1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–6

Non-Factorised Variational Inference in Dynamical Systems

Alessandro D. Ialongo ADI24@CAM.AC.UK
University of Cambridge and *Max Planck Institute for Intelligent Systems*

Mark van der Wilk MARK@PROWLER.IO
PROWLER.io

James Hensman JAMES@PROWLER.IO
PROWLER.io

Carl E. Rasmussen CER54@CAM.AC.UK
University of Cambridge

Keywords: Gaussian processes, variational inference, state-space models, time-series, GPSSM.

1. Introduction

Many real world systems exhibit complex non-linear dynamics, and a wide variety of approaches have been proposed to model them. Characterising these dynamics is essential to the analysis of a system's behaviour, as well as to prediction. Autoregressive models map past observations to future ones and are very common in diverse fields such as economics and model predictive control. State-space models (SSMs), lift the dynamics from the observations to a set of auxiliary, unobserved variables (the latent states $\{\mathbf{x}_t\}_{t=1}^T$) which fully describe the state of the system at each time-step, and thus take the system to evolve as a Markov chain. The "transition function" maps a latent state to the next. We place a Gaussian process (GP) prior on the transition function, obtaining the Gaussian process state-space model (GPSSM). This Bayesian non-parametric approach allows us to: 1) obtain uncertainty estimates and predictions from the posterior over the transition function, 2) handle increasing amounts of data without the model saturating, and 3) maintain high uncertainty estimates in regions with little data.

While many approximate inference schemes have been proposed (Frigola et al., 2013), we focus on variational "inducing point" approximations (Titsias, 2009), as they offer a particularly elegant framework for approximating GP models without losing key properties of the non-parametric model. Since a non-parametric Gaussian process is used as the approximate posterior (Matthews et al., 2016), the properties of the original model are maintained. Increasing the number of inducing points, we add capacity to the approximation, and the quality of the approximation is measured by the marginal likelihood lower bound (or evidence lower bound – ELBO).

The accuracy of variational methods is fundamentally limited by the class of approximate posteriors, with independence assumptions being particularly harmful for time-series models (Turner et al., 2010). Several variational inference schemes that factorise the states and transition function in the approximate posterior have been proposed (Frigola et al., 2014; McHutchon et al., 2014; Ialongo et al., 2017; Eleftheriadis et al., 2017). Here, we investigate the design choices which are available in specifying a non-factorised posterior.

2. Gaussian Process State Space Models

Conceptually, a GPSSM is identical to other SSMs. We model discrete-time sequences of observations $\mathbf{Y} = \{\mathbf{y}_t\}_{t=1}^T$, where $\mathbf{y}_t \in \mathbb{R}^E$, by a corresponding latent Markov chain of states $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^T$ where $\mathbf{x}_t \in \mathbb{R}^D$. All state-to-state transitions are governed by the same transition function f. For simplicity, we take the transition $p(\mathbf{x}_{t+1} \,|\, f, \mathbf{x}_t)$ and observation $p(\mathbf{y}_t \,|\, \mathbf{x}_t)$ densities to be Gaussians, although any closed form density function could be chosen. Without loss of generality, we also assume a linear mapping between $\mathbf{x}_t$ and the mean of $\mathbf{y}_t \,|\, \mathbf{x}_t$ to eliminate any non-identifiabilities (Frigola et al., 2014; Frigola, 2015). The generative model is specified by the following equations:

$$\begin{aligned} f &\sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot)) & \mathbf{x}_1 &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ \mathbf{x}_{t+1} \,|\, f, \mathbf{x}_t &\sim \mathcal{N}(f(\mathbf{x}_t), \mathbf{Q}) & \mathbf{y}_t \,|\, \mathbf{x}_t &\sim \mathcal{N}(\mathbf{C}\mathbf{x}_t + \mathbf{d}, \mathbf{R}) \end{aligned} \tag{1}$$

the function values $f(\mathbf{x}_t)$ are given by the GP as:

