^{-N} \gamma^N e^{\frac{-\gamma ||\mathbf{y}_t - \mathbf{A}_t \mathbf{x}_t||^2}{2}}. \tag{7}$$

To make these priors non-informative, we choose them to be small values $a = c = b = d = 10^{-5}$.

3.1 Variational Bayesian Inference

The computation of the posterior distribution of the parameters is usually intractable. In order to address this issue, in variational Bayesian framework, the posterior distribution $p(\mathbf{x}_t, \boldsymbol{\Lambda}, \gamma/\mathbf{y}_{1:t})$ is approximated by a variational distribution $q(\mathbf{x}_t, \boldsymbol{\Lambda}, \gamma)$ that has the factorized form:

$$q(\mathbf{x}_t, \boldsymbol{\Lambda}, \gamma) \;=\; q_\gamma(\gamma) \prod_{i=1}^{M} q_{x_{i,t}}(x_{i,t}) \prod_{i=1}^{M} q_{\lambda_i}(\lambda_i), \tag{8}$$

where $\mathbf{y}_{1:t}$ represents the observations till the time t $(\mathbf{y}_1, ..., \mathbf{y}_t)$, similarly we define $\mathbf{x}_{1:t}$. Variational Bayes compute the factors q by minimizing the Kullback-Leibler distance between the true posterior distribution $p(\mathbf{x}_t, \boldsymbol{\Lambda}, \gamma/\mathbf{y}_{1:t})$ and the $q(\mathbf{x}_t, \boldsymbol{\Lambda}, \gamma)$. From [20],

$$KLD_{VB} \;=\; KL\left(p(\mathbf{x}_t, \boldsymbol{\Lambda}, \gamma/\mathbf{y}_{1:t}) || q(\mathbf{x}_t, \boldsymbol{\Lambda}, \gamma)\right) \tag{9}$$

The KL divergence minimization is equivalent to maximizing the evidence lower bound (ELBO) [21]. To elaborate on this, we can write the marginal probability of the observed data as,

$$\begin{aligned} &\ln p(\mathbf{y}_t/\mathbf{y}_{1:t-1}) = L(q) + KLD_{VB}, \;\; \text{where,} \\ &L(q) = \int q(\mathbf{x}_t, \boldsymbol{\theta}) \ln \tfrac{p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1})}{q(\boldsymbol{\theta})} d\mathbf{x}_t d\boldsymbol{\theta}, \\ &KLD_{VB} = -\int q(\mathbf{x}_t, \boldsymbol{\theta}) \ln \tfrac{p(\mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t})}{q(\mathbf{x}_t, \boldsymbol{\theta})} d\mathbf{x}_t d\boldsymbol{\theta}, \end{aligned} \tag{10}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\Lambda}, \gamma\}$ and θ_i represents each scalar in $\boldsymbol{\theta}$. Since $KLD_{VB} \geq 0$, it implies that $L(q)$ is a lower bound on $\ln p(\mathbf{y}_t/\mathbf{y}_{1:t-1})$. Moreover, $\ln p(\mathbf{y}_t/\mathbf{y}_{1:t-1})$ is independent of $q(\mathbf{x}_t, \boldsymbol{\theta})$ and therefore maximizing $L(q)$ is equivalent to minimizing KLD_{VB}. This is called as ELBO maximization and doing this in an alternating fashion for each variable in $\mathbf{x}_t, \boldsymbol{\theta}$ leads to,

$$\begin{aligned}
&\ln(q_i(\theta_i)) = < \ln p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) >_{\boldsymbol{\theta}_{\bar{i}}, \mathbf{x}_t} + c_i, \\
&\ln(q_i(x_{i,t})) = < \ln p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) >_{\boldsymbol{\theta}, \mathbf{x}_{\bar{i},t}} + c_i, \\
&p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) = p(\mathbf{y}_t/\mathbf{x}_t, \gamma, \mathbf{y}_{1:t-1})\, p(\mathbf{x}_t/\boldsymbol{\Lambda}, \mathbf{y}_{1:t-1}) p(\boldsymbol{\Lambda}) p(\gamma).
\end{aligned} \tag{11}$$

Here $<>_{k \neq i}$ represents the expectation operator over the distributions q_k for all $k \neq i$. $\mathbf{x}_{\bar{i},t}$ represents $\mathbf{x}_t$ without x_i and $\boldsymbol{\theta}_{\bar{i}}$ represents $\boldsymbol{\theta}$ without θ_i. In section 5, we consider another variant where the components of $\mathbf{x}_t$ are treated jointly, where the approximate posterior becomes $q(\mathbf{x}_t, \boldsymbol{\Lambda}, \gamma) = q_\gamma(\gamma) q_{\mathbf{x}_t}(\mathbf{x}_t) \prod_{i=1}^{M} q_{\lambda_i}(\lambda_i)$.

3.2 Gaussian Posterior Minimizing the KL Divergence

In [22], for any distribution $p(\mathbf{x})$, the Gaussian distribution $q(\mathbf{x}) \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which minimizes the Kullback-Leibler divergence, $KL(p||q)$, reduces to matching the mean and covariance,

$$\boldsymbol{\mu} = < \mathbf{x} >_{p(x)}, \quad \boldsymbol{\Sigma} = < \mathbf{x}\mathbf{x}^H >_{p(x)} - < \mathbf{x} >_{p(x)} < \mathbf{x} >_{p(x)}^H . \tag{12}$$

4 SAVE Sparse Bayesian Learning and Kalman Filtering

In this section, we propose a Space Alternating Variational Estimation (SAVE) based alternating optimization between each element of $\mathbf{x}_t$ or γ. For SAVE, no particular structure of $\mathbf{A}_t$ is assumed, in contrast to AMP which performs poorly when $\mathbf{A}_t$ is not i.i.d or is sub-Gaussian. The joint distribution w.r.t the observation of (2) can be written as,

$$p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) = p(\mathbf{y}_t/\mathbf{x}_t, \boldsymbol{\theta}) p(\mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}). \tag{13}$$

In the following, $c_{x_{k,t}}, c'_{x_{k,t}}, c_{\alpha_k}$, c_{λ_k}, c_{x-1}, c_{x_t}, c'_{x_t} and c_γ represents normalization constants for the respective pdfs.

