Probabilistic filter and smoother for variational inference of Bayesian linear dynamical systems

Julian Neri, Roland Badeau, Philippe Depalle

PROBABILISTIC FILTER AND SMOOTHER FOR VARIATIONAL INFERENCE OF BAYESIAN LINEAR DYNAMICAL SYSTEMS

Julian Neri[*] *Roland Badeau*[†] *Philippe Depalle*[*]

[*]McGill University, CIRMMT, Montréal, Canada
[†]LTCI, Télécom Paris, Institut Polytechnique de Paris, France.

ABSTRACT

Variational inference of the Bayesian linear dynamical system is a powerful method for estimating latent variable sequences and learning sparse dynamic models in domains ranging from neuroscience to audio processing. The hardest part of the method is inferring the model's latent variable sequence. Here, we propose a solution using matrix inversion lemmas to derive what may be considered as the Bayesian counterparts to the Kalman filter and smoother, which are particular forms of the forward-backward algorithm that have known properties of numerical stability and efficiency. Opposed to existing methods, we do not augment the model dimensionality, use Cholesky decompositions or inaccurate numerical matrix inversions. We provide mathematical proof and empirical evidence that the new algorithm respects parameter expected values to more accurately infer hidden state statistics. An application to Bayesian frequency estimation of a stochastic sum of sinusoids model is presented and compared with state-of-the-art estimators.

Index Terms— Time-series, Kalman filter, Variational inference

1. INTRODUCTION

Bayesian linear dynamical systems [1] are central to probabilistic machine learning of sequential data and have proven beneficial in a wide range of disciplines including control systems, audio processing, target tracking, finance, autonomous navigation, and neuroscience [2–8]. The system assumes that the linear dynamical system's parameters are stochastic with prior probabilities. Inferring the model with Bayes' theorem provides not only an estimate of the latent state sequence but also of the model structure itself [9]. While exact inference in the Bayesian linear dynamical system is intractable, variational inference has been successful in inferring an approximation to the posterior, employing a structured mean-field factorization of the latent variable sequence from the parameters.

The most challenging aspect of the variational approach is inferring the statistics of the latent variable sequence. For non-Bayesian models, this is completed through the Kalman filter [10] and Rauch-Tung-Striebel (RTS) smoother [11], being particularly efficient, numerically accurate, and well-studied forms of the forward-backward algorithm for state space models [12]. However, in the Bayesian model the parameters are stochastic and summarized by statistical moments, complicating the forward-backward algorithm. Several approaches have been proposed to infer the sufficient statistics of the hidden state sequence. Originally, [1] developed a specialized routine based on belief propagation that avoided the use of matrix inversion lemmas as such operations appeared to violate parameter expectations. Later, [2] found that Kalman filtering equations could be used at the cost of using Cholesky decompositions of the parameter covariances and augmenting the state space. Lastly, [13] used a Cholesky decomposition of a block-banded matrix akin to [14], departing from the forward-backward algorithm.

In this paper, we use matrix inversion principles to derive a Bayesian Kalman filter and RTS smoother, a numerically stable and efficient form of the forward-backward algorithm for inferring the state sequence of fully Bayesian linear dynamical systems. Opposed to existing methods, we do not rely on augmenting the dimensionality of the state space, Cholesky decompositions, or undesirable numerical matrix inversions, while enjoying a cost that only grows linearly with the sequence length. Crucially, we incorporate uncertainty about parameter values intuitively and respect all the parameter covariances and expectations as in the original belief propagation formulation. Overall, the resulting algorithm is faster and more numerically accurate than the original. An application to Bayesian frequency estimation in a sum of sinusoids model is presented and compared to state-of-the-art deterministic estimators.

The paper is organized as follows. Section 2 details the probabilistic model for the Bayesian linear dynamical system. The proposed inference routine is described in Section 3, then validated along with existing methods in Section 4. The variational approach is applied to the problem of frequency estimation in Section 5. Finally, Section 6 concludes the paper and proposes future work.

The following notation is used throughout the paper: bold lowercase denotes vectors and bold uppercase denotes matrices; $\mathbf{x}_{1:n}$ denotes the set $(\mathbf{x}_1, \dots, \mathbf{x}_n)$; $\mathbf{I}$ is the identity matrix; $|\mathbf{A}|$ is the determinant and $\mathrm{Tr}(\mathbf{A})$ is the trace of matrix $\mathbf{A}$; $\mathrm{diag}(\mathbf{a})$ denotes a diagonal matrix formed from the elements of $\mathbf{a}$; column vector $\mathbf{a}_{(i)}$ denotes the transposed ith row of $\mathbf{A}$; $\langle \mathbf{x} \rangle$ denotes the expected value of $\mathbf{x}$ and covariance $\mathrm{cov}[\mathbf{x}, \mathbf{y}] = \langle \mathbf{x}\mathbf{y}^\mathsf{T} \rangle - \langle \mathbf{x} \rangle \langle \mathbf{y} \rangle^\mathsf{T}$; $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; $\mathrm{Gam}(x|a, b)$ is a Gamma distribution with shape a and rate b [15].

