Scalable Bayesian Learning for State Space Models using Variational Inference with SMC Samplers

Marcel Hirt
University College of London, UK

Petros Dellaportas
University College of London, UK,
Athens University of Economics and Business, Greece
and The Alan Turing Institute, UK

Abstract

We present a scalable approach to performing approximate fully Bayesian inference in generic state space models. The proposed method is an alternative to particle MCMC that provides fully Bayesian inference of both the dynamic latent states and the static parameters of the model. We build up on recent advances in computational statistics that combine variational methods with sequential Monte Carlo sampling and we demonstrate the advantages of performing full Bayesian inference over the static parameters rather than just performing variational EM approximations. We illustrate how our approach enables scalable inference in multivariate stochastic volatility models and self-exciting point process models that allow for flexible dynamics in the latent intensity function.

1 Introduction

We deal with generic state-space models (SSM) which may be nonlinear and non-Gaussian. Inference for this important and popular family of statistical models presents tremendous challenges that has prohibited their widespread applicability. The key difficulty is that inference on the latent process of the model depends crucially on unknown static parameters that need to be also estimated. While MCMC samplers are unsatisfactory because they both fail to produce high dimensional, efficiently mixing Markov chains and because they are inappropriate for on-line inference, sequential Monte Carlo (SMC) methods (Kantas et al., 2015) provide

the tools to construct successful viable implementation strategies. In particular, particle MCMC (Andrieu et al., 2010) utilises SMC to build generic efficient MCMC algorithms that provide inferences for both static parameters and latent paths. We provide a scalable alternative to these methods via an approximation that combines SMC and variational inference.

We introduce a new variational distribution that unlike recent strand of literature (Maddison et al., 2017; Naesseth et al., 2018; Le et al., 2018) performs variational inference also on the static parameters of the SSM. This is essential for various reasons. First, when there is dependency between static and dynamic parameters posterior inference may be inaccurate if the joint posterior density is approximated by conditioning on fixed values of static parameters. Second, inferring the static parameter is often the primary problem of interest: for example, for biochemical networks and models involving Lotka Voltera equations, we are not interested in the population of the species per se, but we want to infer some chemical rate constants (such as reaction rates or predation/growth rates), which are parameters of the transition density; in neuroscience, Bayesian decoding of neural spike trains is often made via a state-space representation of point processes in which inference for static parameters is of great importance. Finally, for complex dynamic systems it is often advisable to improve model compression or interpretability by encouraging sparsity and such operations may require inference for the posterior densities of the static parameters.

Sampling from the new variational distribution involves running a SMC algorithm which yields an unbiased estimate of the likelihood for a fixed static parameter value. Importantly, we show that the SMC algorithm constructs a computational graph that allows for optimisation of the variational bound using stochastic gradient descent. We provide some empirical evidence that variational inference on static parameters can give better predictive performance, either out-of sample in

the linear Gaussian state space model or in-sample for predictive distributions in a multivariate stochastic volatility model. We also illustrate our method by modelling fairly general intensity functions in a multivariate Hawkes process model.

2 Background

Let us begin by introducing the standard inference problem in a generic SSM, followed by a review of the SMC approach to sample from a sequence of distributions arising in such probabilistic structures. SSMs are characterized by a latent Markov state process $\{X_n\}_{n \geq 0}$ on $\mathbb{R}^{d_x}$ and an observable process $\{Y_n\}_{n \geq 0}$ on $\mathbb{R}^{d_y}$. We follow the standard convention of using capital letters for random variables and the corresponding lower case letter to denote their values. The dynamics of the latent states is determined, conditional on a static parameter vector $\theta \in \Theta$, by a transition probability density

$$X_n | (\theta, X_{n-1} = x_{n-1}, Y_{n-1} = y_{n-1}) \sim f_\theta(\cdot | x_{n-1}, y_{n-1}),$$

along with an initial density $X_0 \sim f_\theta(\cdot)$. The observations are assumed to be conditionally iid given the states with density given by

$$Y_n | (\theta, X_{0:n} = x_{0:n}, Y_{0:n-1} = y_{0:n-1}) \sim g_\theta(\cdot | x_n),$$

for any $n \geq 0$ with the generic notation $x_{0:n} = (x_0, ..., x_n)$.