4.1 Diagonal AR(1) (DAR(1)) Prediction Stage

In this stage, we compute the prediction about $\mathbf{x}_t$ given the observations till time $t-1$, $\widehat{x}_{k,t|t-1}$. The joint distribution for the state space model can be written as,

$$\begin{aligned}
&\ln p(x_{k,t}, x_{k,t-1}, f_k, \lambda_k \mid \mathbf{y}_{1:t-1}) = -\lambda_k (x_{k,t} - f_k x_{k,t-1})^H (x_{k,t} - f_k x_{k,t-1}) - \\
&\frac{1}{\sigma^2_{k,t-1|t-1}} |x_{k,t-1} - \widehat{x}_{k,t-1|t-1}|^2 + ((a-1)\ln\lambda_k + a \ln b - b\lambda_k) .
\end{aligned} \tag{14}$$

The prediction about $\mathbf{x}_t$ can be computed from the time update equation of the standard Kalman filter,

$$x_{k,t} = \widehat{f}_{k|t-1} x_{k,t-1} + \widetilde{f}_{k|t-1} x_{k,t-1} + w_{k,t}. \tag{15}$$

Here we denote $\widehat{f}_{k|t-1}$ as the estimate of f_k given the observations till $t-1$ and $\widetilde{f}_{k|t-1}$ represents the error in the estimation. Similary we can represent $x_{k,t-1} = \widehat{x}_{k,t-1|t-1} + \widetilde{x}_{k,t-1|t-1}$, $\widetilde{x}_{k,t-1|t-1}$ being the estimation error.

$$\begin{aligned} &\widehat{x}_{k,t|t-1} = \widehat{f}_{k|t-1}\widehat{x}_{k,t-1|t-1},\ \ \widetilde{x}_{k,t|t-1} = \widehat{f}_{k|t-1}\widetilde{x}_{k,t-1|t-1} + \widetilde{f}_{k|t-1}x_{k,t-1} + w_{k,t}, \\ &\implies \sigma^2_{k,t|t-1} \overset{(a)}{=} |\widehat{f}_{k|t-1}|^2 \sigma^2_{k,t-1|t-1} + \sigma^2_{f_k}(|\widehat{x}_{k,t-1|t-1}|^2 + \sigma^2_{k,t-1|t-1}) + \tfrac{1}{\widehat{\lambda}_{k|t-1}}, \end{aligned} \tag{16}$$

In the variational approximation, we assume that the posterior of f_k and $x_{k,t}$ are independent. (a) in (16) follows from this argument. Further the predictive distribution $p(\mathbf{x}_t/\mathbf{y}_{1:t-1})$ can be approximated to be Gaussian distributed (refer to the discussion in section 3.2) with mean $\widehat{\mathbf{x}}_{t|t-1} = [\widehat{x}_{1,t|t-1}, ..., \widehat{x}_{M,t|t-1}]^T$ and diagonal error covariance $\widehat{\mathbf{P}}_{t|t-1} = \text{diag}\,(\sigma^2_{1,t|t-1}, ..., \sigma^2_{M,t|t-1})$. Further the joint distribution in (13) can be obtained as,

$$\begin{aligned} &\ln p(\mathbf{y}_t, \mathbf{x}_t, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) = N \ln \gamma - \gamma\, ||\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t|\,|^2 - M \ln \det(\widehat{\mathbf{P}}_{t|t-1}) - \\ &\left(\mathbf{x}_t - \widehat{\mathbf{x}}_{t|t-1}\right)^H \widehat{\mathbf{P}}^{-1}_{t|t-1} \left(\mathbf{x}_t - \widehat{\mathbf{x}}_{t|t-1}\right) + (c-1)\ln\gamma + c \ln d - d\gamma + \text{constants}, \end{aligned} \tag{17}$$

4.2 Measurement or Update Stage

Update of $q_{x_{k,t}}(x_{k,t})$: Using (11), $\ln q_{x_{k,t}}(x_{k,t})$ turns out to be quadratic in $x_{k,t}$ and thus can be represented as a Gaussian distribution as follows,

$$\begin{aligned} &\ln q_{x_{k,t}}(x_{k,t}) = - < \gamma > \Big\{ (\mathbf{y}_t - \mathbf{A}_{t,\overline{k}} < \mathbf{x}_{\overline{k},t} >)^H \mathbf{A}_{t,k} x_{k,t} \ - \ x^H_{k,t} \mathbf{A}^H_{t,k} \\ &(\mathbf{y}_t - \mathbf{A}_{t,\overline{k}} < \mathbf{x}_{\overline{k},t} >) \ + \ ||\mathbf{A}_{t,k}|\,|^2 \, |x_{k,t}|^2 \Big\} \ - \ \tfrac{1}{\sigma^2{}_{k,t|t-1}} \Big(|x_{k,t}|^2 - x^H_{k,t}\widehat{x}_{k,t|t-1} - \\ &x_{k,t}\widehat{x}^H_{k,t|t-1} \Big) + c_{x_{k,t}} = -\tfrac{1}{\sigma^2_{k,t|t}} \left| x_{k,t} \ - \ \widehat{x}_{k,t|t} \right|^2 + c'_{x_{k,t}}. \end{aligned} \tag{18}$$