2. PROBABILISTIC MODEL

Linear dynamical systems assume that $H \times 1$ hidden state $\mathbf{x}_n$ evolves linearly according to a first-order Markov process and emits a $V \times 1$ observation $\mathbf{y}_n$ at each time $n \in (1, N)$. A state is linearly transformed over adjacent times by $H \times H$ system matrix $\mathbf{A}$, and to the observable space by $V \times H$ output matrix $\mathbf{C}$. An optional $U \times 1$ input $\mathbf{u}_n$ is transformed by $H \times U$ input matrix $\mathbf{B}$ to drive the state and by $V \times U$ feedforward matrix $\mathbf{D}$ to drive the observation. Additive Gaussian noise is assumed for both the state $\boldsymbol{\eta}_n^x \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ and observation $\boldsymbol{\eta}_n^y \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, where positive semi-definite covariance matrix $\mathbf{Q}$ is $H \times H$ and $\mathbf{R}$ is $V \times V$. The linear dynamical

[*]Thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding.

system has the following state space representation:

$$\mathbf{x}_n = \mathbf{A}\mathbf{x}_{n-1} + \mathbf{B}\mathbf{u}_n + \boldsymbol{\eta}_n^x \tag{1}$$
$$\mathbf{y}_n = \mathbf{C}\mathbf{x}_n + \mathbf{D}\mathbf{u}_n + \boldsymbol{\eta}_n^y \tag{2}$$

Initial state $\mathbf{x}_1$ is Gaussian-distributed with $H \times 1$ mean $\mathbf{m}_0$ and $H \times H$ covariance $\mathbf{P}_0$. The joint probability of the linear dynamical system is

$$p(\mathbf{Y}, \mathbf{X}) = p(\mathbf{y}_1|\mathbf{x}_1)p(\mathbf{x}_1)\prod_{n=2}^{N} p(\mathbf{y}_n|\mathbf{x}_n)p(\mathbf{x}_n|\mathbf{x}_{n-1}) \tag{3}$$

where the probability of transition $p(\mathbf{x}_n|\mathbf{x}_{n-1})$, emission $p(\mathbf{y}_n|\mathbf{x}_n)$, and initial state $p(\mathbf{x}_1)$ are Gaussian (due to additive Gaussian noise):

$$p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\mathbf{m}_0, \mathbf{P}_0) \tag{4}$$
$$p(\mathbf{x}_n|\mathbf{x}_{n-1}) = \mathcal{N}(\mathbf{x}_n|\mathbf{A}\mathbf{x}_{n-1} + \mathbf{B}\mathbf{u}_n, \mathbf{Q}) \tag{5}$$
$$p(\mathbf{y}_n|\mathbf{x}_n) = \mathcal{N}(\mathbf{y}_n|\mathbf{C}\mathbf{x}_n + \mathbf{D}\mathbf{u}_n, \mathbf{R}). \tag{6}$$

Bayesian linear dynamical systems assume that the parameters of the model $\boldsymbol{\theta} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{R}, \mathbf{Q})$ are also stochastic and prescribed with a prior distribution $p(\boldsymbol{\theta})$. While the prior can take any form, placing Normal-Gamma conditional distributions over the rows of dynamics matrices and elements of diagonal noise covariances is a logical choice motivated by conjugacy, reducing the number of induced factorizations needed to make the variational approach analytically feasible. Moreover, scaling the variance of each element of $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ with automatic relevance determination (ARD) hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}$, respectively, has proven beneficial for pruning superfluous parameters and learning sparser models [1]. Here, we do not explicitly define $p(\boldsymbol{\theta})$ because the new method generally supports any prior. An example of parameter priors is detailed in Section 5.

2.1. Posterior Approximation

Exact inference in the fully Bayesian model is intractable; however, structured mean-field variational inference has been successfully used to infer an approximate posterior distribution q that assumes a single factorization between variables and parameters:

$$p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y}) \approx q(\mathbf{X}, \boldsymbol{\theta}) = q_{\mathbf{x}}(\mathbf{X})q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{7}$$

Variational inference circumvents the intractable integral involved in minimizing the Kullback-Leibler (KL) divergence from q to p by instead maximizing the lower bound

$$\mathcal{L}(q) = \ln p(\mathbf{Y}) - \mathrm{KL}(q(\mathbf{X}, \boldsymbol{\theta}) \| p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y})) \tag{8}$$
$$= \langle \ln p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \rangle_q - \langle \ln q(\mathbf{X}, \boldsymbol{\theta}) \rangle_q \tag{9}$$

From the calculus of variations, the log optimal distributions are

$$\ln q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \langle \ln p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \rangle_{q_{\mathbf{x}}(\mathbf{X})} + \text{const.} \tag{10}$$
$$\ln q_{\mathbf{x}}(\mathbf{X}) = \langle \ln p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} + \text{const.} \tag{11}$$

Each distribution is thus updated in turn to maximize $\mathcal{L}(q)$ [9].

