Probabilistic Filter and Smoother for Variational Inference of Bayesian Linear Dynamical Systems - Slides

Julian Neri[1] Philippe Depalle[1] Roland Badeau[2]

[1]McGill University, CIRMMT, Montréal, Canada
[2]LTCI, Télécom Paris, Institut Polytechnique de Paris, France

May 6, 2020

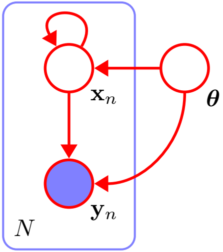
Overview

Introduction - Model

Bayesian linear dynamical systems (BLDS) are widely applicable probabilistic models for sequential data.

- Parameters $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{B}, \mathbf{Q}, \mathbf{C}, \mathbf{D}, \mathbf{R}, \mathbf{m}_0, \mathbf{P}_0\}$ are stochastic (essentially latent variables),

$$\begin{aligned}
\mathbf{y}_n &\sim \mathcal{N}\left(\mathbf{C}\mathbf{x}_n + \mathbf{D}\mathbf{u}_n, \mathbf{R}\right) \\
\mathbf{x}_n &\sim \mathcal{N}\left(\mathbf{A}\mathbf{x}_{n-1} + \mathbf{B}\mathbf{u}_n, \mathbf{Q}\right) \\
\mathbf{x}_1 &\sim \mathcal{N}\left(\mathbf{m}_0, \mathbf{P}_0\right) \\
\boldsymbol{\theta} &\sim p(\boldsymbol{\theta})
\end{aligned}$$



Fully Bayesian treatment discourages over-fitting and enables the automatic learning of the model's structure, i.e. the state space's optimal dimensionality.

Introduction - Variational inference

✗ Exact inference of BLDS is intractable

✓ **Structured mean-field variational inference**

- factor latent variable sequence from parameters

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

Variational inference turns inference into an optimization problem, maximizing the lower bound on model evidence:

$$\mathcal{L}(q) = \ln p(\mathbf{Y}) - \mathsf{KL}(q(\mathbf{X}, \boldsymbol{\theta}) \| p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y}))$$

Log optimal distributions that maximize $\mathcal{L}(q)$:

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

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

Introduction - Variational inference

✗ Exact inference of BLDS is intractable

✓ **Structured mean-field variational inference**

- factor latent variable sequence from parameters

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

Variational inference turns inference into an optimization problem, maximizing the lower bound on model evidence:

$$\mathcal{L}(q) = \ln p(\mathbf{Y}) - \mathsf{KL}(q(\mathbf{X}, \boldsymbol{\theta}) \| p(\mathbf{X}, \boldsymbol{\theta}|\mathbf{Y}))$$

Log optimal distributions that maximize $\mathcal{L}(q)$:

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

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

Introduction - Existing methods

Hardest part is inferring the statistics of the marginal posterior $q^{\star}_{\mathbf{x}}(\mathbf{x}_n)$ and $q^{\star}_{\mathbf{x}}(\mathbf{x}_n, \mathbf{x}_{n+1})$, as needed to update $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.

- Must care for terms in $\ln q^{\star}_{\mathbf{x}}(\mathbf{X})$ that involve expectations of quadratics taken w.r.t. $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$
- For example, $\langle \ln p(\mathbf{x}_n|\mathbf{x}_{n-1}) \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$ includes the term $\langle \mathbf{A}^{\mathsf{T}} \mathbf{Q}^{-1} \mathbf{A} \rangle_{q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$
- Cannot use the usual Kalman filter/smoother algorithms.

Belief propagation was first approach [Beal, 2003], followed by several alternative methods.

✗ Existing methods do not scale linearly with time, are not numerically accurate/stable, or are not equivalent to the original belief propagation algorithm.

Proposed method

Decompose problematic second moment quadratic terms into first moment quadratics plus covariances.

For *generic* matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$, and covariance matrix $\boldsymbol{\Lambda}$, taken w.r.t. $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, the following decomposition holds for arbitrary $p(\boldsymbol{\theta})$[1]:

$$\langle \boldsymbol{\Psi}^{\mathsf{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Omega} \rangle = \langle \boldsymbol{\Psi} \rangle^{\mathsf{T}} \langle \boldsymbol{\Lambda}^{-1} \rangle \langle \boldsymbol{\Omega} \rangle + \boldsymbol{\Sigma}_{\Psi\Lambda\Omega}$$

$$\boldsymbol{\Sigma}_{\Psi\Lambda\Omega} = \sum_i \sum_j \langle \boldsymbol{\Lambda}^{-1} \rangle_{(i,j)} \mathsf{cov}[\boldsymbol{\psi}_{(i)}, \boldsymbol{\omega}_{(j)}].$$

