SEQUENTIAL INFERENCE METHODS FOR NON-HOMOGENEOUS POISSON PROCESSES WITH STATE-SPACE PRIOR

Chenhao Li

University of Cambridge
Department of Engineering
Cambridge, UK

Simon J. Godsill

University of Cambridge
Department of Engineering
Cambridge, UK

ABSTRACT

The Non-homogeneous Poisson process is a point process with time-varying intensity across its domain, the use of which arises in numerous areas in signal processing and machine learning. However, applications are largely limited by the intractable likelihood function and the high computational cost of existing inference schemes. We present a sequential inference framework that utilises generative Poisson data and sequential Markov Chain Monte Carlo (SMCMC) algorithm to enable online inference in various applications. The proposed model is compared to competing methods on synthetic datasets and tested with real-world financial data.

Index Terms— Bayesian inference, non-homogeneous Poisson process, state-space model, sequential-MCMC.

1. INTRODUCTION

Since the original development of the Poisson process, the model has been widely applied to various applications with point data in both temporal and spatial settings, such as neuronal spike trains [1], emissions from radioactive sources and traders' operations in an open limit order book market [2]. Poisson processes have provided the possibility for describing the intensity of occurrences, which is useful in both the study of behaviour of the stochastic process itself and in the prediction of future occurrences. The non-homogeneous Poisson process (NHPP), an important variant of the standard Poisson process, allows the intensity of the process to vary with time, giving more general and realistic modelling of the data.

Intensity inference for the NHPP has become an important research topic. The classic approach developed in early years [3] uses kernel densities to construct a non-parametric estimator with empirical choice of bandwidth. The authors of [4] presents the first tractable approach for performing Bayesian inference on the Gaussian Cox process with a sigmoid transformation function. Based on these ideas, the model was extended in [5] to include variational sampling of some hyperparameters and parallel inference for multiple correlated processes. Both [4] and [6] scale poorly with the number of data due to the high computational complexity of the Gaussian process prior. Inspired by sparse Gaussian process models, [5] uses *inducing* points to perform tractable variational inference with the standard likelihood function, which achieves reduced complexity.

In this paper, we propose a new variant of the Cox process, in which the intensity function is a transformation of the state-space model (SSM). We then explore the SMCMC algorithm [7][8] as the inference method under our formulated sequential framework. Furthermore, we extend the standard SMCMC algorithm with a mixture sampling scheme and sequential batch scheme to improve both accuracy and computational efficiency. Results obtained from both synthetic datasets and a real dataset are shown and discussed in Section 4 with an analysis on hyperparameter settings.

2. THE MODEL

In this section, we briefly describe the NHPP before reviewing the model in [4] as the Sigmoidal Gaussian Cox Process (SGCP), which achieves tractable Bayesian inference via thinning from a generative prior. Later we present our model that allows for the first time sequential online inference of the intensity function, by integrating the SSM and SMCMC algorithm into the generative framework of the SGCP model.

2.1. The Non-homogeneous Poisson Process

For a domain $\mathcal{S} = \mathbb{R}^D$ of arbitrary dimension D, we can define a non-homogeneous Poisson process with an intensity function $\lambda(s)$, $s \in \mathcal{S}$, for which the counting measure $N(\mathcal{T})$ evaluated over a subregion $\mathcal{T} \subset \mathcal{S}$ follows a Poisson distribution with parameter $\lambda_{\mathcal{T}} = \int_{\mathcal{T}} \lambda(s)\, \mathrm{d}s$. Moreover, the number of events in any disjoint subsets $\{\mathcal{T}_i\}_i \subset \mathcal{S}$ are independent random variables. With this definition, we can obtain the likelihood function [9] of such a model given a set of K input

events denoted as $\{s_k\}_{k=1}^K$ in a region $\mathcal{T}$ as:

$$p(\{s_k\}_{k=1}^K \,|\, \lambda(s), \mathcal{T}) = \exp\Big\{ - \int_{\mathcal{T}} \mathrm{d}s\, \lambda(s) \Big\} \prod_{k=1}^{K} \lambda(s_k) \quad (1)$$