$$\begin{aligned} f(\mathbf{x}_t) \,|\, \mathbf{x}_{1:t}, f(\mathbf{x}_{1:t-1}) &\sim \mathcal{N}(\boldsymbol{\mu}_{f_t}, \mathbf{C}_{f_t}) & f(\mathbf{x}_{1:t-1}) &\triangleq \left[f(\mathbf{x}_1)^\top, \ldots, f(\mathbf{x}_{t-1})^\top\right]^\top \\ \boldsymbol{\mu}_{f_t} &\triangleq K_{\mathbf{x}_t,\mathbf{x}_{1:t-1}} K^{-1}_{\mathbf{x}_{1:t-1},\mathbf{x}_{1:t-1}} f(\mathbf{x}_{1:t-1}) & \mathbf{C}_{f_t} &\triangleq K_{\mathbf{x}_t,\mathbf{x}_t} - K_{\mathbf{x}_t,\mathbf{x}_{1:t-1}} K^{-1}_{\mathbf{x}_{1:t-1},\mathbf{x}_{1:t-1}} K_{\mathbf{x}_{1:t-1},\mathbf{x}_t} \end{aligned} \tag{2}$$

3. Design choices in variational inference

We want a variational approximation to $p(\mathbf{X}, f \,|\, \mathbf{Y})$. We begin by writing our approximate posterior as $q(\mathbf{X}, f) = q(\mathbf{X} \,|\, f) q(f)$ For $q(f)$ we choose an inducing point posterior according to Titsias (2009) and Hensman et al. (2013) and write $q(f) = p(f^{\setminus \mathbf{u}} \,|\, \mathbf{u}) q(\mathbf{u})$ by splitting up the function into the inducing outputs $\mathbf{u}$ and all other points $f^{\setminus \mathbf{u}}$. For our $q(\mathbf{X} \,|\, f)$, we will consider Markovian distributions, following the structure of the true posterior[1]:

$$q(\mathbf{X} \,|\, f) = q(\mathbf{x}_1) \prod_{t=1}^{T-1} q(\mathbf{x}_{t+1} \,|\, f, \mathbf{x}_t) \tag{3}$$

This allows us to write down the general form of the variational lower bound:

$$\log p(\mathbf{Y}) \geq \int q(f, \mathbf{X}) \log \left[\frac{p(f^{\setminus \mathbf{u}} \,|\, \mathbf{u}) p(\mathbf{u}) p(\mathbf{x}_1) \prod_{t=1}^{T-1} p(\mathbf{x}_{t+1} \,|\, f, \mathbf{x}_t)}{p(f^{\setminus \mathbf{u}} \,|\, \mathbf{u}) q(\mathbf{u}) q(\mathbf{x}_1) \prod_{t=1}^{T-1} q(\mathbf{x}_{t+1} \,|\, f, \mathbf{x}_t)} \prod_{t=1}^{T} p(\mathbf{y}_t \,|\, \mathbf{x}_t) \right] \mathrm{d}f \, \mathrm{d}\mathbf{X} \tag{4}$$

$$= \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{x}_t)}[\log p(\mathbf{y}_t \,|\, \mathbf{x}_t)] - \mathrm{KL}[\, q(\mathbf{u}) \,\|\, p(\mathbf{u}) \,] - \mathrm{KL}[\, q(\mathbf{x}_1) \,\|\, p(\mathbf{x}_1) \,]$$

$$- \sum_{t=1}^{T-1} \mathbb{E}_{q(f,\mathbf{x}_t)}[\mathrm{KL}[\, q(\mathbf{x}_{t+1} \,|\, f, \mathbf{x}_t) \,\|\, p(\mathbf{x}_{t+1} \,|\, f, \mathbf{x}_t)) \,]] \tag{5}$$

We are now left to specify the form of $q(\mathbf{x}_1)$, $q(\mathbf{u})$, and $q(\mathbf{x}_{t+1} \,|\, f, \mathbf{x}_t)$. For the first two, we choose Gaussian distributions and optimise variationally their means and covariances, while, for the last, several choices are available. We follow the form of the true filtering distribution (for Gaussian emissions) but treat $\mathbf{A}_t, \mathbf{b}_t, \mathbf{S}_t$ as free variational parameters to be optimised (thus approximating the smoothing distribution):

$$q(\mathbf{x}_{t+1} \,|\, f, \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t+1} \,|\, \mathbf{A}_t \mathbf{f}_t + \mathbf{b}_t, \mathbf{S}_t^*) \tag{6}$$

1. More precisely, the true conditional posterior is: $p(\mathbf{X} \,|\, f, \mathbf{Y}) = p(\mathbf{x}_1 \,|\, f, \mathbf{Y}) \prod_{t=1}^{T-1} p(\mathbf{x}_{t+1} \,|\, f(\mathbf{x}_t), \mathbf{Y}_{t+1:T})$.