As a consequence of the model's first-order latent structure, updating $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ does not require statistics from the full $q_{\mathbf{x}}(\mathbf{X})$ but only from the marginal $q_{\mathbf{x}}(\mathbf{x}_n)$ and cross-time posterior $q_{\mathbf{x}}(\mathbf{x}_n, \mathbf{x}_{n+1})$. Such statistics can be computed efficiently with the forward-backward algorithm [9]. Given deterministic parameters, the Kalman filter and smoother can compute such statistics. However, in our case parameters are stochastic and described by their sufficient statistics under $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, complicating the problem.

3. PROPOSED METHOD

This section describes the new algorithm for inferring the Bayesian linear dynamical system's latent statistics. It involves a forward and backward pass that may be considered as the Bayesian counterparts to the Kalman filter and RTS smoother. A full derivation of the new filter and smoother is provided in the supplemental material [16].

We propose a solution by decomposing problematic second moment terms like $\langle \mathbf{A}^\mathsf{T}\mathbf{Q}^{-1}\mathbf{B} \rangle$ into sums of first moments and covariances, for example $\langle \mathbf{A} \rangle^\mathsf{T} \langle \mathbf{Q}^{-1} \rangle \langle \mathbf{B} \rangle + \boldsymbol{\Sigma}_{AB}$ where

$$\boldsymbol{\Sigma}_{AB} = \sum_{i=1}^{H} \sum_{j=1}^{H} \langle Q_{ij}^{-1} \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \mathrm{cov}[\mathbf{a}_{(i)}, \mathbf{b}_{(j)}]_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \tag{12}$$

This general expression applies to $\boldsymbol{\Sigma}_{AA}$, $\boldsymbol{\Sigma}_{AB}$, $\boldsymbol{\Sigma}_{CC}$ and $\boldsymbol{\Sigma}_{CD}$. These covariances encode parameter uncertainty. In practice this expression simplifies. For example, a common assumption is that rows are independent and the noise covariance is diagonal [1,13,17]. This decomposition holds for arbitrary priors, since non-conjugate priors will require an induced variational factorization $q_{\boldsymbol{\theta}}(\mathbf{A}, \mathbf{B}) q_{\boldsymbol{\theta}}(\mathbf{Q})$, and conjugate priors require that $\mathbf{Q}$ is diagonal, each element scaling a row of $\mathbf{A}$ and $\mathbf{B}$. In the following, we drop the subscript $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ from the parameter expectations for notational convenience.

3.1. Forward Pass (Filtering)

The forward pass calculates the mean $\boldsymbol{\mu}_n$ and covariance $\mathbf{V}_n$ of the marginal probability, $\forall n \in (1, N)$:

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n}) = \frac{p(\mathbf{y}_n|\mathbf{x}_n) q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1})}{q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1})} \tag{13}$$
$$= \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_n, \mathbf{V}_n) \tag{14}$$

The predictive distribution is given by

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1}) = \mathcal{N}(\mathbf{x}_n|\mathbf{m}_{n-1}, \mathbf{P}_{n-1}) \tag{15}$$

with mean and covariance

$$\mathbf{m}_{n-1} = \langle \mathbf{A} \rangle \overline{\boldsymbol{\mu}}_{n-1} + \langle \mathbf{B} \rangle \mathbf{u}_n - \langle \mathbf{A} \rangle \overline{\mathbf{V}}_{n-1} \boldsymbol{\Sigma}_{AB} \mathbf{u}_n \tag{16}$$
$$\mathbf{P}_{n-1} = \langle \mathbf{A} \rangle \overline{\mathbf{V}}_{n-1} \langle \mathbf{A} \rangle^\mathsf{T} + \langle \mathbf{Q} \rangle \tag{17}$$

The filtered output probability is given by

$$q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1}) = \mathcal{N}(\mathbf{y}_n|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \tag{18}$$

with mean and covariance

$$\boldsymbol{\mu}_y = \langle \mathbf{C} \rangle \overline{\mathbf{m}}_{n-1} + \langle \mathbf{D} \rangle \mathbf{u}_n - \langle \mathbf{C} \rangle \overline{\mathbf{P}}_{n-1} \boldsymbol{\Sigma}_{CD} \mathbf{u}_n \tag{19}$$
$$\boldsymbol{\Sigma}_y = \langle \mathbf{C} \rangle \overline{\mathbf{P}}_{n-1} \langle \mathbf{C} \rangle^\mathsf{T} + \langle \mathbf{R} \rangle \tag{20}$$