2.2. Thinning & Tractable Joint Distribution

Inference for the posterior intensity is clearly intractable using the likelihood in (1). However, a *thinning* procedure provides a tractable approach for simulating a NHPP from a particular intensity function $\lambda(s)$ [10], and this idea forms the basis of the SGCP inference method presented by [4]. A sequence of events $\{s_n\}_{n=1}^N$ is generated from a homogeneous Poisson process (HPP) with intensity $\lambda^* \geq \lambda(s)$, $s \in \mathcal{T}$, which is an upper bound on the desired NHPP intensity function $\lambda(s)$. Points from the desired NHPP can then be generated unbiasedly by thinning the homogeneous Poisson points independently with probability $\lambda(s_n)/\lambda^*$. This constructive generation process leads to a tractable augmented likelihood function. If we now include a mapping from a real-valued stochastic process $\{g(t),\, t \in \mathcal{T}\}$ to the intensity function, for example the scaled sigmoidal function $\lambda(s) = \lambda^* \sigma(g(s))$, then a prior may be included over the latent values of $\{g(t)\}$. Define $i_n \in \{0, 1\}$ as an indicator associated with each Poisson event, taking value 0 for a event selected by the thinning process, and 1 for a 'latent' event rejected by the thinning process. A tractable joint likelihood is now obtained as:

$$\begin{aligned} &p(\{s_n\}_{n=1}^N, \mathbf{g}_{1:N}, \{i_n\}_{n=1}^N \,|\, \lambda^*, \mathcal{T}) \\ =&(\lambda^*)^N \,\mathrm{e}^{-\lambda^* |\mathcal{T}|}\, p(\mathbf{g}_{1:N} | \{s_n\}_{n=1}^N) \prod_{n=1}^{N} \sigma\{(-1)^{i_n} g(s_n)\} \end{aligned} \quad (2)$$

where $\{s_n\}_{n=1}^N$ is the time-ordered list of homogeneous Poisson points and $\mathbf{g}_{1:N}$ is the vector of corresponding $g(s_n)$ stochastic process values, with $p(\mathbf{g}_{1:N} ...)$ its prior probability. In [4] a Gaussian process prior [11] is adopted for $p(\mathbf{g}_{1:N} ...)$ and batch-based MCMC methods are applied for inference. In our new model however a continuous-time SSM is adopted and sequential Monte Carlo inference methods are provided, which allow for on-line inference as new data points arrive.

2.3. New Model

Despite the tractability achieved in the SGCP method [4], its application is limited by its computational load, which demands $\mathcal{O}(N^3)$ for the Gaussian process [11], and scales poorly with the value of λ^* (choice of λ^* has also been demonstrated to impact heavily the accuracy of the method). In order to alleviate these limitations, we propose a new model under the tractable framework which allows efficient sequential Bayesian inference for the intensity function. We replace the fully correlated Gaussian process prior with a continuous-time SSM. This structure has the benefit that the prior is readily computed for arbitrary sets of Poisson points $\{s_n\}$ and also has a sequential Markovian property that avoids the $O(N^3)$ computational complexity and aids sequential inference formulations. We employ a SSM of the form $d\mathbf{g}(t) = \mathbf{A}\mathbf{g}(t)dt + \mathbf{h}dW(t)$ where $\{W(t)\}$ is the Wiener process. Under this framework transition densities are readily computed to be Gaussian and Markovian, $p(\mathbf{g}(Q)|\mathbf{g}(P)) = \mathcal{N}(\mu(Q,P), C(Q,P))$ for $Q > P$, see e.g. [12], and hence the joint prior $p(\mathbf{g}_{1:N})$ is readily computed by the probability chain rule.

While any Gaussian SSM could be applied in our framework, we adopt a Langevin dynamics model similar to that in [12] in which $\mathbf{g}_t = [g_{1,t}\ g_{2,t}]^T$ contains a stochastic trend term $g_{2,t}$ that is integrated to give the value $g_{1,t}$ which is input to the sigmoidal function of the SGCP framework. In this model $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & \theta \end{bmatrix}$ and $\mathbf{h} = [0\ \sigma]^T$, where $\theta \geq 0$ is a mean-reverting term on the trend and σ is a scale parameter for the noise $W(t)$. The joint prior under this model is then computed as in [12] and denoted $p(\mathbf{g}_{1:N}|\{s_n\}_{n=1}^N) = \mathcal{LD}(\mathbf{g}_{1:N}|\{s_n\}_{n=1}^N)$.