We apply matrix inversion lemmas to derive new forward (filter) and backward (smoother) algorithms that

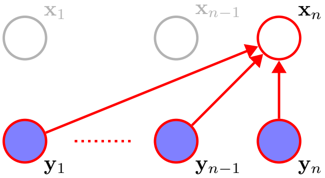
- scale linearly with time
- respect parameter covariances
- have the same desirable forms of the Kalman filter/smoother

[1] See [Neri et al., 2020] and the Appendix for details: http://www.music.mcgill.ca/~julian/vblds.

Proposed method - forward pass

The forward pass calculates the mean $\boldsymbol{\mu}_n$ and covariance $\mathbf{V}_n$ of the marginal probability[2], $\forall n \in [1..N]$:

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n}) = \frac{p(\mathbf{y}_n|\mathbf{x}_n)q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1})}{q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1})}$$



[2] Full derivations of the forward and backward pass are included in the appendix: http://www.music.mcgill.ca/~julian/vblds.

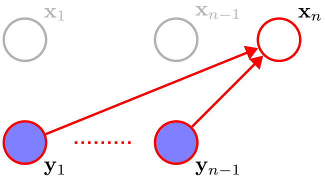
Proposed method - forward pass

The predictive distribution is

$$
\begin{aligned}
q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1}) &= \int p(\mathbf{x}_n|\mathbf{x}_{n-1}) q_{\mathbf{x}}(\mathbf{x}_{n-1}|\mathbf{y}_{1:n-1}) d\mathbf{x}_{n-1} \\
&= \mathcal{N}(\mathbf{x}_n|\mathbf{m}_{n-1}, \mathbf{P}_{n-1})
\end{aligned}
$$

where the mean $\mathbf{m}_{n-1}$ and covariance $\mathbf{P}_{n-1}$ are

$$
\begin{aligned}
\mathbf{G}_{n-1} &= \mathbf{I} - \mathbf{V}_{n-1} \left(\mathbf{I} + \mathbf{\Sigma}_{AQA} \mathbf{V}_{n-1}\right)^{-1} \mathbf{\Sigma}_{AQA} \\
\mathbf{m}_{n-1} &= \langle \mathbf{A} \rangle \mathbf{G}_{n-1} \left(\boldsymbol{\mu}_{n-1} - \mathbf{V}_{n-1} \mathbf{\Sigma}_{AQB} \mathbf{u}_n\right) + \langle \mathbf{B} \rangle \mathbf{u}_n \\
\mathbf{P}_{n-1} &= \langle \mathbf{A} \rangle \mathbf{G}_{n-1} \mathbf{V}_{n-1} \langle \mathbf{A} \rangle^{\mathsf{T}} + \langle \mathbf{Q} \rangle
\end{aligned}
$$



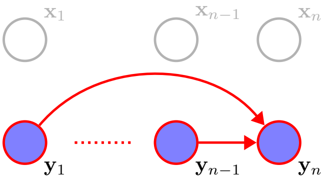
Proposed method - forward pass

The filtered output probability is

$$\begin{aligned} q_{\mathbf{x}}(\mathbf{y}_n|\mathbf{y}_{1:n-1}) &= \int p(\mathbf{y}_n|\mathbf{x}_n) q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n-1}) d\mathbf{x}_n \\ &= \mathcal{N}(\mathbf{y}_n|\widehat{\mathbf{y}}_n, \mathbf{S}_n) \end{aligned}$$

where the mean $\widehat{\mathbf{y}}_n$ and covariance $\mathbf{S}_n$ are

$$\begin{aligned} \mathbf{L}_{n-1} &= \mathbf{I} - \mathbf{P}_{n-1} \left(\mathbf{I} + \mathbf{\Sigma}_{CRC}\mathbf{P}_{n-1}\right)^{-1} \mathbf{\Sigma}_{CRC} \\ \widehat{\mathbf{y}}_n &= \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\left(\mathbf{m}_{n-1} - \mathbf{P}_{n-1}\mathbf{\Sigma}_{CRD}\mathbf{u}_n\right) + \langle\mathbf{D}\rangle\mathbf{u}_n \\ \mathbf{S}_n &= \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}\rangle^{\mathsf{T}} + \langle\mathbf{R}\rangle \end{aligned}$$