Note that we split $\mathbf{A}_t\mathbf{x}_t$ as, $\mathbf{A}_t\mathbf{x}_t \ = \ \mathbf{A}_{t,k}x_{k,t} \ + \ \mathbf{A}_{t,\overline{k}}\mathbf{x}_{\overline{k},t}$, where $\mathbf{A}_{t,k}$ represents the k^{th} column of $\mathbf{A}_t$, $\mathbf{A}_{t,\overline{k}}$ represents the matrix with k^{th} column of $\mathbf{A}_t$ removed. Clearly, the mean and the variance of the resulting Gaussian distribution becomes,

$$\begin{aligned} &\sigma^{-2,(i)}_{k,t|t} = < \gamma > ||\mathbf{A}_{t,k}|\,|^2 \ + \ \sigma^{-2,(i-1)}_{k,t|t}, \\ &< x^{(i)}_{k,t|t} > = \ \sigma^{2,(i)}_{k,t|t} \left(\mathbf{A}^H_{t,k} \left(\mathbf{y}_t \ - \ \mathbf{A}_{t,\overline{k}} < \mathbf{x}^{(i-1)}_{\overline{k},t} > \right) < \gamma > + \tfrac{\widehat{x}_{k,t|t-1}}{\sigma^2_{k,t|t-1}} \right), \end{aligned} \tag{19}$$

where i represents the iteration stage with $\lim_{i\to\infty} < x^{(i)}_{k,t|t} > = \widehat{x}_{k,t|t}$ represents the point estimate of $x_{k,t}$. However, in (19) the computation of $< x^{(i)}_{k,t|t} >$ requires

the knowledge of $< \mathbf{x}_{k,t}^{(i)} >$. So we need to perform enough iterations between the components of $< x_{k,t|t} >$ till convergence. Moreover, we initialize $< x_{k,t|t}^{(0)} >$ by $\widehat{x}_{k,t|t-1}$and $\sigma_{k,t}^{-2,(0)} = \sigma_{k,t|t-1}^{-2}$, which is obtained in the prediction stage. One remark is that forcing a Gaussian posterior q with diagonal covariance matrix on the original Kalman measurement equations gives the same result as SAVE. Note that the derivations in [23] for VB-KF are not correct as it does not have the correct variance expressions that vary with iteration! For the convenience of the derivations in the following sections, we define $\widehat{\mathbf{P}}_{t|t} = diag(\sigma_{1,t|t}^2, ..., \sigma_{M,t|t}^2)$, $\widehat{\mathbf{x}}_{t|t} = [\widehat{x}_{1,t|t}, ..., \widehat{x}_{M,t|t}]^T$.

4.3 Fixed Lag Smoothing

Kalman filtering in the EM-KF is not enough to adapt the hyper parameters, instead we need atleast a lag 1 smoothing [24]. Motivated by this result, we propose fixed lag smoothing with delay 1 for SAVE-KF. We rewrite the state space model as follows,

$$\begin{aligned} &\mathbf{y}_t = \mathbf{A}_t\mathbf{F}\mathbf{x}_{t-1} + \underbrace{\mathbf{A}_t\mathbf{w}_{t-1} + \mathbf{v}_t}_{\widetilde{\mathbf{v}}_t}, \\ &p(\mathbf{y}_t, \mathbf{x}_{t-1}, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) = p(\mathbf{y}_t/\mathbf{x}_{t-1}, \boldsymbol{\theta})p(\mathbf{x}_{t-1}, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}), \end{aligned} \tag{20}$$

where $\widetilde{\mathbf{v}}_t \sim \mathcal{CN}(\mathbf{0}, \widetilde{\mathbf{R}}_t)$, $\widetilde{\mathbf{R}}_t = \mathbf{A}_t\boldsymbol{\Lambda}\mathbf{A}_t^H + \frac{1}{\gamma}\mathbf{I}$. The posterior distribution $p(\mathbf{x}_{t-1}/\mathbf{y}_{1:t-1})$ is approximated using variational approximation as $q(\mathbf{x}_{t-1}/\mathbf{y}_{1:t-1})$ with mean and covariance as $\widehat{\mathbf{x}}_{t-1|t-1}$ and $\widehat{\mathbf{P}}_{t-1|t-1}$.

$$\begin{aligned} &\ln p(\mathbf{y}_t, \mathbf{x}_{t-1}, \boldsymbol{\theta}/\mathbf{y}_{1:t-1}) = \tfrac{-1}{2}\ln\det\widetilde{\mathbf{R}}_t - (\mathbf{y}_t - \mathbf{A}_t\mathbf{F}\mathbf{x}_{t-1})^H\widetilde{\mathbf{R}}_t^{-1}(\mathbf{y}_t - \mathbf{A}_t\mathbf{F}\mathbf{x}_{t-1}) \\ &-\tfrac{1}{2}\det(\widehat{\mathbf{P}}_{t-1|t-1}) - \left(\mathbf{x}_{t-1} - \widehat{\mathbf{x}}_{t-1|t-1}\right)^H\widehat{\mathbf{P}}_{t-1|t-1}^{-1}\left(\mathbf{x}_{t-1} - \widehat{\mathbf{x}}_{t-1|t-1}\right) + c_{x-1}. \end{aligned} \tag{21}$$