Finally, the updated state mean and covariance are as follows:

$$\mathbf{K} = \overline{\mathbf{P}}_{n-1} \langle \mathbf{C} \rangle^\mathsf{T} \boldsymbol{\Sigma}_y^{-1} \tag{21}$$
$$\boldsymbol{\mu}_n = \overline{\mathbf{m}}_{n-1} + \mathbf{K}\left(\mathbf{y}_n - \boldsymbol{\mu}_y\right) - \overline{\mathbf{P}}_{n-1} \boldsymbol{\Sigma}_{CD} \mathbf{u}_n \tag{22}$$
$$\mathbf{V}_n = \left(\mathbf{I} - \mathbf{K}\langle \mathbf{C} \rangle\right) \overline{\mathbf{P}}_{n-1} \tag{23}$$

where $\mathbf{K}$ is called the Kalman gain and

$$\mathbf{G} = \mathbf{I} - \mathbf{V}_{n-1}\left(\mathbf{I} + \boldsymbol{\Sigma}_{AA}\mathbf{V}_{n-1}\right)^{-1} \boldsymbol{\Sigma}_{AA} \tag{24}$$
$$\mathbf{L} = \mathbf{I} - \mathbf{P}_{n-1}\left(\mathbf{I} + \boldsymbol{\Sigma}_{CC}\mathbf{P}_{n-1}\right)^{-1} \boldsymbol{\Sigma}_{CC} \tag{25}$$

$$\overline{\boldsymbol{\mu}}_{n-1} = \mathbf{G}\boldsymbol{\mu}_{n-1}, \qquad \overline{\mathbf{V}}_{n-1} = \mathbf{G}\mathbf{V}_{n-1} \tag{26}$$
$$\overline{\mathbf{m}}_{n-1} = \mathbf{L}\mathbf{m}_{n-1}, \qquad \overline{\mathbf{P}}_{n-1} = \mathbf{L}\mathbf{P}_{n-1} \tag{27}$$

The forward pass is initialized at time $n = 1$ with $\overline{\mathbf{m}}_{n-1} = \mathbf{m}_0$ and $\overline{\mathbf{P}}_{n-1} = \mathbf{P}_0$. Pleasingly, when the parameter covariances are zero ($\boldsymbol{\Sigma}_{AA} = \mathbf{0}$, etc.), these equations exactly match the Kalman filter.

3.2. Backward Pass (Smoothing)

The backward pass calculates the mean $\widehat{\boldsymbol{\mu}}_n$ and covariance $\widehat{\mathbf{V}}_n$ of the marginal posterior

$$q_{\mathbf{x}}(\mathbf{x}_n) = q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n}) \int \frac{p(\mathbf{x}_{n+1}|\mathbf{x}_n) q_{\mathbf{x}}(\mathbf{x}_{n+1})}{q_{\mathbf{x}}(\mathbf{x}_{n+1}|\mathbf{y}_{1:n})} d\mathbf{x}_{n+1} \quad (28)$$

$$= \mathcal{N}(\mathbf{x}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n) \quad (29)$$

After initializing the statistics at time $n = N$ with $\widehat{\boldsymbol{\mu}}_N = \boldsymbol{\mu}_N$ and $\widehat{\mathbf{V}}_N = \mathbf{V}_N$, the mean and covariance are propagated analytically from time $n = N-1$ back to $n = 1$ with the following equations

$$\mathbf{J}_n = \overline{\mathbf{V}}_n \langle \mathbf{A} \rangle^{\mathsf{T}} \mathbf{P}_n^{-1} \quad (30)$$

$$\widehat{\boldsymbol{\mu}}_n = \overline{\boldsymbol{\mu}}_n + \mathbf{J}_n \left(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n \right) - \overline{\mathbf{V}}_n \boldsymbol{\Sigma}_{AB} \mathbf{u}_{n+1} \quad (31)$$

$$\widehat{\mathbf{V}}_n = \overline{\mathbf{V}}_n + \mathbf{J}_n \left(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n \right) \mathbf{J}_n^{\mathsf{T}}. \quad (32)$$

When the covariances of the parameters are zero, these equations exactly match the RTS smoother. Finally, the following expected sufficient statistics are required for updating $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$\langle \mathbf{x}_n \rangle = \widehat{\boldsymbol{\mu}}_n \quad (33)$$

$$\langle \mathbf{x}_n \mathbf{x}_n^{\mathsf{T}} \rangle = \widehat{\boldsymbol{\mu}}_n \widehat{\boldsymbol{\mu}}_n^{\mathsf{T}} + \widehat{\mathbf{V}}_n \quad (34)$$

$$\langle \mathbf{x}_n \mathbf{x}_{n+1}^{\mathsf{T}} \rangle = \widehat{\boldsymbol{\mu}}_n \widehat{\boldsymbol{\mu}}_{n+1}^{\mathsf{T}} + \mathbf{J}_n \widehat{\mathbf{V}}_{n+1} \quad (35)$$

3.3. Lower bound evaluation

The log partition function $\ln q_{\mathbf{x}}(\mathbf{Y})$ is an optional quantity to compute during the forward pass that is useful for evaluating $\mathcal{L}(q)$ [1].