	$q(\mathbf{X} \mid f)$	$\mathbf{f}_t$	$\mathbf{S}_t^*$	sampling
1) Factorised - linear	$q(\mathbf{X})$	$\mathbf{x}_t$	$\mathbf{S}_t$	$\mathcal{O}(T)$
2) Factorised - non-linear	$q(\mathbf{X})$	$K_{\mathbf{x}_t,\mathbf{Z}}K_{\mathbf{Z},\mathbf{Z}}^{-1}\boldsymbol{\mu}_\mathbf{u}$	$\mathbf{S}_t + \mathbf{A}_t\mathbf{C}_\mathbf{u}(\mathbf{x}_t)\mathbf{A}_t^\top$	$\mathcal{O}(T)$
3) U-factorised - non-linear	$q(\mathbf{X} \mid \mathbf{u})$	$K_{\mathbf{x}_t,\mathbf{Z}}K_{\mathbf{Z},\mathbf{Z}}^{-1}\mathbf{u}$	$\mathbf{S}_t$	$\mathcal{O}(T)$
4) Non-Factorised - non-linear	$q(\mathbf{X} \mid f)$	$f(\mathbf{x}_t)$	$\mathbf{S}_t$	$\mathcal{O}(T^4)$

Table 1: Variations of approximate posteriors. $\mathbf{C}_\mathbf{u}(\mathbf{x}_t)$ is the sparse GP posterior variance:

$$\mathbf{C}_\mathbf{u}(\mathbf{x}_t) = K_{\mathbf{x}_t,\mathbf{x}_t} + K_{\mathbf{x}_t,\mathbf{Z}}K_{\mathbf{Z},\mathbf{Z}}^{-1}(\boldsymbol{\Sigma}_\mathbf{u} - K_{\mathbf{Z},\mathbf{Z}})K_{\mathbf{Z},\mathbf{Z}}^{-1}K_{\mathbf{Z},\mathbf{x}_t}$$

Each posterior is identified by whether it factorises the distribution between states and transition function, and by whether it is linear in the latent states (i.e. only the first one, which corresponds to a joint Gaussian over the states). The quartic sampling cost associated with the last choice of posterior ("Non-Factorised - non-linear", i.e. the full GP) derives from having to condition, at every time-step, on a sub-sequence of length $t \in \{0, \dots, T-1\}$, giving T operations, each of $\mathcal{O}(t^3)$ complexity (i.e. inverting a $t \times t$ matrix). The cost of sampling is crucial, because evaluating and optimising our variational bound requires obtaining samples of $q(\mathbf{x}_t)$ and $q(f, \mathbf{x}_t)$ to compute expectations by Monte Carlo integration. We now review two options to side-step the quartic cost associated with the fully non-parametric GP.

3.1. Dependence on the entire process f - "Chunking"

If we wish to retain the full GP as well as a non-factorised posterior, but the data does not come in short independent sequences, one approach is to "cut" the posterior $q(\mathbf{X}|f)$ into sub-sequences of lengths $\tau_1, \dots, \tau_n$:

$$q(\mathbf{X} \mid f) = \left[q(\mathbf{x}_1)\prod_{t=1}^{\tau_1-1} q(\mathbf{x}_{t+1} \mid f, \mathbf{x}_t)\right]\left[q(\mathbf{x}_{\tau_1+1})\prod_{t=\tau_1+1}^{\tau_1+\tau_2-1} q(\mathbf{x}_{t+1} \mid f, \mathbf{x}_t)\right]\dots\left[\dots\right] \tag{7}$$

for $\tau_1 = \dots = \tau_n = \tau$ this reduces the cost of sampling to $\mathcal{O}\big(\frac{T}{\tau}\tau^3\big)$. Conditioning (which has cubic cost) now only needs to extend as far back as the beginning of the current chunk, where the marginal $q(\mathbf{x}_i)$ is explicitly represented and can be sampled directly. Moreover we can now "minibatch" over different chunks, evaluating our bound in a "doubly stochastic manner" (Salimbeni and Deisenroth, 2017).