Prediction of $\mathbf{x}_{t-1}$: Using (11), $\ln q_{\mathbf{x}_{t-1}}(\mathbf{x}_{t-1}/\mathbf{y}_{1:t})$ turns out to be quadratic in $\mathbf{x}_{t-1}$ and thus can be represented as a Gaussian distribution with mean and covariance as $\widehat{\mathbf{x}}_{t-1|t}$ and $\widehat{\mathbf{P}}_{t-1|t}$ respectively,

$$\begin{aligned} &\widehat{\mathbf{P}}_{t-1|t}^{(i)} = (< \mathbf{F}^H\mathbf{A}_t^H\widetilde{\mathbf{R}}_t^{-1}\mathbf{A}_t\mathbf{F} > + \widehat{\mathbf{P}}_{t-1|t}^{-(i-1)})^{-1}, \\ &\widehat{\mathbf{x}}_{t-1|t}^{(i)} = \widehat{\mathbf{P}}_{t-1|t}^{(i)}(\widehat{\mathbf{P}}_{t-1|t-1}^{-1}\widehat{\mathbf{x}}_{t-1|t}^{(i-1)} + < \mathbf{F}^H\mathbf{A}_t^H\widetilde{\mathbf{R}}_t^{-1}\mathbf{y}_t >). \end{aligned} \tag{22}$$

To simplify further, we substitute $\mathbf{F} = \widehat{\mathbf{F}}_{|t} + \widetilde{\mathbf{F}}_{|t}$ and the following expressions can be obtained,

$$\begin{aligned} &\widehat{\mathbf{P}}_{t-1|t}^{(i)} = (\widehat{\mathbf{F}}_{|t}^H\mathbf{A}_t^H\widetilde{\mathbf{R}}_t^{-1}\mathbf{A}_t\widehat{\mathbf{F}}_{|t} + \text{ diag}\,(\mathbf{A}_t^H\widetilde{\mathbf{R}}_t^{-1}\mathbf{A}_t)\widehat{\mathbf{P}}_{\mathbf{F}|t} + \widehat{\mathbf{P}}_{t-1|t}^{-(i-1)})^{-1}, \\ &\widehat{\mathbf{x}}_{t-1|t}^{(i)} = \widehat{\mathbf{P}}_{t-1|t}^{(i)}(\widehat{\mathbf{P}}_{t-1|t-1}^{-1}\widehat{\mathbf{x}}_{t-1|t}^{(i-1)} + \widehat{\mathbf{F}}_{|t}^H\mathbf{A}_t^H\widetilde{\mathbf{R}}_t^{-1}\mathbf{y}_t). \end{aligned} \tag{23}$$

Note that, in the algorithm implementation as shown in Algorithm 1 below, we introduce an iterative procedure (with i denoting the stage number) for the smoothing updates unlike [23] where there is no iteration for the covariance part. Note that we initialize the mean and variance in (22) from the converged values from the filtering stage.

4.4 Estimation of Hyper-Parameters

Update of $q_\gamma(\gamma)$: The Gamma distribution from the variational Bayesian approximation for the $q_\gamma(\gamma)$ can be written as,

$$\begin{aligned} &\ln q_\gamma(\gamma) \;=\; (c-1+N)\ln\gamma \;-\; \gamma\left(<||\mathbf{y}_t \;-\; \mathbf{A}_t\mathbf{x}_t||^2> + d\right) \;+\; c_\gamma, \\ &q_\gamma(\gamma) \;\propto\; \gamma^{c+N-1} e^{-\gamma\left(<||\mathbf{y}_t - \mathbf{A}_t\mathbf{x}_t||^2> + d\right)}. \end{aligned} \tag{24}$$

The mean of the Gamma distribution for γ is given by,

$$\begin{aligned} &<\gamma> = \widehat{\gamma}_t = \tfrac{c+\frac{N}{2}}{(\zeta_t + d)},\;\; \zeta_t = \beta\zeta_{t-1} + (1-\beta)<||\mathbf{y}_t \;-\; \mathbf{A}_t\mathbf{x}_t||^2>,\;\; \text{where,} \\ &<||\mathbf{y}_t \;-\; \mathbf{A}_t\mathbf{x}_t||^2> = ||\mathbf{y}_t||^2 \;-\; 2\Re(\mathbf{y}_t^H\mathbf{A}_t\widehat{\mathbf{x}}_{t|t}) \;+\; \mathrm{tr}\left(\mathbf{A}_t^H\mathbf{A}_t(\widehat{\mathbf{x}}_{t|t}\widehat{\mathbf{x}}_{t|t}^H \;+\; \widehat{\mathbf{P}}_{t|t})\right), \end{aligned} \tag{25}$$

where we introduced temporal averaging also and β denotes the weighting coefficients which are less than one.