3. SEQUENTIAL INFERENCE

Inference can be made sequential by inputting short batches of data, delineated by times t_k, $i = 0, 1, ...$ The kth time interval is $\mathcal{E}_k = (t_{k-1}, t_k] \subset \mathcal{T}$, and for example we can use regularly spaced intervals $t_k = k\delta_T$, or we can space according to the arrival times of observed points $\{s_n;\, i_n = 0\}$. We further define the notation $\{s, \mathbf{g}, i\}_{\mathcal{E}_k} = \{s_n, \mathbf{g}_n, i_n;\, s_n \in \mathcal{E}_k\}$ as the locations, the state vectors and the indicators corresponding to all events in the interval $\mathcal{E}_k$. We can hence write the joint distribution recursion as:

$$\begin{aligned} p(\{s, \mathbf{g}, i\}_{\mathcal{E}_{1:k}} | \lambda^*, \mathcal{E}_{1:k}) = p(\{s, \mathbf{g}, i\}_{\mathcal{E}_{1:k-1}} | \lambda^*, \mathcal{E}_{1:k-1}) \\ \times\, p(\{s, \mathbf{g}, i\}_{\mathcal{E}_k} | \lambda^*, \{s, \mathbf{g}, i\}_{\mathcal{E}_{k-1}}, \mathcal{E}_k) \end{aligned} \quad (3)$$

where the conditional propagation can be factorised based on (2) as:

$$\begin{aligned} p(\{s, \mathbf{g}, i\}_{\mathcal{E}_k} | \lambda^*, \{s, \mathbf{g}, i\}_{\mathcal{E}_{k-1}}, \mathcal{E}_k) = (\lambda^*)^{N_k} \mathrm{e}^{-\lambda^* |\mathcal{E}_k|} \\ \times\, \mathcal{LD}(\{\mathbf{g}\}_{\mathcal{E}_k} | \{\mathbf{g}\}_{\mathcal{E}_{k-1}}) \times \prod_{n: s_n \in \mathcal{E}_k} \sigma\{(-1)^{i_n} g_{1,n}\} \end{aligned} \quad (4)$$

where $N_k = |\{n;\, s_n \in \mathcal{E}_k\}|$ is the total number of events in $\mathcal{E}_k$. A standard (variable rate) particle filter can be used for inference but is not efficient as the proposal in (4) involves multiple latent variables in a single propagation. We address this high-dimensional proposal with the SMCMC algorithm where samples are obtained with local or global MCMC moves followed by a Metropolis-Hastings (MH) accept-reject step as bias correction [13][14]. Furthermore, a mixture sampling procedure is adopted: at each MCMC iteration, a decision is made on performing either a joint MH proposal step with probability P_J or a sequence of individual

refinement Metropolis-within-Gibbs transitions with probability $1-P_J$. This scheme provides trade-offs between the efficiency and the accuracy of the inference.

3.1. Joint Proposal

A joint MH kernel consists of a discrete uniform draw of $\{\mathbf{g}, s\}^{(p)}_{\mathcal{E}_{k-1}}$ from the particle collection of $\mathcal{E}_{k-1}$ followed by the proposals of $\{\mathbf{g}, s\}_{\mathcal{E}_k}$ conditioned on the drawn particle p. More specifically, the latter is achieved by three samplers in combination: 1) the total number of thinned events $\tilde{M}$ in $\mathcal{E}_k$ from a Poisson distribution; 2) the locations of thinned events $\{\tilde{s}_m\}_{m=1}^{\tilde{M}}$ uniformly from $\mathcal{E}_k$; and 3) state vectors $\{\mathbf{g}\}_{\mathcal{E}_k}$ of all events in $\mathcal{E}_k$ from the $\mathcal{LD}$ prior conditioned on all event locations $\{s\}_{\mathcal{E}_k}$ and the drawn particle $\{\mathbf{g}, s\}^{(p)}_{\mathcal{E}_{k-1}}$. This joint proposal has density:

$$q_J = \frac{\text{Poi}(\tilde{M} \mid \lambda^*, \mathcal{E}_k)}{|\mathcal{E}_k|^{\tilde{M}}} \mathcal{LD}(\{\mathbf{g}\}_{\mathcal{E}_k} | \{\mathbf{g}\}_{\mathcal{E}_{k-1}}) \tag{5}$$

