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

Julian Neri, Roland Badeau, Philippe Depalle

To cite this version:
Julian Neri, Roland Badeau, Philippe Depalle. Probabilistic filter and smoother for variational inference of Bayesian linear dynamical systems. 45th International Conference on Acoustics, Speech, and Signal Processing, May 2020, Barcelona, Spain. hal-02456651

HAL Id: hal-02456651
https://telecom-paris.hal.science/hal-02456651
Submitted on 17 May 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$; $I$ is the identity matrix; $|\mathbf{A}|$ is the determinant and $\text{Tr}(\mathbf{A})$ is the trace of matrix $\mathbf{A}$; $\text{diag}(\mathbf{a})$ denotes a diagonal matrix formed from the elements of $\mathbf{a}$; column vector $\mathbf{a}^{(i)}$ denotes the transposed $i$th row of $\mathbf{A}$; $\langle x \rangle$ denotes the expected value of $x$ and covariance $\text{cov}[x, y] = \langle xy^T \rangle - \langle x \rangle \langle y \rangle^T$; $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ is a multivariate Normal distribution with mean $\mu$ and covariance $\Sigma$; $\text{Gam}(x|a, b)$ 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 $\eta^u_n \sim \mathcal{N}(0, \mathbf{Q})$ and observation $\eta^v_n \sim \mathcal{N}(0, \mathbf{R})$, where positive semi-definite covariance matrix $\mathbf{Q}$ is $H \times H$ and $\mathbf{R}$ is $V \times V$. The linear dynamical
system has the following state space representation:

\[
x_n = Ax_{n-1} + Bu_n + \eta^x_n \\
y_n = Cx_n + Du_n + \eta^y_n
\]  

(1)

(2)

Initial state \( x_1 \) is Gaussian-distributed with \( H \times 1 \) mean \( m_0 \) and \( H \times H \) covariance \( P_0 \). The joint probability of the linear dynamical system is

\[
p(Y, X) = p(y_1|x_1)p(x_1)\prod_{n=2}^{N}p(y_n|x_n)p(x_n|x_{n-1})
\]  

(3)

where the probability of transition \( p(x_n|x_{n-1}) \), emission \( p(y_n|x_n) \), and initial state \( p(x_1) \) are Gaussian (due to additive Gaussian noise):

\[
p(x_1) = \mathcal{N}(x_1|m_0, P_0)
\]  

(4)

\[
p(x_n|x_{n-1}) = \mathcal{N}(x_n|Ax_{n-1} + Bu_n, Q)
\]  

(5)

\[
p(y_n|x_n) = \mathcal{N}(y_n|Cx_n + Du_n, R).
\]  

(6)

Bayesian linear dynamical systems assume a single factorization between variables and parameters: \( q \) is diagonal, each element scalable by \( \alpha \) (see Section 5). 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 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 \( (A^T Q A)^{-1} \) into sums of first moments and covariances, for example \( (A^T Q A)^{-1} = B + \Sigma_{AB} \) where

\[
\Sigma_{AB} = \sum_{t=1}^{H} \sum_{j=1}^{H} (Q_{ij})^{-1} q_y(\theta_i) \text{cov}[\theta_i, \theta_j] q_y(\theta)
\]  

(12)

This general expression applies to \( \Sigma_{AA}, \Sigma_{AB}, \Sigma_{CC} \) and \( \Sigma_{CD} \). 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 \( q_y(A, B|\varphi(Q)) \), and conjugate priors require that \( Q \) is diagonal, each element scaling a row of \( A \) and \( B \). In the following, we drop the subscript \( y \) from the parameter expectations for notational convenience.

3. PROPOSED METHOD

3.1. Forward Pass (Filtering)

The forward pass calculates the mean \( \mu_n \) and covariance \( V_n \) of the marginal probability, \( \forall n \in (1, N) \):

\[
q_x(x_n | y_{1:n-1}) = \mathcal{N}(x_n | \mu_{n-1}, V_{n-1})
\]  

(13)

\[
q_x(x_n | y_{1:n-1}) = \mathcal{N}(x_n | \mu_{n-1}, V_{n-1})
\]  

(14)

The predictive distribution is given by

\[
q_x(x_{n+1} | y_{1:n}) = \mathcal{N}(x_{n+1}|m_{n-1}, P_{n-1})
\]  

(15)

with mean and covariance

\[
m_{n-1} = \langle A \rangle m_{n-1} + \langle B \rangle u_n - \langle A \rangle \bar{V}_{n-1} \Sigma_{AB} u_n
\]  

(16)

\[
P_{n-1} = \langle A \rangle \bar{V}_{n-1}^T + \langle Q \rangle
\]  

(17)

The filtered output probability is given by

\[
q_x(y_n | y_{1:n-1}) = \mathcal{N}(y_n | \mu_y, \Sigma_y)
\]  