$$\begin{aligned} \ln q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1}) = \frac{1}{2}\Big(& - \langle \ln|2\pi\mathbf{R}| \rangle - \langle \ln|\mathbf{Q}| \rangle \\ & - \ln|\mathbf{V}_{n-1}| + \ln|\boldsymbol{\Xi}| + \ln|\mathbf{V}_n| + \boldsymbol{\xi}_{n-1}^{\mathsf{T}} \boldsymbol{\Xi}_{n-1}^{-1} \boldsymbol{\xi}_{n-1} \\ & - \mathbf{y}_n^{\mathsf{T}} \langle \mathbf{R}^{-1} \rangle \mathbf{y}_n + 2 \mathbf{y}_n^{\mathsf{T}} \langle \mathbf{R}^{-1} \mathbf{D} \rangle \mathbf{u}_n - \boldsymbol{\mu}_{n-1}^{\mathsf{T}} \mathbf{V}_{n-1}^{-1} \boldsymbol{\mu}_{n-1} \\ & - \mathbf{u}_n^{\mathsf{T}} \big(\langle \mathbf{B}^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{B} \rangle + \langle \mathbf{D}^{\mathsf{T}} \mathbf{R}^{-1} \mathbf{D} \rangle \big) \mathbf{u}_n + \boldsymbol{\mu}_n^{\mathsf{T}} \mathbf{V}_n^{-1} \boldsymbol{\mu}_n \Big) \end{aligned} \quad (36)$$

where $\mathbf{W} = \mathbf{I} - \mathbf{J}_n \langle \mathbf{A} \rangle$ and

$$\boldsymbol{\xi}_{n-1} = \mathbf{W} \big(\overline{\boldsymbol{\mu}}_{n-1} - \overline{\mathbf{V}}_{n-1} \boldsymbol{\Sigma}_{AB} \mathbf{u}_n \big) - \mathbf{J}_n \langle \mathbf{B} \rangle \mathbf{u}_n \quad (37)$$

$$\boldsymbol{\Xi}_{n-1} = \mathbf{W} \overline{\mathbf{V}}_{n-1} \quad (38)$$

The log partition function is $\ln q_{\mathbf{x}}(\mathbf{Y}) = \sum_{n=1}^{N} \ln q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1})$.

4. VALIDATION

Experiments were conducted to validate the accuracy of the proposed and existing filtering and smoothing methods. The KL divergence from the ground truth distribution $q_{\mathbf{x}}(\mathbf{X})$ provided by belief propagation [1] to $\widehat{q}_{\mathbf{x}}(\mathbf{X})$ provided by the proposed and existing methods was calculated for a range of parameter covariances, embedded in the model through $\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_C = \mathbf{I}\sigma$, where the variance σ took values in the range $(10^{-10}, 10^{10})$. The KL divergence was averaged over 100 Monte Carlo runs for each variance setting. Figure 1 shows that the augmented and non-augmented versions that rely on Cholesky factors are in fact not numerically equivalent to the original belief propagation version once the parameter variance exceeds approximately 10^{-7}. The KL divergence levels off for variances larger than 10^4, likely because the model is saturated with parameter uncertainty. The proposed method proves to be mathematically consistent with belief propagation, having a consistently small KL divergence regardless of the amount of propagated parameter uncertainty.

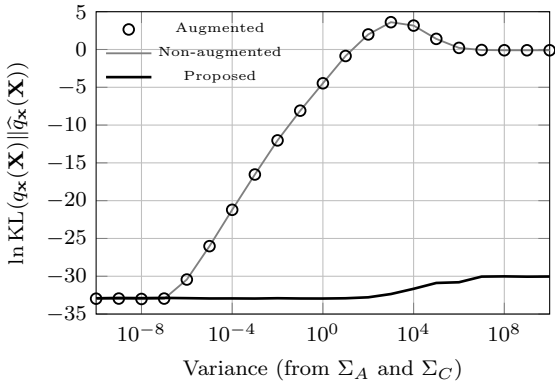

Fig. 1. Log KL divergence from $q_{\mathbf{x}}(\mathbf{X})$ provided by belief propagation [1] to $\widehat{q}_{\mathbf{x}}(\mathbf{X})$ provided by the proposed and existing augmented/non-augmented methods [18].

5. APPLICATION: BAYESIAN FREQUENCY ESTIMATION

In this section, variational inference of a Bayesian linear dynamical system is applied to the problem of frequency estimation. System matrices and initial conditions are designed to reflect a sum of sinusoids model. Inferring the fully Bayesian model provides high-resolution frequency estimations that are accompanied by measures of uncertainty about the inferred values. Moreover, the level of noise present in the signal is estimated.

5.1. System structure

This section describes a linear dynamical system that generates a univariate observation y_n from a sum of K noisy sinusoids:

$$y_n = \sum_{k=1}^{K} g_k \sin(2\pi f_k n T + \phi_k) + \eta_n^y \quad (39)$$

where the kth sinusoid has amplitude g_k, frequency f_k in Hz, and initial phase ϕ_k in radians.