Update of $q_{f_k}(f_k)$: Using variational approximation we get a quadratic expression for $\ln q(f_k|\mathbf{y}_{1:t}) \sim \mathrm{E}_{q(\mathbf{x}_t,\mathbf{x}_{t-1},\boldsymbol{\Lambda}/\mathbf{y}_{1:t})} \ln p(\mathbf{x}_t,\mathbf{x}_{t-1},\boldsymbol{\Lambda},\mathbf{y}_{1:t})$. Finally we write the mean and variance of the resulting Gaussian distribution as,

$$\sigma^2_{f_k|t} = \tfrac{1}{\lambda_k <x^2_{k,t-1}>_{|t}},\;\; \widehat{f}_{k|t} = \tfrac{<x_{k,t|t}x^H_{k,t-1|t}>_{|t}}{<x^2_{k,t-1}>_{|t}} \tag{26}$$

Here $<>_{|t}$ represents the temporal average given the observations till time t. We introduce temporal averaging here to approximate terms of the form $< x_{k,t|t}x^H_{k,t-1|t} >$. This is done using the orthogonality property of LMMSE. So $< x_{k,t|t}x^H_{k,t-1|t} >=< \widehat{x}_{k,t|t}\widehat{x}^H_{k,t-1|t} > + < \widetilde{x}_{k,t|t}\widetilde{x}^H_{k,t-1|t} >$. The Kalman filter (in linear state-space models and Gaussian noise) provides instantaneous $\widehat{x}_{k,t|t}, \widehat{x}^H_{k,t-1|t}$ and $\sigma^2_{k,t|t}, \sigma^2_{k,t-1|t}$. This explains why we do temporal averaging (sample average replacing statistical average). We define $\widehat{\mathbf{P}}_{\mathbf{F}|t} = \mathrm{diag}\left(\sigma^2_{f_1|t}, ..., \sigma^2_{f_M|t}\right)$. Also we define the following covariance matrices, $R^{m,n}_t =< \mathbf{x}_{t-n}\mathbf{x}^H_{t-m} >_{|t}$ and ξ_t represents the temporal weighting coefficient which is less than one [24],

$$\begin{aligned} &\mathbf{R}^{0,0}_t = (1-\xi_t)\,\mathbf{R}^{0,0}_{t-1} + \xi_t\,(\widehat{\mathbf{x}}_{t|t}\widehat{\mathbf{x}}^H_{t|t} + \widehat{\mathbf{P}}_{t|t}),\;\; \mathbf{R}^{1,0}_t = (\mathbf{R}^{0,1}_t)^H = (1-\xi_t)\,\mathbf{R}^{1,0}_{t-1} + \\ &\xi_t\,\mathbf{F}(\widehat{\mathbf{x}}_{t-1|t}\widehat{\mathbf{x}}^H_{t-1|t} + \widehat{\mathbf{P}}_{t-1|t}),\;\; \mathbf{R}^{1,1}_t = (1-\xi_t)\,\mathbf{R}^{1,1}_{t-1} + \xi_t(\widehat{\mathbf{x}}_{t-1|t}\widehat{\mathbf{x}}^H_{t-1|t} + \widehat{\mathbf{P}}_{t-1|t}). \end{aligned} \tag{27}$$

Further, we denote the $(i,j)^{th}$ element of $\mathbf{R}^{mn}_t$ as $\mathbf{R}^{mn}_t(i,j)$.

Update of $q_{\lambda_k}(\lambda_k)$: Using variational approximation
$\ln q(\lambda_k|\mathbf{y}_{1:t}) \sim \mathrm{E}_{q(\mathbf{x}_t,\mathbf{x}_{t-1},f_k/\mathbf{y}_{1:t})} \ln p(\mathbf{x}_t,\boldsymbol{\Lambda},f_k \mid \mathbf{y}_{1:t})$, leading to

$$\begin{aligned} &\ln\lambda_k - \lambda_k(<|x_{k,t} - f_k x_{k,t-1}|^2> + b) + (a-1)\ln\lambda_k + c_{\lambda_k}, \\ &q_{\lambda_k}(\lambda_k) \propto \lambda_k^a e^{-\lambda_k\left(<|x_{k,t}-f_k x_{k,t-1}|^2> + b\right)}. \end{aligned} \tag{28}$$

The resulting gamma distribution is parameterized just by one quantity, the mean value, which gets used in the prediction stage and can be written as,

$$< \lambda_k >= \tfrac{(a+1)}{(<|x_{k,t}-f_k x_{k,t-1}|^2>_{|t} + b)}. \tag{29}$$

The temporal average $< |x_{k,t} - f_k x_{k,t-1}|^2 >_{|t}$ can be written as,

$$\begin{aligned} &< |x_{k,t} - f_k x_{k,t-1}|^2 >_{|t} = \\ &\mathbf{R}_t^{0,0}(k,k) - 2\Re\{\widehat{f}_{k|t}^H \mathbf{R}_t^{1,0}(k,k)\} + (|\widehat{f}_{k|t}|^2 + \sigma_{f_k|t}^2)\mathbf{R}_t^{1,1}(k,k). \end{aligned} \tag{30}$$

In Algorithm 1, we describe the GSAVE-KF algorithm in detail.

Algorithm 1 The GSAVE-KF Algorithm

Given: $\mathbf{A}_t, \mathbf{y}_t, N, M, \lambda_{k|0} = a/b \ \forall k, \gamma_0 = c/d, \sigma_{k,0|0}^2 = 0, \widehat{x}_{k,0|0} = 0 \ \forall k, t > 0.$

Prediction Stage

$\sigma_{k,t|t-1}^2 = (|\widehat{f}_{k|t-1}|^2 + \sigma_{f_{k|t-1}}^2)\sigma_{k,t-1|t-1}^2 + \frac{1}{\widehat{\lambda}_{k|t-1}}, \ \widehat{x}_{k,t|t-1} = \widehat{f}_{k|t-1}\widehat{x}_{k,t-1|t-1},$

Update Stage

Initialization: $\sigma_{k,t|t}^{2,(0)} = \sigma_{k,t|t-1}^{2,(0)}, \widehat{x}_{t,\overline{k}|t}^{(0)} = \widehat{x}_{t,\overline{k}|t-1}$

for $i = 1, ...$until convergence

$\sigma_{k,t|t}^{2,(i)} = \sigma_{k,t|t}^{2,(i-1)}(\sigma_{k,t|t}^{2,(i-1)}\widehat{\gamma}_{t-1} ||\mathbf{A}_{t,k}||^2 + 1)^{-1},$

