Supplementary material

This is the supplementary material for 'Computationally Efficient Bayesian Learning of Gaussian Process State Space Models' by Svensson, Solin, Särkkä and Schön, presented at AISTATS 2016. The references in this document point to the bibliography in the article.

1 Proofs

Proof of Theorem 4.1. Let us start by considering the GP approximation to $\mathbf{f}(\mathbf{x})$, $\mathbf{x} \in [-L_1, L_1] \times \cdots \times [-L_d, L_d]$. By Theorem 4.4 of Solin and Särkkä (2014), when domain size $\inf_i L_i \to \infty$ and the number of basis functions $m \to \infty$, the approximate covariance function $\boldsymbol{\kappa}_m(\mathbf{x}, \mathbf{x}')$ converges point-wise to $\boldsymbol{\kappa}(\mathbf{x}, \mathbf{x}')$. As the prior means of the exact and approximate GPs are both zero, the means thus converge as well. By similar argument as is used in the proof of Theorem 2.2 in Särkkä and Piché (2014) it follows that the posterior mean and covariance functions will converge point-wise as well.

Now, consider the random variables defined by

$$\mathbf{x}_{t+1} = \mathbf{f}(\mathbf{x}_t) + \mathbf{w}_t, \tag{17}$$

$$\hat{\mathbf{x}}_{t+1} = \mathbf{f}_m(\mathbf{x}_t) + \mathbf{w}_t, \tag{18}$$

where $\mathbf{f}_m$ is an m-term series expansion approximation to the GP. It now follows that for any fixed $\mathbf{x}_t$ the mean and covariance of $\mathbf{x}_{t+1}$ and $\hat{\mathbf{x}}_{t+1}$ coincide when $L_i, m \to \infty$. However, because these random variables are Gaussian, the first two moments determine the whole distribution and hence we can conclude that $\hat{\mathbf{x}}_{t+1} \to \mathbf{x}_{t+1}$ in distribution.

For the measurement model we can similarly consider the random variables

$$\mathbf{y}_t = \mathbf{g}(\mathbf{x}_t) + \mathbf{e}_t, \tag{19}$$

$$\hat{\mathbf{y}}_t = \mathbf{g}_m(\mathbf{x}_t) + \mathbf{e}_t, \tag{20}$$

With similar argument as above, we can conclude that the approximation converges in distribution. □

Proof of Theorem 4.2. Provided the reduced-rank approximation of the Gram matrix, the reduction in the computational load directly follows from application of the matrix inversion lemma. □

Proof of Theorem 4.3. Using fundamental properties of the Gibbs sampler (see, *e.g.*, Tierney (1994)), the claim holds if all steps of Algorithm 1 are leaving the right conditional probability density invariant. Step 3 is justified by Lindsten et al. (2014) (even for a finite N), and step 4–5 by Wills et al. (2012). Further, step 6 can be seen as a Metropolis-within-Gibbs procedure (Tierney, 1994). □

2 Details on Matrix Normal and Inverse Wishart distributions

As presented in the article, the matrix normal inverse Wishart (MNIW) distribution is the conjugate prior for state space models linear in its parameters $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{Q} \in \mathbb{R}^{n \times n_x}$ Wills et al. (2012). The MNIW distribution can be written as $\mathcal{MN}\mathcal{IW}(\mathbf{A}, \mathbf{Q} \mid M, \mathbf{V}, \ell, \mathbf{\Lambda}) = \mathcal{MN}(\mathbf{A} \mid M, \mathbf{Q}, \mathbf{V}) \times \mathcal{IW}(\mathbf{Q} \mid \ell, \mathbf{\Lambda})$, where each part is defined as follows:

- The pdf for the Inverse Wishart distribution with ℓ degrees of freedom and positive definite scale matrix $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$:

$$\mathcal{IW}(\mathbf{Q} \mid \ell, \mathbf{\Lambda}) = \frac{|\mathbf{\Lambda}|^{\ell/2} |\mathbf{Q}|^{-(n+\ell+1)/2}}{2^{\ell n/2} \Gamma_n(\ell/2)} \exp\left(-\frac{1}{2} \mathrm{tr}(\mathbf{Q}^{-1}\mathbf{\Lambda})\right) \tag{21}$$

  with $\Gamma_n(\cdot)$ being the multivariate gamma function.

- The pdf for the Matrix Normal distribution with mean $\mathbf{M} \in \mathbb{R}^{n \times m}$, right covariance $\mathbf{Q} \in \mathbb{R}^{n \times n}$ and left precision $\mathbf{V} \in \mathbb{R}^{m \times m}$:

$$\mathcal{MN}(\mathbf{A} \mid \mathbf{M}, \mathbf{Q}, \mathbf{V}) = \frac{|\mathbf{V}|^{n/2}}{(2\pi)^{nm} |\mathbf{Q}|^{m/2}} \exp\left(-\frac{1}{2} \mathrm{tr}\left((\mathbf{A} - \mathbf{M})^{\mathsf{T}} \mathbf{Q}^{-1} (\mathbf{A} - \mathbf{M}) \mathbf{V}\right)\right) \tag{22}$$

To sample from the MN distribution, one may sample a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ of i.i.d. $\mathcal{N}(0, 1)$ random variables, and obtain $\mathbf{A}$ as $\mathbf{A} = \mathbf{M} + \mathrm{chol}(\mathbf{Q})\, \mathbf{X}\, \mathrm{chol}(\mathbf{V})$, where chol denotes the Cholesky factor ($\mathbf{V} = \mathrm{chol}(\mathbf{V})\, \mathrm{chol}(\mathbf{V})^{\mathsf{T}}$).

3 Eigenfunctions for Multi-Dimensional Spaces

The eigenfunctions for a d-dimensional space with a rectangular domain $[-L_1, L_1] \times \cdots \times [-L_d, L_d]$, used in Example 5.2 and Example 5.3, are on the form

$$\phi^{(j_1,\ldots,j_d)}(x) = \prod_{k=1}^{d} \frac{1}{\sqrt{L_k}} \sin\left(\frac{\pi j_k (x_k + L_k)}{2L_k}\right) \qquad \text{with} \quad \lambda_{j_1,\ldots,j_d} = \sum_{k=1}^{d} \left(\frac{\pi j_k}{2L_k}\right)^2. \tag{23}$$

Note how this for $d = 1$ reduces to the univariate case presented in Section 5.1. For further details we refer to Section 4.2 in Solin and Särkkä (2014).

4 Provided Matlab Software

The following Matlab files are available via the first authors homepage:

File	Use	Comments
`synthetic_example_1.m`	First synthetic example (including Figure 1)	
`synthetic_example_2.m`	Second synthetic example	
`damper.m`	MR damper example	For other results, see The MathWorks, Inc. (2015)
`energy_forecast.m`	Energy consumption forecasting example	
`iwishpdf.m`	Implements (21)	
`mvnpdf_log.m`	Logarithm of normal distribution pdf	
`systematic_resampling.m`	Systematic resampling (Step 5, Algorithm 2)	

All files are published under the GPL license.