3.2. Dependence on the inducing points u only - "U-factorisation"

In order to avoid "cutting" long-ranging temporal dependences in our data by "chunking", we could instead use the inducing points to represent the dependence between $\mathbf{X}$ and f. We can take $q(\mathbf{X} \mid f) = q(\mathbf{X} \mid \mathbf{u})$ and, constraining $q(\mathbf{X} \mid \mathbf{u})$ to be Markovian (which is not optimal in general, but is required for efficiency), we write:

$$q(\mathbf{X} \mid f) = q(\mathbf{x}_1)\prod_{t=1}^{T-1} q(\mathbf{x}_{t+1} \mid \mathbf{u}, \mathbf{x}_t) \triangleq \mathcal{N}\big(\boldsymbol{\mu}_{\mathbf{x}_1}, \boldsymbol{\Sigma}_{\mathbf{x}_1}\big)\prod_{t=1}^{T-1}\mathcal{N}\Big(\mathbf{A}_t K_{\mathbf{x}_t,\mathbf{Z}}K_{\mathbf{Z},\mathbf{Z}}^{-1}\mathbf{u} + \mathbf{b}_t, \mathbf{S}_t\Big) \tag{8}$$

The intuition behind this choice of posterior (corresponding to the third one in the table) is to represent the GP with samples from the inducing point variational distribution (each sample effectively being a different transition function), and to generate trajectories $\mathbf{X}$ for each of the samples. Because these functions are represented parametrically (with finite "resolution" corresponding to the number of inducing points), and our posterior is Markovian, our sampling complexity does not grow with the length of the trajectory. If we were to "integrate out" $\mathbf{u}$ (as in the second posterior in the table), however, the dependence between $\mathbf{X}$ and f would be severed (recall $\mathbf{u}$ is "part" of f).

4. Experiments

In order to test the effect of the factorisation on the approximate GP posterior (i.e. the learned dynamics), we perform inference on data[2] generated by the "kink" transition function (see Figure 1's "true function"). The models whose fit is shown are "Factorised - non-linear" and "U-factorised - non-linear", both using an RBF kernel. Factorisation leads to an overconfident, mis-calibrated posterior and this is the same for both factorised models (they gave virtually the same fit). Using 100 inducing points, the U-factorised and Non-Factorised posteriors were also indistinguishable, the transition function being precisely "pinned-down" by the inducing points.

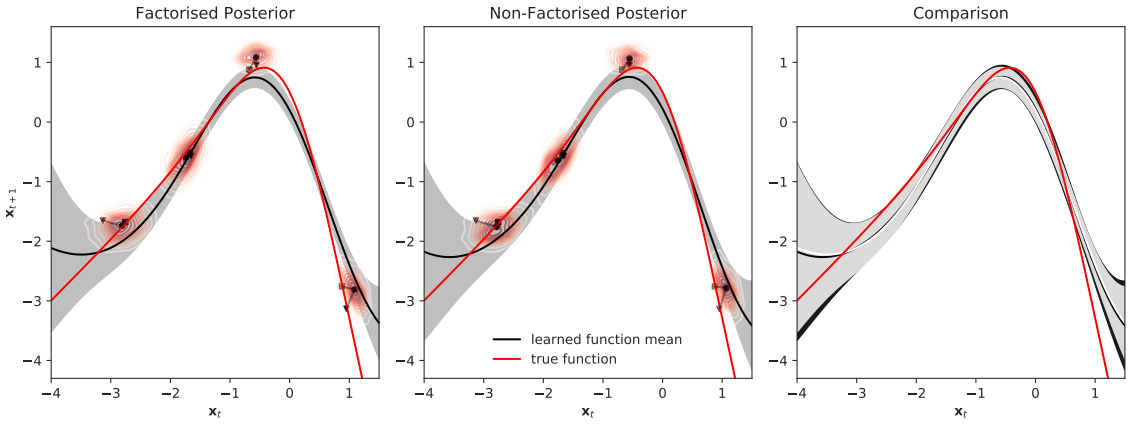

Figure 1: Uncertainty shrinking due to factorisation. **Left**: fit for the factorised $q(\mathbf{X})q(f)$; **Middle**: fit for the U-factorised $q(\mathbf{X} \mid \mathbf{u})q(f)$; **Right**: superimposed posteriors (grey is factorised, black is U-factorised). **All subplots**: the shaded regions indicate a 2σ confidence interval and the contour plots (left and middle) show the pairwise posteriors $q(\mathbf{x}_t, \mathbf{x}_{t+1})$ over some test latent states. The (x, y) coordinates of the squares, triangles and circles correspond to, respectively: the true latent states $(\mathbf{x}_t, \mathbf{x}_{t+1})$, the observed states $(\mathbf{y}_t, \mathbf{y}_{t+1})$ and the means of the marginal posteriors $(\mathbb{E}_{q(\mathbf{x}_t)}[\mathbf{x}_t], \mathbb{E}_{q(\mathbf{x}_{t+1})}[\mathbf{x}_{t+1}])$.