Proposed method - forward pass

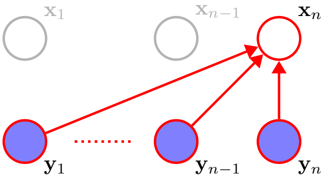
The updated marginal posterior probability is

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

where the state's mean and covariance are

$$\begin{aligned}
\mathbf{K}_n &= \mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}\rangle^{\mathsf{T}}\mathbf{S}_n^{-1} \\
\boldsymbol{\mu}_n &= \mathbf{L}_{n-1}\left(\mathbf{m}_{n-1} - \mathbf{P}_{n-1}\boldsymbol{\Sigma}_{CRD}\mathbf{u}_n\right) + \mathbf{K}_n\left(\mathbf{y}_n - \widehat{\mathbf{y}}_n\right) \\
\mathbf{V}_n &= \left(\mathbf{I} - \mathbf{K}_n\langle\mathbf{C}\rangle\right)\mathbf{L}_{n-1}\mathbf{P}_{n-1}
\end{aligned}$$

$\rightarrow \mathbf{K}_n$ is the (Bayesian) Kalman gain.



Proposed method - forward pass

Complete forward pass equations:

$$
\begin{aligned}
\mathbf{G}_{n-1} &= \mathbf{I} - \mathbf{V}_{n-1}\left(\mathbf{I} + \boldsymbol{\Sigma}_{AQA}\mathbf{V}_{n-1}\right)^{-1}\boldsymbol{\Sigma}_{AQA} \\
\mathbf{m}_{n-1} &= \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\left(\boldsymbol{\mu}_{n-1} - \mathbf{V}_{n-1}\boldsymbol{\Sigma}_{AQB}\mathbf{u}_n\right) + \langle\mathbf{B}\rangle\mathbf{u}_n \\
\mathbf{P}_{n-1} &= \langle\mathbf{A}\rangle\mathbf{G}_{n-1}\mathbf{V}_{n-1}\langle\mathbf{A}\rangle^{\mathsf{T}} + \langle\mathbf{Q}\rangle
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{L}_{n-1} &= \mathbf{I} - \mathbf{P}_{n-1}\left(\mathbf{I} + \boldsymbol{\Sigma}_{CRC}\mathbf{P}_{n-1}\right)^{-1}\boldsymbol{\Sigma}_{CRC} \\
\widehat{\mathbf{y}}_n &= \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\left(\mathbf{m}_{n-1} - \mathbf{P}_{n-1}\boldsymbol{\Sigma}_{CRD}\mathbf{u}_n\right) + \langle\mathbf{D}\rangle\mathbf{u}_n \\
\mathbf{S}_n &= \langle\mathbf{C}\rangle\mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}\rangle^{\mathsf{T}} + \langle\mathbf{R}\rangle
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{K}_n &= \mathbf{L}_{n-1}\mathbf{P}_{n-1}\langle\mathbf{C}\rangle^{\mathsf{T}}\mathbf{S}_n^{-1} \\
\boldsymbol{\mu}_n &= \mathbf{L}_{n-1}\left(\mathbf{m}_{n-1} - \mathbf{P}_{n-1}\boldsymbol{\Sigma}_{CRD}\mathbf{u}_n\right) + \mathbf{K}_n\left(\mathbf{y}_n - \widehat{\mathbf{y}}_n\right) \\
\mathbf{V}_n &= \left(\mathbf{I} - \mathbf{K}_n\langle\mathbf{C}\rangle\right)\mathbf{L}_{n-1}\mathbf{P}_{n-1}
\end{aligned}
$$

When all parameter covariances are zero ($\boldsymbol{\Sigma}_{AQA} = \mathbf{0}$, etc.) the forward pass reduces exactly to the Kalman filter!

Proposed method - forward pass

Complete forward pass equations:

$$\begin{aligned}
\mathbf{G}_{n-1} &= \mathbf{I} \\
\mathbf{m}_{n-1} &= \langle \mathbf{A} \rangle \boldsymbol{\mu}_{n-1} + \langle \mathbf{B} \rangle \mathbf{u}_n \\
\mathbf{P}_{n-1} &= \langle \mathbf{A} \rangle \mathbf{V}_{n-1} \langle \mathbf{A} \rangle^{\mathsf{T}} + \langle \mathbf{Q} \rangle
\end{aligned}$$

$$\begin{aligned}
\mathbf{L}_{n-1} &= \mathbf{I} \\
\widehat{\mathbf{y}}_n &= \langle \mathbf{C} \rangle \mathbf{m}_{n-1} + \langle \mathbf{D} \rangle \mathbf{u}_n \\
\mathbf{S}_n &= \langle \mathbf{C} \rangle \mathbf{P}_{n-1} \langle \mathbf{C} \rangle^{\mathsf{T}} + \langle \mathbf{R} \rangle
\end{aligned}$$