(18)

with mean and covariance

\[
\mu_y = \langle C \rangle m_{n-1} + \langle D \rangle u_n - \langle C \rangle \bar{V}_{n-1} \Sigma_{CD} u_n
\]  

(19)

\[
\Sigma_y = \langle C \rangle \bar{P}_{n-1}^T \langle C \rangle + \langle R \rangle
\]  

(20)

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

\[
K = \bar{P}_{n-1} \langle C \rangle \Sigma_y^{-1}
\]  

(21)

\[
\mu_n = \bar{m}_{n-1} + K (y_n - \mu_y) - \bar{P}_{n-1} \Sigma_{CD} u_n
\]  

(22)

\[
V_n = (I - K \langle C \rangle) \bar{P}_{n-1}
\]  

(23)

where \( K \) is called the Kalman gain and

\[
G = I - V_{n-1} (I + \Sigma_{AA} V_{n-1})^{-1} \Sigma_{AA}
\]  

(24)

\[
L = I - P_{n-1} (I + \Sigma_{CC} P_{n-1})^{-1} \Sigma_{CC}
\]  

(25)

\[
\bar{P}_{n-1} = G \mu_{n-1}, \quad \bar{V}_{n-1} = G V_{n-1}
\]  

(26)

\[
\bar{m}_{n-1} = L m_{n-1}, \quad \bar{P}_{n-1} = L P_{n-1}
\]  

(27)

The forward pass is initialized at time \( n = 1 \) with \( \bar{m}_{n-1} = m_0 \) and \( \bar{P}_{n-1} = P_0 \). Please, when the parameter covariances are zero (\( \Sigma_{AA} = 0 \), etc.), these equations exactly match the Kalman filter.
3.2. Backward Pass (Smoothing)

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

$$q_x(x_n) = q_x(x_n|y_{1:n}) \int p(x_{n+1}|x_n)q_x(x_{n+1}) \, dx_{n+1}$$

$$= \mathcal{N}(x_n|\tilde{\mu}_n, \tilde{V}_n)$$

We now formalize these results mathematically. From time $n = N$ to $n = 1$, the mean and covariance are propagated according to the following equations:

$$\tilde{\mu}_n = \mathbf{p}_n + J_n (\tilde{\mu}_{n+1} - \mathbf{m}_n) - V_n \Sigma_{AB} u_{n+1}$$

$$\tilde{V}_n = V_n + J_n (\tilde{V}_{n+1} - P_n) J_n^T$$

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 $q_\theta(\theta)$:

$$\langle x_n \rangle = \tilde{\mu}_n$$

$$\langle x_n x_n^T \rangle = \tilde{\mu}_n \tilde{\mu}_n^T + \tilde{V}_n$$

$$\langle x_n x_{n+1} \rangle = \tilde{\mu}_n \tilde{\mu}_{n+1} + J_n \tilde{V}_{n+1}$$

3.3. Lower bound evaluation

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

$$\ln q_x(y_{1:n}|y_{1:n-1}) = \frac{1}{2} \left( - \langle \ln |2\pi R| \rangle - \langle \ln |Q| \rangle 
- \ln |V_{n-1}| + \ln |\Xi| + \ln |V_n| + \epsilon_{n-1}^T \Xi^{-1} \epsilon_{n-1}
- y_n^T (R^{-1}) y_n + 2 y_n^T (R^{-1} D) u_n - \epsilon_{n-1}^T V_{n-1} \epsilon_{n-1} - u_n^T (D^T R^{-1} D) u_n + \mu_n^T V_{n-1} \mu_n \right)$$

$$W = I - J_n (A)$$

$$\Xi_{n-1} = W \Xi_{n-1}^T$$

$$\epsilon_{n-1} = W \epsilon_{n-1}^T$$

The log partition function is $\ln q_x(Y) = \sum_{n=1}^N \ln q_x(y_n|y_{1:n-1})$.

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 $q_κ(X)$ provided by belief propagation [1] to $\tilde{q}_κ(X)$ provided by the proposed and existing methods was calculated for a range of parameter covariances, embedded in the model through $\Sigma_A = \Sigma_C = I\sigma^2$, where the variance $\sigma$ took values in the range $(10^{-10}, 10^{0})$. 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 $10^{-5}$. 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.

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. 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:

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

where the $k$th sinusoid has amplitude $g_k$, frequency $f_k$ in Hz, and initial phase $\phi_k$ in radians.

