Bayesian Inference and Learning in Gaussian Process State-Space Models with Particle MCMC – Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

1 Implementation Details – Cholesky updates of the Matrix $\widetilde{K}$

The Cholesky decomposition of a positive-definite matrix K is a decomposition into a lower triangular matrix L, according to $K = LL^{\mathsf{T}}$, where we refere to L as the Cholesky factor of K. Consider a symmetric positive-definite matrix K with Cholesky factor L, both with block entries according to

$$K = \begin{pmatrix} K_{11} & K_{12} & K_{13} \\ \times & K_{22} & K_{23} \\ \times & \times & K_{33} \end{pmatrix} \qquad L = \begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{pmatrix}, \tag{1}$$

where K_{12} is a column vector, $K_2 2$ is a scalar and K_{23} is a row vector. We construct the matrix $\widetilde{K}$ by replacing one row and one column in K. The task is now to make use of this decomposition in order to find the Cholesky factor M of the matrix $\widetilde{K}$,

$$\widetilde{K} = \begin{pmatrix} K_{11} & \widetilde{K}_{12} & K_{13} \\ \times & \widetilde{K}_{22} & \widetilde{K}_{23} \\ \times & \times & K_{33} \end{pmatrix} \qquad \widetilde{K} = MM^{\mathsf{T}}, \qquad M = \begin{pmatrix} M_{11} & 0 & 0 \\ M_{21} & M_{22} & 0 \\ M_{31} & M_{32} & M_{33} \end{pmatrix}. \tag{2}$$

without having to explicitly compute yet another Cholesky decomposition. Put in slightly different words we want to express the block entries of the Cholesky factor M in terms of the block entries in L that we already have available to us in (1).

From (1) we have

$$K = LL^{\mathsf{T}} = \begin{pmatrix} L_{11}L_{11}^{\mathsf{T}} & L_{11}L_{21}^{\mathsf{T}} & L_{11}L_{31}^{\mathsf{T}} \\ \times & L_{21}L_{21}^{\mathsf{T}} + L_{22}L_{22}^{\mathsf{T}} & L_{21}L_{31}^{\mathsf{T}} + L_{22}L_{32}^{\mathsf{T}} \\ \times & \times & L_{31}L_{31}^{\mathsf{T}} + L_{32}L_{32}^{\mathsf{T}} + L_{33}L_{33}^{\mathsf{T}} \end{pmatrix} \tag{3}$$

allowing us to identify the following relationships

$$K_{11} = L_{11}L_{11}^{\mathsf{T}}, \tag{4a}$$
$$K_{21} = L_{21}L_{11}^{\mathsf{T}}, \tag{4b}$$
$$K_{31} = L_{31}L_{11}^{\mathsf{T}}, \tag{4c}$$
$$K_{22} = L_{21}L_{21}^{\mathsf{T}} + L_{22}L_{22}^{\mathsf{T}}, \tag{4d}$$
$$K_{32} = L_{31}L_{21}^{\mathsf{T}} + L_{32}L_{22}^{\mathsf{T}}, \tag{4e}$$
$$K_{33} = L_{31}L_{31}^{\mathsf{T}} + L_{32}L_{32}^{\mathsf{T}} + L_{33}L_{33}^{\mathsf{T}}. \tag{4f}$$

From (4a) and (2) we have

$$M_{11} = L_{11}. \tag{5a}$$

By similar identifications of terms we have

$$M_{21} = \widetilde{K}_{21} L_{11}^{-\mathsf{T}}, \tag{5b}$$
$$M_{31} = L_{31}, \tag{5c}$$
$$M_{22}^2 = \widetilde{K}_{22} - M_{21} M_{21}^{\mathsf{T}}, \tag{5d}$$
$$M_{32} = \frac{1}{M_{22}} \left(\widetilde{K}_{32} - M_{31} M_{21}^{\mathsf{T}} \right), \tag{5e}$$
$$M_{33} M_{33}^{\mathsf{T}} = K_{33} - M_{31} M_{31}^{\mathsf{T}} - M_{32} M_{32}^{\mathsf{T}}, \tag{5f}$$