The system matrix $\mathbf{A}$ is block diagonal. Each 2×2 block $\mathbf{A}_k$ is parametrized by a scalar $\nu_k \in (-2, 0)$ that controls the frequency of the kth latent oscillation,

$$\mathbf{A}_k = \mathbf{F} + \mathbf{E}\nu_k \quad (40)$$

where $\mathbf{F}$ and $\mathbf{E}$ are constants given by

$$\mathbf{F} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{E} = \begin{pmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{pmatrix} \quad (41)$$

Indeed, the elements of the 2×1 state subvector $\mathbf{x}_{nk}$ will oscillate at $f_k = \arccos(a_k + 1)/(2\pi T)$ Hz, where the sampling period $T = 1/Fs$ seconds. The initial mean $\mathbf{m}_0$ over initial state $\mathbf{x}_{1k}$ sets the amplitude g_k and initial phase ϕ_k:

$$\mathbf{m}_{0k} = g_k \begin{pmatrix} \sin(\phi_k) \\ 2\cos(\phi_k)\tan(\pi f_k T) \end{pmatrix} \quad (42)$$

and the initial covariance is $\mathbf{P}_0 = \mathbf{I}\tau_0^{-1}$. The output matrix $\mathbf{C} = (1, 0, 1, 0, \ldots)$ sums the first dimension of each subvector. The latent noise covariance $\mathbf{Q}$ is block diagonal. Each block is parametrized by a scalar precision τ_k, $\mathbf{Q}_k = \mathbf{I}\tau_k^{-1}$. The output noise covariance is

parametrized by a scalar precision $R = \rho^{-1}$. The prior distribution $p(\boldsymbol{\theta}) = p(\boldsymbol{\nu}, \boldsymbol{\tau})p(\rho)$ where

$$p(\boldsymbol{\nu}, \boldsymbol{\tau}) = \prod_{k=1}^{K} \mathcal{N}(\nu_k | 0, \alpha_k^{-1}\tau_k^{-1}) \text{Gam}(\tau_k | e_0, i_0) \quad (43)$$

$$p(\rho) = \text{Gam}\big(\rho | r_0, s_0\big) \quad (44)$$

Using the Normal-Gamma distribution is beneficial for reducing the number of induced variational factorizations.

5.2. Variational M step

The variational M step computes the optimal factor $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = q_{\boldsymbol{\nu}}(\boldsymbol{\nu}|\boldsymbol{\tau})q_{\boldsymbol{\tau}}(\boldsymbol{\tau})q_{\rho}(\rho)$ where

$$q_{\boldsymbol{\nu}}(\boldsymbol{\nu}|\boldsymbol{\tau})q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=1}^{K} \mathcal{N}(\nu_k | \widehat{\nu}_k, \sigma_k\tau_k^{-1}) \text{Gam}(\tau_k | \widehat{e}_k, \widehat{g}_k) \quad (45)$$

$$q_{\rho}(\rho) = \text{Gam}(\rho | \widehat{r}, \widehat{s}) \quad (46)$$

The moments of these optimal distributions are as follows.

$$\sigma_k = \left(\text{Tr}\left(\mathbf{E}\textstyle\sum_{n=2}^{N} \langle \mathbf{x}_{n-1k}\mathbf{x}_{n-1k}^{\mathrm{T}} \rangle \mathbf{E}^{\mathrm{T}}\right) + \alpha_k\right)^{-1} \quad (47)$$

$$\widehat{\nu}_k = \sigma_k \text{Tr}\left(\mathbf{E}\left(\textstyle\sum_{n=2}^{N} \langle \mathbf{x}_{n-1k}\mathbf{x}_{nk}^{\mathrm{T}} \rangle - \sum_{n=2}^{N} \langle \mathbf{x}_{n-1k}\mathbf{x}_{n-1k}^{\mathrm{T}} \rangle \mathbf{F}^{\mathrm{T}}\right)\right) \quad (48)$$

$$\widehat{e}_k = e_0 + N - 1 \quad (49)$$

$$\widehat{i}_k = i_0 - \frac{1}{2}\nu_k^2\sigma_k^{-1} + \frac{1}{2}\text{Tr}\Big(\mathbf{F}\textstyle\sum_{n=2}^{N} \langle \mathbf{x}_{n-1k}\mathbf{x}_{n-1k}^{\mathrm{T}} \rangle \mathbf{F}^{\mathrm{T}} - 2\mathbf{F}\sum_{n=2}^{N} \langle \mathbf{x}_{n-1k}\mathbf{x}_{nk}^{\mathrm{T}} \rangle + \sum_{n=2}^{N} \langle \mathbf{x}_{nk}\mathbf{x}_{nk}^{\mathrm{T}} \rangle\Big) \quad (50)$$

$$\widehat{r} = r_0 + N/2 \quad (51)$$

$$\widehat{s} = s_0 + \frac{1}{2}\Big(\textstyle\sum_{n=1}^{N} \mathbf{y}_n\mathbf{y}_n^{\mathrm{T}} - 2\mathbf{C}\sum_{n=1}^{N} \langle \mathbf{x}_n \rangle \mathbf{y}_n^{\mathrm{T}} + \mathbf{C}\sum_{n=1}^{N} \langle \mathbf{x}_n\mathbf{x}_n^{\mathrm{T}} \rangle \mathbf{C}^{\mathrm{T}}\Big) \quad (52)$$