The system matrix $A$ is block diagonal. Each $2 \times 2$ block $A_k$ is parametrized by a scalar $\nu_k \in (-2, 0)$ that controls the frequency of the $k$th latent oscillation,

$$A_k = F + E \nu_k$$

where $F$ and $E$ are constants given by

$$F = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad E = \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}$$

Indeed, the elements of the $2 \times 1$ state subvector $x_{nk}$ will oscillate at $f_k = \arccos(\nu_k + 1)/(2\pi T)$ Hz, where the sampling period $T = 1/F_s$ seconds. The initial mean $m_0$ over initial state $x_{nk}$ sets the amplitude $g_k$ and initial phase $\phi_k$:

$$m_{0k} = g_k \begin{pmatrix} \sin(\phi_k) \\ 2 \cos(\phi_k) \tan(\pi f_k T) \end{pmatrix}$$

and the initial covariance is $P_0 = I\tau_0^{-1}$. The output matrix $C = (1, 0, 1, 0, \ldots)$ sums the first dimension of each subvector. The latent noise covariance $Q$ is block diagonal. Each block is parametrized by a scalar precision $\tau_k$, $Q_k = I\tau_k^{-1}$. The output noise covariance is
parametrized by a scalar precision \( R = \rho^{-1} \). The prior distribution
\( p(\theta) = p(\nu, \tau) p(\rho) \) where
\[
p(\nu, \tau) = \prod_{k=1}^{K} \mathcal{N}(\nu_k|0, \alpha_k^{-1} \tau_k^{-1}) \Gamma(\tau_k|c_0, i_o) \tag{43}
p(\rho) = \Gamma(\rho|\tau_0, s_0) \tag{44}
\]
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 \( q_\theta(\theta) = q_\theta(\nu|\tau) q_\tau(\tau) q_\rho(\rho) \) where
\[
q_\nu(\nu|\tau) q_\tau(\tau) = \prod_{k=1}^{K} \mathcal{N}(\nu_k|\tilde{\nu}_k, \sigma_k \tau_k^{-1}) \Gamma(\tau_k|\tilde{\tau}_k, \tilde{\gamma}_k) \tag{45}
q_\rho(\rho) = \Gamma(\rho|\tilde{\rho}, \tilde{s}) \tag{46}
\]
The moments of these optimal distributions are as follows.
\[
\sigma_k = \left( \text{Tr} \left( E \sum_{n=2}^{N} (x_{n-1}x_{n-1}^T) E^T \right) + \alpha_k \right)^{-1} \tag{47}
\tilde{\nu}_k = \sigma_k \text{Tr} \left( E \sum_{n=2}^{N} (x_{n-1}x_{n-1}^T) - \sum_{n=2}^{N} (x_{n-1}x_{n-1}^T) F^T \right) \tag{48}
\tilde{\sigma}_k = \epsilon_0 + N - 1 \tag{49}
\hat{\tau}_k = i_0 - \frac{1}{2} \nu_k^2 \sigma_k^{-1} + \frac{1}{2} \text{Tr} \left( F \sum_{n=2}^{N} (x_{n-1}x_{n-1}^T) F^T \right) - 2F \sum_{n=2}^{N} (x_{n-1}x_{n-1}^T) + \sum_{n=2}^{N} (x_{n-1}x_{n-1}^T) \tag{50}
\hat{\rho} = \tau_0 + N/2 \tag{51}
\hat{s} = s_0 + \frac{1}{2} \sum_{n=1}^{N} Y_n Y_n^T - 2C \sum_{n=1}^{N} (x_n y_n^T) + C \sum_{n=1}^{N} (x_n x_n^T) C^T \tag{52}
\]
The following statistics are needed for the E step: \( \langle \nu_k \rangle = \tilde{\nu}_k \), \( \langle \nu_k^{-1} \rangle = \tilde{\tau}_k^2 / \tilde{\sigma}_k \), and \( \langle \rho^{-1} \rangle = \hat{s} / \hat{F} \).

The initial state’s prior mean and covariance are simply updated to their maximum likelihood values: \( \mu_0 = \tilde{\mu}_1 \) and \( \Sigma_0 = \tilde{\Sigma}_1 \). Alternatively, they can be given priors and included in \( q_\theta \).

5.3. Variational E step
The proposed algorithm in Section 3.1-3.2 is used to calculate the sufficient statistics of \( q_\theta(X) \). Using the results of the M step:
\[
\langle A_k \rangle = F + E \langle \nu_k \rangle, \quad \langle Q_k \rangle = I / \langle \tau_k^{-1} \rangle, \quad \langle R \rangle = \langle \rho^{-1} \rangle \tag{53}
\]
Covariance \( \Sigma_{AA} \) is block diagonal. The \( k \)-th block is \( E \Sigma_k E^T \). Since \( C \) is deterministic, \( \Sigma_{CC} \) 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 \( F_s = 44100Hz \), 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 \( (F_s/N, F_s/4) \) range, with uniformly sampled phases \( \phi \in (-\pi, \pi) \), and computed the variance of estimation error \( \frac{2(\rho)^{-1}}{M} \sum_{i=1}^{M} (f_i - f_i) \). 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 \( 2F_s/N \) 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.](image-url)

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