Incorporating (3) and (4), we obtain the MH acceptance ratio for the jth MCMC iteration as:

$$\rho_J = \frac{q_J^{j-1}(\lambda^*)^{N_k^*} \mathcal{LD}(\{\mathbf{g}\}^*_{\mathcal{E}_k} | \{\mathbf{g}\}^*_{\mathcal{E}_{k-1}}) \prod_n \sigma\{(-1)^{i_n^*} g^*_{1,n}\}}{q_J^*(\lambda^*)^{N_k^{j-1}} \mathcal{LD}(\{\mathbf{g}\}^{j-1}_{\mathcal{E}_k} | \{\mathbf{g}\}^{j-1}_{\mathcal{E}_{k-1}}) \prod_n \sigma\{(-1)^{i_n^{j-1}} g^{j-1}_{1,n}\}} \tag{6}$$

where the superscript $*$ indicates the variables proposed in current iteration and j–1 indicates the accepted variables from last iteration. N_{burn} iterations are run before before include the accepted particle into new particle set to ensure convergence. Tuning proposals with domain knowledge can certainly improve convergence rate and performance.

3.2. Refinement Metropolis-within-Gibbs

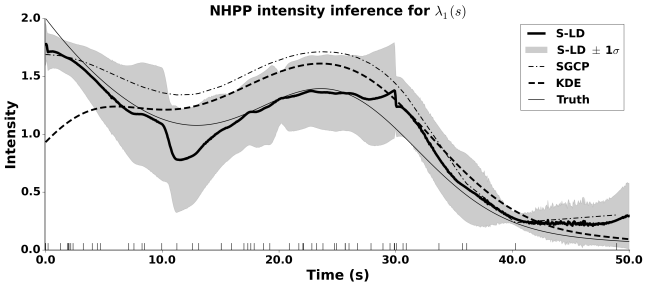
A Joint proposal can sometimes result in low acceptance ratios and ultimately low particle diversity. We hence use three local conditional samplers to individually move the three types of latent variables. Taking advantages from the SGCP model, we use a similar construction as presented in [4]: a MH move to perturb the number of thinned events $\tilde{M}$ in $\mathcal{E}_k$; a MH move to change the locations of the thinned events $\{\tilde{s}_m\}_{m=1}^{\tilde{M}}$ conditioned on $\tilde{M}$; and a Metropolis-Adjusted-Langevin-Algorithm (MALA) for state vectors $\{\mathbf{g}\}_{\mathcal{E}_k}$ to make efficient use of the gradient information available. The refinement moves improve inference accuracy at a cost of extra computation.

3.3. Sequential Batch Scheme

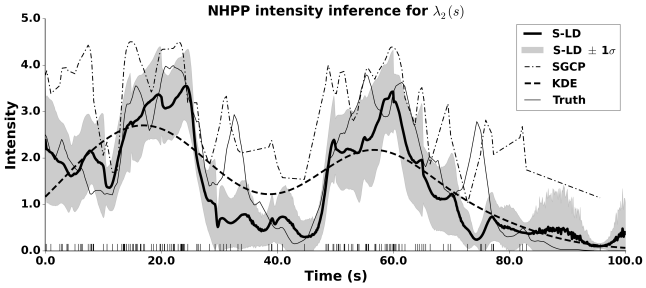
We defined earlier the interval $\mathcal{E}_k$. If the times of the batch intervals are fixed at the observed event times then this can be non-ideal for inference as it decouples the local location information among non-neighbouring points. Hence, in this section we experiment with regular sized batches of duration δT. With an appropriate choice of the batch size, the scheme recovers the temporal correlation among points in the same batch and thus improves the sequential inference accuracy. Furthermore by placing a Gamma prior over the maximum intensity λ^*, we obtain a posterior Gibbs update of the local λ_k^* for each batch, which provides considerable computational saving compared to a global λ^*: $\alpha_{\text{post}} = \alpha + N_k$, $\beta_{\text{post}} = \beta + |\mathcal{E}_k|$.

Table 1. Numerical results for models. Bold is the best.