The following statistics are needed for the E step: $\langle \nu_k \rangle = \widehat{\nu}_k$, $\langle \tau_k^{-1} \rangle = \widehat{i}_k/\widehat{e}_k$, and $\langle \rho^{-1} \rangle = \widehat{s}/\widehat{r}$.

The initial state's prior mean and covariance are simply updated to their maximum likelihood values: $\mathbf{m}_0 = \widehat{\boldsymbol{\mu}}_1$ and $\mathbf{P}_0 = \widehat{\mathbf{V}}_1$. Alternatively, they can be given priors and included in $q_{\boldsymbol{\theta}}$.

5.3. Variational E step

The proposed algorithm in Section 3.1-3.2 is used to calculate the sufficient statistics of $q_{\mathbf{x}}(\mathbf{X})$. Using the results of the M step:

$$\langle \mathbf{A}_k \rangle = \mathbf{F} + \mathbf{E}\langle \nu_k \rangle, \quad \langle \mathbf{Q}_k \rangle = \mathbf{I}\langle \tau_k^{-1} \rangle, \quad \langle R \rangle = \langle \rho^{-1} \rangle \quad (53)$$

Covariance $\boldsymbol{\Sigma}_{AA}$ is block diagonal. The kth block is $\mathbf{E}^{\mathrm{T}}\mathbf{E}\sigma_k$. Since $\mathbf{C}$ is deterministic, $\boldsymbol{\Sigma}_{CC}$ is zero.

5.4. Results

The proposed Bayesian frequency estimator (BFE) was tested alongside three existing sinusoidal model estimators: the Reassignment method (RM) [19], the Derivative method (DM) [20], and a high-resolution subspace method (ESPRIT) [21].

A test signal comprised of a sinusoid with additive zero-mean Gaussian noise sampled at $Fs = 44100Hz$, sample length $N = 63$, and the signal-to-noise ratio (SNR) expressed in dB as $10\log_{10}(g^2/R)$, and ranged from -20dB to +120dB by increments of 20dB, with amplitude $g = 1$ of the sinusoid. For each SNR and analysis method, we tested $M = 1000$ frequencies (f) linearly distributed in the $(Fs/N,\ Fs/4)$ range, with uniformly sampled phases $\phi \in (-\pi, \pi)$, and computed the variance of estimation error $\frac{(2\pi T)^2}{M-1}\sum_{i=1}^{M}(\hat{f}_i - f_i)^2$.

The performance of each estimator was compared to the Cramér-Rao lower bound (CRLB) given in [22], defining the limit of the best possible variance of estimation error achievable by an unbiased estimator for a particular sample length and noise level. Results are shown in Figure 2. The Bayesian frequency estimator achieves close to the same resolution as ESPRIT, outperforming the phase-based RM and DM. The second test involved estimating two superimposed sinusoids of the same amplitude with frequencies spaced randomly within $2Fs/N$ Hz of each other, for which ESPRIT has the best accuracy and the Bayesian method is intermediate. While more computationally intensive, the Bayesian frequency estimator quantifies uncertainty about the frequency estimation, simultaneously estimates the observed noise variance, and can prune superfluous sinusoids from the model.

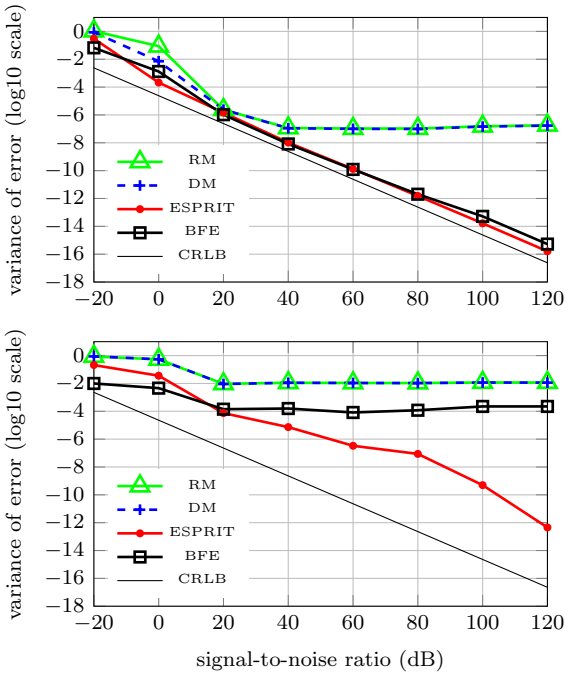

Fig. 2. Estimation of one (above) and two (below) frequencies.