Kalman Gain $\mathbf{K}_{k,t} = \sigma_{k,t|t}^{2,(i)} \mathbf{A}_{t,k}^H \widehat{\gamma}_{t-1},$

$\widehat{x}_{k,t|t}^{(i)} = \frac{\sigma_{k,t|t}^{2,(i)}}{\sigma_{k,t|t-1}^2}\widehat{x}_{k,t|t-1} + \mathbf{K}_{k,t}\left(\mathbf{y}_t - \mathbf{A}_{t,\overline{k}}\widehat{\mathbf{x}}_{t,\overline{k}|t}^{(i-1)}\right),$

end for

Smoothing Stage

Initialization: $\widehat{\mathbf{P}}_{t-1|t}^{(0)} = \widehat{\mathbf{P}}_{t-1|t-1}, \widehat{\mathbf{x}}_{t-1|t}^{(0)} = \widehat{\mathbf{x}}_{t-1|t-1}$

for $i = 1, ...,$ until convergence

$\widehat{\mathbf{P}}_{t-1|t}^{-(i)} = (\widehat{\mathbf{F}}_{|t}^H \mathbf{A}_t^H \widetilde{\mathbf{R}}_t^{-1} \mathbf{A}_t \widehat{\mathbf{F}}_{|t} + \text{diag}\,(\mathbf{A}_t^H \widetilde{\mathbf{R}}_t^{-1} \mathbf{A}_t)\widehat{\mathbf{P}}_{\mathbf{F}|t} + \widehat{\mathbf{P}}_{t-1|t}^{-(i-1)}),$

$\widehat{\mathbf{x}}_{t-1|t}^{(i)} = \widehat{\mathbf{P}}_{t-1|t}^{(i)}(\widehat{\mathbf{P}}_{t-1|t-1}^{-1}\widehat{\mathbf{x}}_{t-1|t}^{(i-1)} + \widehat{\mathbf{F}}^H \mathbf{A}_t^H \widetilde{\mathbf{R}}_t^{-1}\mathbf{y}_t).$

end for

Estimation of Hyper-Parameters

Compute $\zeta_t, \mathbf{R}_t^{m,n}$ from (25), (27).

$\sigma_{f_k|t}^2 = \frac{1}{\lambda_k \mathbf{R}_t^{1,1}(k,k)}, \ \widehat{f}_{k|t} = \frac{\mathbf{R}_t^{1,0}(k,k)}{\mathbf{R}_t^{1,1}(k,k)}.$

$\widehat{\gamma}_t = \frac{c+\frac{N}{2}}{(\zeta_t + d)}, \ \widehat{\lambda}_{k|t} = \frac{a+1}{(\mathbf{R}_t^{0,0}(k,k) - 2\Re\{\widehat{f}_{k|t}^H \mathbf{R}_t^{1,0}(k,k)\} + (|\widehat{f}_{k|t}|^2 + \sigma_{f_k|t}^2)\mathbf{R}_t^{1,1}(k,k) + b)}.$

5 VB-KF for Diagonal AR(1) (DAR(1))

In this section, we treat the components of the state $\mathbf{x}_t$ jointly, with all the hyper-parameters λ_k, f_k, γ assumed to be independent in the q's. So the expressions for the estimates of the hyper-parameters can be shown to be the same as in the previous section on SAVE-KF.

5.1 DAR(1) Prediction Stage

The prediction about $\mathbf{x}_t$ can be computed from the time update equation of the standard Kalman filter,

$$\mathbf{x}_t = \widehat{\mathbf{F}}_{|t-1}\mathbf{x}_{t-1|t-1} + \widetilde{\mathbf{F}}_{|t-1}\mathbf{x}_{t-1|t-1} + \mathbf{v}_t, \ \ \mathbf{F} = \widehat{\mathbf{F}}_{|t-1} + \widehat{\mathbf{F}}_{|t-1}, \tag{31}$$

where $\widehat{\mathbf{F}}_{|t-1} = \text{diag}\,(\widehat{f}_{1|t-1}, ..., \widehat{f}_{M|t-1})$. We also define $\widehat{\boldsymbol{\Lambda}}_{|t-1} = \text{diag}\,(\frac{1}{\widehat{\lambda}_{1|t-1}}, ..., \frac{1}{\widehat{\lambda}_{M|t-1}})$. Substituting $\mathbf{x}_{t-1|t-1} = \widehat{\mathbf{x}}_{t-1|t-1} + \widetilde{\mathbf{x}}_{t-1|t-1}$,

$$\begin{aligned}&\widehat{\mathbf{x}}_{t|t-1} = \widehat{\mathbf{F}}_{|t-1}\widehat{\mathbf{x}}_{t-1|t-1},\ \widetilde{\mathbf{x}}_{t|t-1} = \widehat{\mathbf{F}}_{|t-1}\widetilde{\mathbf{x}}_{t-1|t-1} + \widetilde{\mathbf{F}}_{|t-1}\mathbf{x}_{t-1|t-1} + \mathbf{w}_t,\ \Longrightarrow \\ &\widehat{\mathbf{P}}_{t|t-1} = \widehat{F}_{|t-1}\widehat{\mathbf{P}}_{t-1|t-1}\widehat{F}^H_{|t-1} + \widehat{\mathbf{P}}_{\mathbf{F}|t-1}\,\text{diag}\,(\widehat{\mathbf{x}}_{t-1|t-1}\widehat{\mathbf{x}}^H_{t-1|t-1} + \widehat{\mathbf{P}}_{t-1|t-1}) + \widehat{\boldsymbol{\Lambda}}_{|t-1}.\end{aligned} \tag{32}$$