2. A sequence of 50 steps was generated with Gaussian emission and process noise (standard deviations of $\sqrt{0.1}$ and $\sqrt{0.01}$ respectively). The emission model was fixed to the generative one to allow comparisons.

References

Andreas Doerr, Christian Daniel, Martin Schiegg, Duy Nguyen-Tuong, Stefan Schaal, Marc Toussaint, and Sebastian Trimpe. Probabilistic recurrent state-space models. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Stefanos Eleftheriadis, Tom Nicholson, Marc Deisenroth, and James Hensman. Identification of gaussian process state space models. In *Advances in Neural Information Processing Systems 30*, 2017. URL http://papers.nips.cc/paper/7115-identification-of-gaussian-process-state-space-models.pdf.

Roger Frigola. Bayesian time series learning with gaussian processes. *University of Cambridge*, 2015.

Roger Frigola, Fredrik Lindsten, Thomas B. Schön, and Carl E. Rasmussen. Bayesian inference and learning in Gaussian process state-space models with particle MCMC. In *Advances in Neural Information Processing Systems 26*, 2013. URL http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips26/1449.pdf.

Roger Frigola, Yutian Chen, and Carl Edward Rasmussen. Variational gaussian process state-space models. In *Advances in Neural Information Processing Systems 27*, 2014. URL http://papers.nips.cc/paper/5375-variational-gaussian-process-state-space-models.pdf.

James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian Processes for Big Data. *Uncertainty in Artificial Intelligence*, 2013.

Alessandro Davide Ialongo, Mark van der Wilk, and Carl Edward Rasmussen. Closed-form inference and prediction in Gaussian process state-space models. *NIPS 2017 Time-Series Workshop*, 2017.

Alexander Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the kullback-leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.

Andrew McHutchon et al. *Nonlinear modelling and control using Gaussian processes*. PhD thesis, PhD thesis, University of Cambridge UK, Department of Engineering, 2014.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

Michalis Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. *Artificial Intelligence and Statistics*, 2009.

Ryan Turner, Marc Deisenroth, and Carl Rasmussen. State-space inference and learning with gaussian processes. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

Appendix A. Comparison to PR-SSM

Doerr et al. (2018) were, to the best of our knowledge, the first to consider a non-factorised variational posterior for the GPSSM[3]. Their work, however, has two significant shortcomings. Firstly, the $q(\mathbf{x}_{t+1}|f, \mathbf{x}_t)$ terms are set to be the same as the corresponding prior terms (i.e. the prior transitions). This posterior fails to exploit information contained in the observations other than by adapting $q(\mathbf{u})$ (it performs no filtering or smoothing on the latent states), and can only be an adequate approximation when the process noise is low and the observed sequence is short (as even low noise levels can compound, and potentially be amplified, in a long sequence). Of course, it would also be an appropriate choice (if somewhat difficult to optimise) when the process noise is zero, but then the latent variables become deterministic, given the transition function, and it is unclear whether modelling them explicitly through a probabilistic state-space model should be beneficial (an auto-regressive model with no latent variables might be sufficient).
Secondly, Doerr et al. (2018) employ a sampling scheme which gives incorrect marginal samples of $q(\mathbf{x}_t)$, even assuming, as they did, that:

$$p(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{T-1})|\mathbf{u}) = \prod_{t=1}^{T-1} p(f(\mathbf{x}_t)|\mathbf{u}) \tag{9}$$

The mistake is predicated on believing that the factorisation in equation 9 produces a Markovian $q(\mathbf{X})$. In general, this is not the case. A mismatch is thus introduced between the form of the variational lower bound and the samples being used to evaluate it, resulting in a spurious objective.

3. They call their approach PR-SSM.