Using back-substitution it is straightforward to find the factors M_{21} and M_{31} by exploiting the fact that the L_{11} matrix is lower triangular. The term M_{22} is scalar and found simply via a square root operation and the term M_{32} is given by (5e). However, computing M_{33} by a direct Cholesky decomposition is too expensive and again it comes down to exploiting the structure inherent in the problem. Let us start by inserting (4f) into (5f), resulting in

$$M_{33} M_{33}^{\mathsf{T}} = L_{33} L_{33}^{\mathsf{T}} + L_{31} L_{31}^{\mathsf{T}} + L_{32} L_{32}^{\mathsf{T}} - M_{31} M_{31}^{\mathsf{T}} - M_{32} M_{32}^{\mathsf{T}} \tag{6a}$$
$$= L_{33} L_{33}^{\mathsf{T}} + L_{32} L_{32}^{\mathsf{T}} - M_{32} M_{32}^{\mathsf{T}}, \tag{6b}$$

where the last equality follows (5c). Since $L_{32} L_{32}^{\mathsf{T}}$ and $M_{32} M_{32}^{\mathsf{T}}$ are both rank one, we can now compute M_{33} by one rank-one update and one rank-one downdate of L_{33}. See [1] for an overview of rank-one update/downdate methods. This concludes our work in finding the decomposition $\widetilde{K} = MM^{\mathsf{T}}$.

2 The CPF-AS Algorithm

The basic idea underlying PMCMC is to use SMC to construct a Markov kernel leaving the exact joint smoothing distribution invariant. Hence, we seek a family of ergodic Markov kernels on X^{T+1},

$$\{M_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}, \tag{7}$$

such that, for each $\boldsymbol{\theta}$, $M_{\boldsymbol{\theta}}(\mathbf{x}_{0:T} \mid \widetilde{\mathbf{x}}_{0:T})$ leaves $p(\mathbf{x}_{0:T} \mid \boldsymbol{\theta}, \mathbf{y}_{0:T})$ invariant. In PGAS, these kernels are constructed using a procedure referred to as a conditional particle filter with ancestor sampling (CPF-AS). This procedure is particularly suitable for non-Markovian latent variable models [2], as it relies only on a forward recursion.

CPF-AS is similar to a standard SMC sampler, but with the important difference that one particle at each time step is specified *a priori*. Let these particles be denoted $\widetilde{\mathbf{x}}_{0:T} = \{\widetilde{\mathbf{x}}_0, \ldots, \widetilde{\mathbf{x}}_T\}$. More precisely, we condition on the event that $\widetilde{\mathbf{x}}_t$ is contained in the collection of particles $\{\mathbf{x}_t^i\}_{i=1}^N$, generated at time t. To accomplish this, we sample according to $\mathbf{x}_t^i \sim p(\mathbf{x}_t \mid \boldsymbol{\theta}, \mathbf{x}_{0:t-1}^{\mathbf{a}_t^i})$ only for $i = 1, \ldots, N-1$. The Nth particle is then set deterministically: $\mathbf{x}_t^N = \widetilde{\mathbf{x}}_t$. The CPF-AS is given in Algorithm 1.

The conditioning on a pre-specified collection of particles implies an invariance property of the CPF-AS, which is key to its applicability in an MCMC sampler.

Proposition 1. *Let the support of the target density be a subset of the support of the proposal density. Then, for any $\boldsymbol{\theta}$ and any $N \geq 2$, the procedure*

(i) Run Algorithm 1 conditionally on $\widetilde{\mathbf{x}}_{0:T}$;

(ii) Sample $\widetilde{\mathbf{x}}'_{0:T}$ with $\mathbb{P}(\widetilde{\mathbf{x}}'_{0:T} = \mathbf{x}_{0:T}^i) = \mathbf{w}_T^i$;

defines an irreducible and aperiodic Markov kernel $M_{\boldsymbol{\theta}}^N$ on X^T, with invariant distribution $p(\mathbf{x}_{0:T} \mid \boldsymbol{\theta}, \mathbf{y}_{0:T})$.

Proof. The invariance property follows by the construction of the CPF-AS in [2], and the fact that the law of $\widetilde{\mathbf{x}}'_{0:T}$ is independent of permutations of the particle indices. This allows us to always place the conditioned particles at the Nth position. Irreducibility and aperiodicity follows from [3, Theorem 5]. □