$$\begin{aligned}
\mathbf{K}_n &= \mathbf{P}_{n-1} \langle \mathbf{C} \rangle^{\mathsf{T}} \mathbf{S}_n^{-1} \\
\boldsymbol{\mu}_n &= \mathbf{m}_{n-1} + \mathbf{K}_n \left(\mathbf{y}_n - \widehat{\mathbf{y}}_n \right) \\
\mathbf{V}_n &= \left(\mathbf{I} - \mathbf{K}_n \langle \mathbf{C} \rangle \right) \mathbf{P}_{n-1}
\end{aligned}$$

When all parameter covariances are zero ($\boldsymbol{\Sigma}_{AQA} = \mathbf{0}$, etc.) the forward pass reduces exactly to the Kalman filter!

Proposed method - backward pass

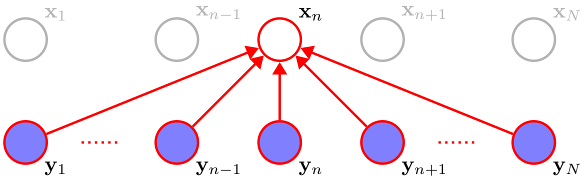
Backward pass computes the marginal posterior given all the data:

$$q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{Y}) = q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n}) \int \frac{p(\mathbf{x}_{n+1}|\mathbf{x}_n) q_{\mathbf{x}}(\mathbf{x}_{n+1}|\mathbf{Y})}{q_{\mathbf{x}}(\mathbf{x}_{n+1}|\mathbf{y}_{1:n})} d\mathbf{x}_{n+1}$$
$$= \mathcal{N}(\mathbf{x}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n).$$

Propagates backwards from $n = N-1$ to $n = 1$:

$$\mathbf{J}_n = \mathbf{G}_n \mathbf{V}_n \langle \mathbf{A} \rangle^{\mathsf{T}} \mathbf{P}_n^{-1}$$
$$\widehat{\boldsymbol{\mu}}_n = \mathbf{G}_n \left(\boldsymbol{\mu}_n - \mathbf{V}_n \boldsymbol{\Sigma}_{AQB} \mathbf{u}_{n+1} \right) + \mathbf{J}_n \left(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n \right)$$
$$\widehat{\mathbf{V}}_n = \mathbf{G}_n \mathbf{V}_n + \mathbf{J}_n \left(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n \right) \mathbf{J}_n^{\mathsf{T}}.$$

Equivalent to RTS smoother when parameter covariances are zero!



Proposed method - backward pass

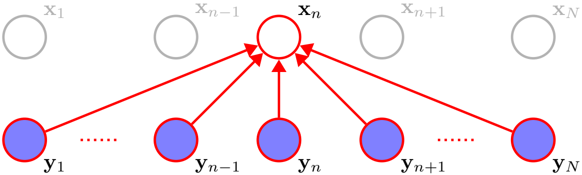
Backward pass computes the marginal posterior given all the data:

$$
\begin{aligned}
q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{Y}) &= q_{\mathbf{x}}(\mathbf{x}_n|\mathbf{y}_{1:n}) \int \frac{p(\mathbf{x}_{n+1}|\mathbf{x}_n) q_{\mathbf{x}}(\mathbf{x}_{n+1}|\mathbf{Y})}{q_{\mathbf{x}}(\mathbf{x}_{n+1}|\mathbf{y}_{1:n})} d\mathbf{x}_{n+1} \\
&= \mathcal{N}(\mathbf{x}_n|\widehat{\boldsymbol{\mu}}_n, \widehat{\mathbf{V}}_n).
\end{aligned}
$$

Propagates backwards from $n = N - 1$ to $n = 1$:

$$
\begin{aligned}
\mathbf{J}_n &= \mathbf{V}_n \langle \mathbf{A} \rangle^{\mathsf{T}} \mathbf{P}_n^{-1} \\
\widehat{\boldsymbol{\mu}}_n &= \boldsymbol{\mu}_n + \mathbf{J}_n \left(\widehat{\boldsymbol{\mu}}_{n+1} - \mathbf{m}_n \right) \\
\widehat{\mathbf{V}}_n &= \mathbf{V}_n + \mathbf{J}_n \left(\widehat{\mathbf{V}}_{n+1} - \mathbf{P}_n \right) \mathbf{J}_n^{\mathsf{T}}.
\end{aligned}
$$

Equivalent to RTS smoother when parameter covariances are zero!