6. CONCLUSION

A new algorithm was presented for inferring the hidden state sequence of a variational Bayesian linear dynamical system, which can be seen as the Bayesian extension to the Kalman filter and smoother. It is more numerically stable than existing routines, respects the statistical moments of the parameters, and enjoys a cost that scales linearly with the data sequence length. This work is applicable to variational inference of Bayesian linear dynamical systems and its extensions, for example, recurrent switching linear dynamical systems [23]. More generally, the proposed Bayesian filter and smoother is useful for embedding uncertainty about parameters into a dynamical model. In future work, we will extend the Bayesian frequency estimation to probabilistic time-frequency analysis and apply the new algorithm to learn switching dynamical models.

7. REFERENCES

[1] M. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, University College London, May 2003.

[2] David Barber, A. Cemgil, and Silvia Chiappa, Eds., *Bayesian Time Series Models*, Cambridge University Press, 2011.

[3] S. Chiappa, "A Bayesian approach to switching linear Gaussian state-space models for unsupervised time-series segmentation," in *International Conference on Machine Learning and Applications*. IEEE Computer Society, Dec. 2008, pp. 3–9.

[4] T. Ardeshiri, E. Özkan, U. Orguner, and F. Gustafsson, "Approximate bayesian smoothing with unknown process and measurement noise covariances," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2450–2454, 2015.

[5] Binbin Li and Armen Der Kiureghian, "Operational modal identification using variational bayes," *Mechanical Systems and Signal Processing*, vol. 88, pp. 377–398, 2017.

[6] Antonio Salmeron, Rafael Rum, Helge Langseth, Thomas D. Nielsen, and Anders L. Madsen, "A review of inference algorithms for hybrid Bayesian networks," *Journal of Artifical Intelligence Research*, vol. 62, pp. 799–828, 2018.

[7] Sezen Cekic, Didier Grandjean, and Olivier Renaud, "Multiscale Bayesian state-space model for Granger causality analysis of brain signal," *Journal of Applied Statistics*, vol. 46, no. 1, pp. 66–84, 2019.

[8] Charul, U. Bhatt, P. Biyani, and K. Rajawat, "Online variational Bayesian subspace filtering," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2019, pp. 5057–5061.

[9] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[10] R. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the American Society for Mechanical Engineering, Series D, Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[11] H. Rauch, F. Tung, and C. Striebel, "Maximum likelihood estimates of linear dynamical systems," *AIAA Journal*, vol. 3, pp. 1445–1450, 1965.

[12] M. Verhaegen and P. Van Dooren, "Numerical aspects of different Kalman filter implementations," *IEEE Transactions on Automatic Control*, vol. 31, pp. 901–917, 1986.

[13] J. Luttinen, "Fast variational Bayesian linear state-space model," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Heidelberg, 2013, pp. 305–320, Springer, Berlin.

[14] R. Eubank and Soujin Wang, "The equivalence between the Cholesky decomposition and the Kalman filter," *The American Statistician*, vol. 56, no. 1, pp. 39–43, Feb. 2002.

[15] Catherine Forbes, Merran Evans, Nicholas Hastings, and Brian Peacock, *Statistical Distributions*, John Wiley & Sons, Inc., 4th edition, 2011.

[16] J. Neri, P. Depalle, and R. Badeau, "Supplemental material," Tech. Rep., McGill University, 2019.

[17] D. Barber and S. Chiappa, "Unified inference for variational bayesian linear gaussian state-space models," in *Advances in Neural Information Processing Systems 19*. 2006, pp. 81–88, MIT Press.

[18] D. Barber, "Expectation correction for smoothed inference in switching linear dynamical systems," *Journal of Machine Learning Research*, vol. 7, pp. 2515–2540, Nov. 2006.

[19] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. Signal Process.*, vol. 43, pp. 1068–1089, 1995.

[20] S. Marchand and P. Depalle, "Generalization of the derivative analysis method to non-stationary sinusoidal modeling," in *Proc. of the 11th Int. Conf. on Digital Audio Effects (DAFx-08)*, Espoo, Finland, September 2008.

[21] Roland Badeau, B. David, and G. Richard, "High-resolution spectral analysis of mixtures of complex exponentials modulated by polynomials," *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1341–1350, April 2006.

[22] Guotong Zhou, Georgios Giannakis, and Ananthram Swami, "On polynomial phase signal with time-varying amplitudes," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 44, no. 4, pp. 848–860, 1996.

[23] S. Linderman, M. Johnson, A. Miller, R. Adams, D. Blei, and L. Paninski, "Bayesian learning and inference in recurrent switching linear dynamical systems," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017, vol. 54, pp. 914–922.