Algorithm 1 CPF-AS, conditioned on $\widetilde{\mathbf{x}}_{0:T}$

1. **Initialize:**
   (a) Draw $\mathbf{x}_0^i \sim p(\mathbf{x}_0 \mid \theta, \mathbf{y}_0)$ for $i = 1, \ldots, N-1$.
   (b) Set $\mathbf{x}_0^N = \widetilde{\mathbf{x}}_0$.
   (c) For $i = 1, \ldots, N$, set $\mathbf{w}_0^i \propto p(\mathbf{y}_0 \mid \theta, \mathbf{x}_0^i)$, where the weights are normalized to sum to 1.
2. **For** $t = 1, \ldots, T$ **do:**
   (a) Draw $\mathbf{a}_t^i$ with $P(\mathbf{a}_t^i = j) = \mathbf{w}_{t-1}^j$ for $i = 1, \ldots, N-1$.
   (b) Draw $\mathbf{x}_t^i \sim p(\mathbf{x}_t \mid \boldsymbol{\theta}, \mathbf{x}_{0:t-1}^{\mathbf{a}_t^i})$ for $i = 1, \ldots, N-1$.
   (c) Draw $\mathbf{a}_t^N$ with $\mathbb{P}(\mathbf{a}_t^N = j) \propto \mathbf{w}_{t-1}^j p(\widetilde{\mathbf{x}}_{t:T} \mid \boldsymbol{\theta}, \mathbf{x}_{1:t-1}^j)$.
   (d) Set $\mathbf{x}_t^N = \widetilde{\mathbf{x}}_t$.
   (e) For $i = 1, \ldots, N$, set $\mathbf{w}_t^i \propto p(\mathbf{y}_t \mid \theta, \mathbf{x}_t^i)$, where the weights are normalized to sum to 1.

Consequently, if $\widetilde{\mathbf{x}}_{0:T} \sim p(\mathbf{x}_{0:T} \mid \boldsymbol{\theta}, \mathbf{y}_{0:T})$ and we sample $\widetilde{\mathbf{x}}'_{0:T}$ according to the procedure given in Proposition 1, then, for any number of particles N, it holds that $\widetilde{\mathbf{x}}'_{0:T} \sim p(\mathbf{x}_{0:T} \mid \boldsymbol{\theta}, \mathbf{y}_{0:T})$. For $N = 1$ we get, by construction, $\widetilde{\mathbf{x}}'_{0:T} = \widetilde{\mathbf{x}}_{0:T}$, i.e. the trajectories are perfectly correlated (this is why we need $N \geq 2$ to get an irreducible kernel). On the other hand, as $N \to \infty$, the conditioning will have a negligible effect on the CPF-AS. Hence, $\widetilde{\mathbf{x}}'_{0:T}$ will be effectively independent of $\widetilde{\mathbf{x}}_{0:T}$ and (with an infinite number of particles) distributed according to the exact smoothing distribution. The number of particles N will thus affect the mixing of the Markov kernel $M_{\boldsymbol{\theta}}^N$. The invariance property of the kernel holds for any N, but the larger we take N, the smaller the correlation will be between $\widetilde{\mathbf{x}}'_{0:T}$ and $\widetilde{\mathbf{x}}_{0:T}$. However, it has been experienced in practice that the correlation drops off very quickly as N increases [2, 4], and for many models a moderate N is enough to obtain a rapidly mixing kernel.

3 An Additional plot for the Nonlinear System Benchmark

In Figure 1 we show the values of the hyper-parameters that are learnt during the experiment.

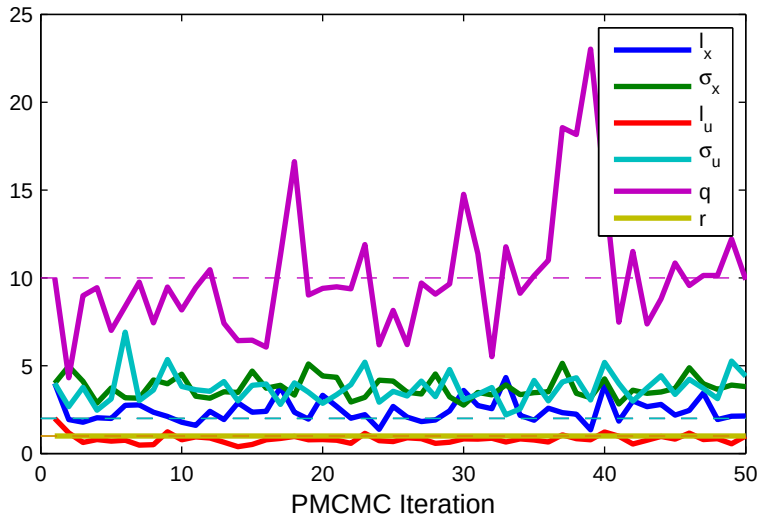

Figure 1: Hyper-parameter samples.

References

[1] P. Gill, G. Golub, W. Murray, and M. Saunders, "Methods for modifying matrix factorizations," *Mathematics of Computation*, vol. 126, no. 28, pp. 505–535, 1974.

[2] F. Lindsten, M. Jordan, and T. B. Schön, "Ancestor sampling for particle Gibbs," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 2600–2608.

[3] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society: Series B*, vol. 72, no. 3, pp. 269–342, 2010.

[4] F. Lindsten and T. B. Schön, "On the use of backward simulation in the particle Gibbs sampler," in *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012.