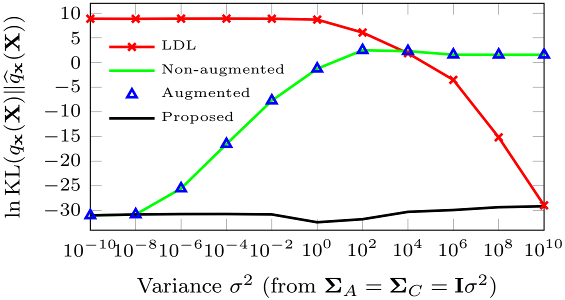


Proposed method - validation

Compared the numerical accuracy across the existing methods:

- belief propagation (ground truth) [Beal, 2003]
- Cholesky factor approach [Barber and Chiappa, 2006]
- Cholesky factor approach with augmented state [Barber, 2006]
- LDL decomposition [Luttinen, 2013]

Proposed filter/smoother's $\widehat{q}_{\mathbf{x}}(\mathbf{X})$ is mathematically equivalent and more numerically stable than the ground truth $q_{\mathbf{x}}(\mathbf{X})$.



Application - Bayesian frequency estimator (BFE)

Structure the dynamics matrix so elements of 2×1 sub-vector $\mathbf{x}_{nk}$ oscillate at $f_k = \arccos(\nu_k + 1)/(2\pi T)$ Hz:

$$\mathbf{A}_k = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 & \frac{1}{2} \\ 2 & 1 \end{pmatrix} \nu_k, \qquad \mathbf{Q}_k = \tau_k^{-1} \mathbf{I}$$

Set the amplitude g_k and initial phase ϕ_k through the initial mean

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

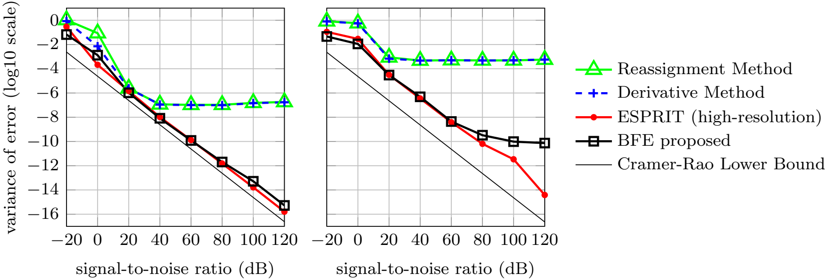
Conjugate Normal-Gamma priors over the model parameters

$$\begin{aligned} p(\boldsymbol{\nu}|\boldsymbol{\tau})p(\boldsymbol{\tau}) &= \textstyle\prod_{k=1}^{K} \mathcal{N}(\nu_k|0, \alpha_k^{-1}\tau_k^{-1})\mathsf{Gam}(\tau_k|e_0, i_0) \\ p(\rho) &= \mathsf{Gam}(\rho|r_0, s_0) \end{aligned}$$

Automatic relevance determination (ARD) hyperparameter α_k can promote sparse solutions when optimized [Beal, 2003].

Application - Results

Compared the proposed BFE with state-of-the-art methods.



(a) One sinusoid. (b) Two sinusoids.

While BFE is more complex than existing methods, it can

- infer the number of relevant oscillations in a signal,
- infer a signal's noise variance,
- measure frequency, amplitude and phase estimate uncertainty,
- separate a noisy signal into a set of filtered oscillations.

Conclusion

We presented a new algorithm for inferring the latent state sequence of a BLDS $\rightarrow$ Bayesian generalization to the Kalman filter and smoother.

- more numerically stable than existing routines and respects the statistical moments of the parameters
- cost that scales linearly with the data sequence length
- applicable to BLDS and its extensions, like recurrent switching linear dynamical systems

References

Barber, D. (2006).
Expectation correction for smoothed inference in switching linear dynamical systems.
Journal of Machine Learning Research, 7:2515–2540.

Barber, D. and Chiappa, S. (2006).
Unified inference for variational Bayesian linear Gaussian state-space models.
In *Advances in Neural Information Processing Systems 19*, pages 81–88. MIT Press.

Beal, M. (2003).
Variational Algorithms for Approximate Bayesian Inference.
PhD thesis, University College London.

Luttinen, J. (2013).
Fast variational Bayesian linear state-space model.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 305–320, Springer, Berlin, Heidelberg.

Neri, J., Badeau, R., and Depalle, P. (2020).
Probabilistic filter and smoother for variational inference of Bayesian linear dynamical systems.