		S-LD	KDE	SGCP
$\lambda_1(s)$	mse	**0.0257**	0.129	0.0704
	$\mathcal{L}(p)$	**1.825**	–	-9.440
	Time (s)	15.86	**0.01**	60.23
$\lambda_2(s)$	mse	**0.6531**	0.8599	1.5257
	$\mathcal{L}(p)$	**-248.1**	–	-326.6
	Time (s)	60.05	**0.05**	1326.28



(1) Data and model fits for $\lambda_1(s)$



(2) Data and model fits for $\lambda_2(s)$

Fig. 1. Model performances on synthetic datasets

4. RESULTS AND DISCUSSIONS

In this section, we present three empirical analyses of our sequential-Langevin (S-LD) model. We use synthetic datasets with ground truth $\lambda(s)$ to compare the relative performance of S-LD, SGCP and a baseline kernel density estimation (KDE) method [3]. We then apply our model to a financial dataset with high frequency input of events. Finally, we examine the effects of hyperparameters on the S-LD model performance.

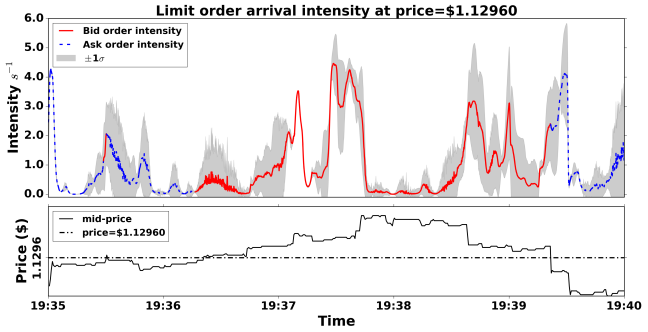


Fig. 2. Results of the S-LD model on limit order book data. The top plot shows the inferred limit order arrival intensity at price \$1.12960. The bottom plot shows the market mid-price and the selected price level for the same duration

4.1. Synthetic Data

Two sets of one-dimensional data are created with the following intensity functions:

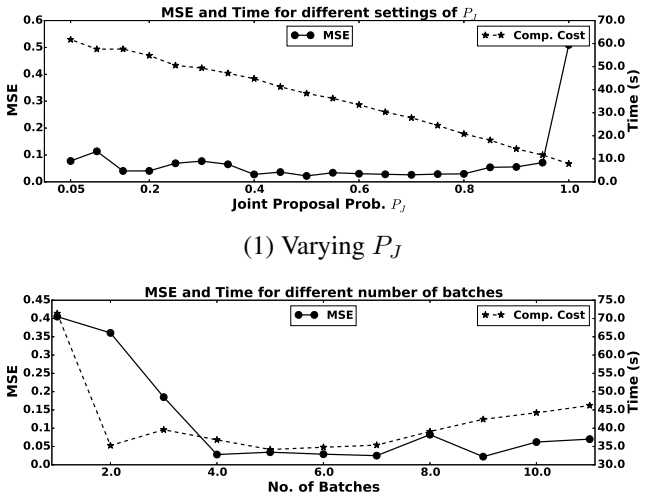
1. A sum of an exponential and a Gaussian bump: $\lambda_1(s) = 2\exp\{-s/15\} + \exp\{-((s-25)/10)^2\}$ on the interval [0, 50] with 55 events.
2. A doubly-stochastic process with $\lambda_2(s)$ governed by Langevin dynamics with parameters $\theta = -0.5$, $\sigma = 0.5$ on interval [0, 100] with 156 events.

We compared the S-LD model to both the SGCP model with a square-exponential covariance function and the KDE approach with Gaussian smoothing kernel. Fig.1 shows the graphical results of the three models and Table.1 quantitatively reports the performance averaged across 10 trials in terms of the mean squared error (MSE), the (log) likelihood of truth under Gaussian assumption and computational time. Visually and numerically, the S-LD outperforms the other two methods in both MSE and likelihood. Computational-wise, our model gives reasonable computational costs even with high number of input points whilst the SGCP model is significantly more expensive with $\lambda_2(s)$. The KDE gives the best computational speed but coarse estimations of the intensity overall.

4.2. Application to Order Book Data

We ran the S-LD model on a small set of limit order book (LOB) data taken from the EUR-USD FOREX market on the 2nd of September 2015[1]. The data describes the limit order arrivals at a fixed price level of \$1.12960 for a duration of 5 minutes (19:35–19:40) from one of the busiest hour of the day. The specific interval and price is picked intentionally to show several transitions between bid and ask sides. Fig.2 shows the outcome from the S-LD model. Note the SGCP model is far too expensive on this dataset.