5.2 Measurement or Update Stage

Using (11),

$$\begin{aligned}&\ln q_{\mathbf{x}_t}(\mathbf{x}_t) = - < \gamma > \Big\{ - \mathbf{y}_t^H\mathbf{A}_t\mathbf{x}_t\ - \ \mathbf{x}_t^H\mathbf{A}_t^H\mathbf{y}_t + \mathbf{x}_t^H\mathbf{A}_t^H\mathbf{A}_t\mathbf{x}_t\Big\}\ - \ \mathbf{x}_t^H\widehat{\mathbf{P}}^{-1}_{t|t-1}\mathbf{x}_t \\ &+\mathbf{x}_t^H\widehat{\mathbf{P}}^{-1}_{t|t-1}\widehat{\mathbf{x}}_{t|t-1} + \widehat{\mathbf{x}}^H_{t|t-1}\widehat{\mathbf{P}}^{-1}_{t|t-1}\mathbf{x}_t + c_{x_t} = -(\mathbf{x}_t\ -\ \widehat{\mathbf{x}}_{t|t})^H\widehat{\mathbf{P}}^{-1}_{t|t}(\mathbf{x}_t\ -\ \widehat{\mathbf{x}}_{t|t}) + c'_{x_t},\end{aligned} \tag{33}$$

where the mean and variance are written as,

$$\widehat{\mathbf{P}}^{-1}_{t|t} =< \gamma > \mathbf{A}_t^H\mathbf{A}_t + \widehat{\mathbf{P}}^{-1}_{t|t-1},\ \widehat{\mathbf{x}}_{t|t} = \widehat{\mathbf{P}}_{t|t}(< \gamma > \mathbf{A}_t^H\mathbf{y}_t + \widehat{\mathbf{P}}^{-1}_{t|t-1}\widehat{\mathbf{x}}_{t|t-1}). \tag{34}$$

6 Simulation Results

For the observation model, $\mathbf{y}_t$ is of dimension 100×1 and $\mathbf{x}_t$ is of size 200×1 with 30 non-zero elements. All signals are considered to be real in the simulation. All the elements of $\mathbf{A}_t$ (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 $\sqrt{30}$ so that the signal part of any scalar observation has unit variance. Taking the SNR to be 20dB, the variance of each element of $\mathbf{v}_t$ (Gaussian with mean 0) is computed as 0.01.

Consider the state update, $\mathbf{x}_t = \mathbf{F}\mathbf{x}_{t-1} + \mathbf{w}_t$. To generate $\mathbf{x}_0$, the first 30 elements are chosen as Gaussian (mean 0 and variance 1) and then the remaining elements of the vector $\mathbf{x}_0$ are put to zero. Then the elements of $\mathbf{x}_0$ are randomly permuted to distribute the 30 non-zero elements across the whole vector. The diagonal elements of $\mathbf{F}$ are chosen uniformly in $[0.9, 1)$. Then the covariance of $\mathbf{w}_t$ can be computed as $\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{F}\mathbf{F}^H)$. Note that $\boldsymbol{\Lambda}$ contains the variances of the elements of $\mathbf{x}_t$ (including $t = 0$), where for the non-zero elements of $\mathbf{x}_0$ the variance is 1 and for the zero elements it is 0. Note that $\mathbf{v}_t$ is Gaussian distributed with mean 0. In Fig. 1, the blue curve corresponds to the case of a standard Kalman Filter with known state-space model parameters. The red curve corresponds to GSAVE-KF with again all these hyper-parameters known. The green curve corresponds to the case of GSAVE-KF with all the hyper parameters also estimated with lag-1 smoothing. Further, we show that filtering for AR(1) coefficients (black curve) doesn't converge to the basic KF. NMSE is the normalized mean squared error at time t computed as $||\mathbf{x}_t - \widehat{\mathbf{x}}_t||^2$, averaged over 100 different realizations of $\mathbf{A}_t$, $\mathbf{F}$, and of course the noise realizations. The simulations show that in the scenario considered, GSAVE-KF exhibits hardly any MSE degradation over the more complex standard Kalman Filter in steady-state, but takes time to reach steady-state. Adding the estimation of the parameters leads to further slight degradations in steady-state and transient.

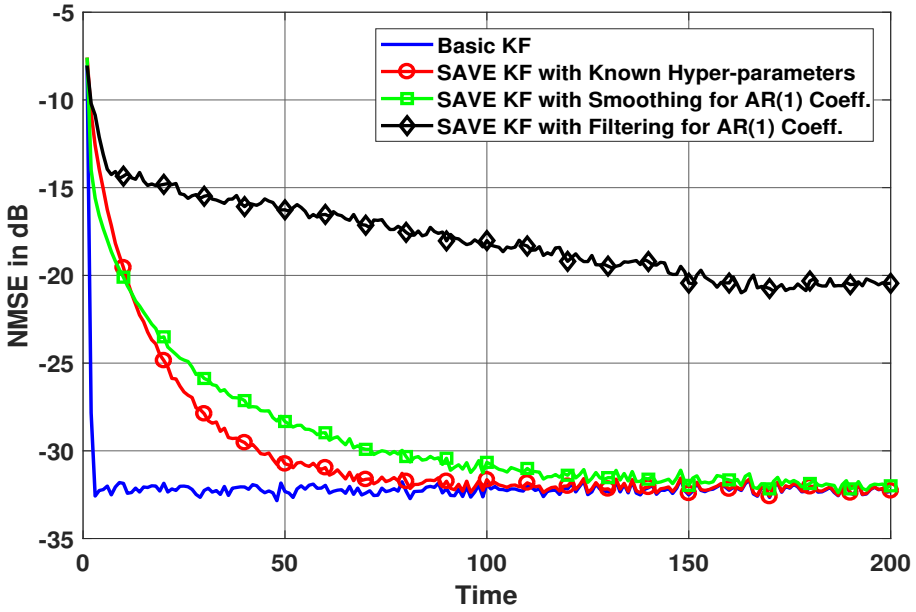

Fig. 1. NMSE as a function of time (i.e. number of measurements or iteration index).