[1]The authors would like to thank Cambridge Capital Management for providing the datasets for these experiments.



(1) Varying P_J

(2) Varying the number of batches

Fig. 3. The S-LD model performances under different settings of hyperparameters under synthetic dataset $\lambda_1(s)$

4.3. Hyperparameter Settings

This analysis focuses on the scheme-related hyperparameters instead of the hyperparameters of the SSM as the latter can be made adaptive with extension of a variational structure [15]. The two plots in Fig.3 show how computational time and MSE change with different settings of P_J and the number of batches respectively while other hyperparameters are held fixed. The inference accuracy is improved considerably upon the use of the refinement samplers. Reducing P_J linearly increases the computation time and eventually raises the MSE due to the lack of sampling from previous particle collection. The sequential batch scheme improves both accuracy and efficiency of the inference. However, too many batches give similar outcomes as the pointwise propagation scheme. The sudden jump between $P_J = 1.0$ and 0.9 indicates that the refinement procedure significantly improves the acceptance ratio by proposing samples in a Gibbs manner.

5. SUMMARY

We have introduced a novel sequential method for inference about the intensity function in a NHPP. By avoiding the intractability with generative prior and latent variables, our model has demonstrated improved accuracy and efficiency compared to the KDE and SGCP approaches. The sequential framework utilising SMCMC algorithm allows our model to stand in contrast to other approaches for the NHPP in that it provides online inference and the possibility of auto-adaptation of hyperparameters to input data, which can be achieved with a variational structure. The sequential batch scheme further improves the model performance by restoring the local location information, which is crucial in the inference of NHPP but usually ignored in standard sequential inference methods.

6. REFERENCES

[1] D. Perkel, G. Gerstein, and G. Moore, "Neuronal spike trains and stochastic point processes: II. Simultaneous spike trains," *Biophysical journal*, vol. 7, no. 4, pp. 419–440, 1967.

[2] M. Avellaneda and S. Stoikov, "High-frequency trading in a limit order book," *Quantitative Finance*, vol. 8, no. 3, pp. 217–224, 2008.

[3] P. Diggle, "A kernel method for smoothing point process data," *Applied statistics*, pp. 138–147, 1985.

[4] R. Adams, I. Murray, and D. MacKay, "Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 9–16.

[5] C. Lloyd, T. Gunter, M. Osborne, and S. Roberts, "Variational inference for Gaussian process modulated Poisson processes," in *International Conference on Machine Learning*, 2015, pp. 1814–1822.

[6] T. Gunter, C. Lloyd, M. Osborne, and S. Roberts, "Efficient Bayesian nonparametric modelling of structured point processes," *arXiv preprint arXiv:1407.6949*, 2014.

[7] S. K. Pang, J. Li, and S. Godsill, "Detection and tracking of coordinated groups," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 47, no. 1, pp. 472–502, 2011.

[8] A. Golightly and D. Wilkinson, "Bayesian sequential inference for nonlinear multivariate diffusions," *Statistics and Computing*, vol. 16, no. 4, pp. 323–338, 2006.

[9] D. Cox and V. Isham, *Point processes*, vol. 12, CRC Press, 1980.

[10] P. Lewis and G. Shedler, "Simulation of nonhomogeneous Poisson processes by thinning," *Naval Research Logistics (NRL)*, vol. 26, no. 3, pp. 403–413, 1979.

[11] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*, vol. 1, MIT press Cambridge, 2006.

[12] H. Christensen, J. Murphy, and S. Godsill, "Forecasting high-frequency futures returns using online Langevin dynamics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 4, pp. 366–380, 2012.

[13] S. K. Pang, S. Godsill, J. Li, F. Septier, and S. Hill, *Sequential inference for dynamically evolving groups of objects*, chapter 12, pp. 245–276, Bayesian Time Series Models. 2011.

[14] F. Septier and G. Peters, "Langevin and Hamiltonian based sequential MCMC for efficient Bayesian filtering in high-dimensional spaces," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 312–327, 2016.

[15] Z. Ghahramani and G. Hinton, "Variational learning for switching state-space models," *Neural computation*, vol. 12, no. 4, pp. 831–864, 2000.