7 Conclusions

We presented a fast SBL algorithm called GSAVE-KF, which uses the variational inference techniques to approximate the posteriors of the data and parameters and track a time varying sparse signal. GSAVE-KF helps to circumvent the matrix inversion operation required in conventional SBL using the EM algorithm. We showed that in spite of the significantly reduced computational complexity, the proposed algorithm with estimation of the unknown model parameters has similar steady-state performance compared to the standard Kalman filter, at the price of a significantly increased transient. The GSAVE-KF algorithm exploits the underlying sparsity in the signal compared to classical Kalman filtering based methods.

References

1. S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit ," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.
2. J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit ," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, December 2007.
3. M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.
4. D. P. Wipf and B. D. Rao, "Sparse Bayesian Learning for Basis Selection ," *IEEE Trans. on Sig. Process.*, vol. 52, no. 8, pp. 2153–2164, August 2004.

5. Z. Zhang and B. D. Rao, "Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning ," *IEEE J. of Sel. Topics in Sig. Process.*, vol. 5, no. 5, pp. 912 – 926, September 2011.
6. D. Angelosan, J. A. Bazerq, and G. B. Giannakis, "Online adaptive estimation of sparse signals: where RLS meets the l1-norm," *IEEE Trans. on Sig. Process.*, vol. 58, no. 7, Jul. 2010.
7. N. Vaswani, "Kalman filtered compressed sensing," in *Int. Conf. Image Process.*, San Diego, CA,, Oct. 2008.
8. C. Couvreur and Y. Bresler, "Decomposition of a mixture of Gaussian AR processes," *IEEE Intl. Conf. on Acous., Speech, and Sig. Process.*, vol. 3, pp. 1605–1608, 1995.
9. M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the em algorithm," *IEEE Transactions on Acous., Speech and Sig. Process.,*, vol. 36, no. 4, pp. 477–489, Apr 1988.
10. R. D. Bass, V. D. Norum, and L. Swartz, "Optimal multichannel nonlinear filtering," *J. Mufh. Anal. Appl.*, vol. 16, pp. 152 – 164, 1966.
11. A. H. Jazwinski, *Stochastic processes and filtering theory*, 1970.
12. R. Henriksen, "The truncated second-order nonlinear filter revisited," *IEEE Transactions on Automatic Control*, vol. 27, no. 1, pp. 247 – 251, feb 1982.
13. M. Athans, R. Wishner, and A. Bertolini, "Suboptimal state estimation for continuous-time nonlinear systems from discrete noisy measurements," *IEEE Transactions on Automatic Control*, vol. 13, no. 5, pp. 504 – 514, oct 1968.
14. J. Villares and G. Vazquez, "The quadratic extended Kalman filter," in *Sens. Arr. and Multichnl. Sig. Process. Wkshp. (SAM), 2004*, july 2004, pp. 480 – 484.
15. C. K. Thomas and D. Slock, "SAVE - space alternating variational estimation for sparse Bayesian learning," in *Data Science Workshop*, 2018.
16. T. Sadiki and D. T. Slock, "Bayesian adaptive filtering: principles and practical approaches," in *EUSIPCO*, 2004.
17. J. B. S. Ciochina, C. Paleologu, "A family of optimized LMS-based algorithms for system identification," in *Proc. EUSIPCO*, 2016, pp. 1803–1807.
18. C. K. Thomas and D. Slock, "Variational Bayesian Learning for Channel Estimation and Transceiver Determination," in *Information Theory and Applications Workshop*, San Diego, USA, February 2018.
19. B. H. Fleury, M. Tschudin, R. Heddergott, D. Dahlhaus, and K. I. Pedersen, "Channel Parameter Estimation in Mobile Radio Environments Using the SAGE Algorithm ," *IEEE J. on Sel. Areas in Commun.*, vol. 17, no. 3, pp. 434–450, March 1999.
20. M. J. Beal, "Variational algorithms for approximate Bayesian inference," in *Thesis, Univeristy of Cambridge, UK*, May 2003.
21. D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Sig. Process. Mag.*, vol. 29, no. 6, pp. 131–146, November 2008.
22. R. Herbrich, "Minimising the Kullback-Leibler Divergence," in *Microsoft Research*, August 2015.
23. B. Ait-El-Fquih and I. Hoteit, "Fast Kalman-like filtering for large-dimensional linear and Gaussian state-space models," *IEEE Trans. on Sig. Process.*, vol. 63, no. 21, Nov. 2015.
24. S. Bensaid and D. Slock, "Comparison of Various Approaches for Joint Wiener/Kalman Filtering and Parameter Estimation with Application to BASS," in *IEEE 45th Asilomar Conference on Sig., Sys. and Comp.*, Pacific Grove, CA, USA, 2011.