We consider a Bayesian framework and assume θ has a prior density $p(\theta)$. Consequently, for observed data $y_{0:M}$, we perform inference using the posterior density

$$\pi(\theta, x_{0:M}) := p(\theta, x_{0:M} | y_{0:M}) \propto p(\theta) p_\theta(x_{0:M}, y_{0:M}), \quad (1)$$

where the joint density of the latent states and observations given a fixed static parameter value θ writes as

$$p_\theta(x_{0:M}, y_{0:M}) = \gamma_\theta(x_{0:M})$$
$$:= f_\theta(x_0) \prod_{n=1}^{M} f_\theta(x_n | x_{n-1}, y_{n-1}) \prod_{n=0}^{M} g_\theta(y_n | x_n). \quad (2)$$

The posterior density $p(\theta, x_{0:M} | y_{0:M})$ is in general intractable, as is

$$p_\theta(x_{0:M} | y_{0:M}) = \frac{\gamma_\theta(x_{0:M})}{p_\theta(y_{0:M})}, \quad (3)$$

where $p_\theta(y_{0:M}) = \int p_\theta(x_{0:M}, y_{0:M}) dx_{0:M}$. However, an SMC algorithm can be used to approximate (3). A brief review of how this sampling algorithm proceeds is as follows and further details can be found in Doucet et al. (2000); Doucet and Johansen (2009).
SMC methods approximate $p_\theta(x_{0:n} | y_{0:n})$ using a set of

K weighted random samples $X_{0:n}^{1:K} = (X_{0:n}^1, ..., X_{0:n}^K)$, also called particles, having positive weights $W_n = W_n^{1:K}$, so that $p_\theta(x_{0:n} | y_{0:n}) \approx \hat{p}_\theta(x_{0:n} | y_{0:n}) = \sum_{k=1}^{K} W_n^k \delta_{X_{0:n}^k}(x_{0:n})$. Here, δ denotes the Dirac delta function. To do so, one starts at $n = 0$ by sampling X_0^k from an importance density $M_0^\phi(\cdot | y_0)$, parametrized with ϕ, where ϕ can depend on the static parameters θ. For any $n \geq 1$, we first resample an ancestor variable A_{n-1}^k that represents the 'parent' of particle $X_{0:n}^k$ according to $A_{n-1}^k \sim r(\cdot | W_{n-1})$, where r is a categorical distribution on $\{1, ..., K\}$ with probabilities W_{n-1}. We then set $W_{n-1} = \frac{1}{K}$ and proceed by extending the path of each particle by sampling from a transition kernel $X_n^k \sim M_n^\phi(\cdot | y_n, X_{0:n-1}^{A_{n-1}^k})$. This yields an updated latent path $X_{0:n}^k = (X_{0:n-1}^{A_{n-1}^k}, X_n^k)$ for which we compute the incremental importance weight

$$\alpha_n(X_{0:n}^k) = \frac{\gamma_\theta(X_{0:n}^k)}{\gamma_\theta(X_{0:n-1}^k) M_n^\phi(X_n^k | y_n, X_{0:n-1}^{A_{n-1}^k})}.$$

We set $w_n(X_{0:n}^k) = W_{n-1}^k \alpha_n(X_{0:n}^k)$ as well as $W_n^k = \frac{w_n(X_{0:n}^k)}{\sum_l w_n(X_{0:n}^l)}$ and define

$$\hat{Z}_n^{\theta, \phi} := \prod_{m=0}^{n} \sum_{k=1}^{K} w_m(X_{0:m}^k),$$

which is an unbiased and strongly consistent estimator of $p_\theta(y_{0:n})$, see Del Moral (1996). A pseudo-code (Algorithm 1) for this standard SMC sampler can be found in Appendix A. It is possible to perform the resampling step only if some condition on W_{n-1} is satisfied, see Algorithm 1. For simplicity, we assume that the particles are resampled at every step. The density of all variables generated by this SMC sampler for a fixed static parameter value θ is given by

$$q_\phi(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l | \theta) = w_M^l \prod_{k=1}^{K} M_0^\phi(x_0^k | y_0)$$
$$\cdot \prod_{n=1}^{M} \prod_{k=1}^{K} r(a_{n-1}^k | w_{n-1}) M_n^\phi(x_n^k | y_n, x_{0:n-1}^{a_{n-1}^k}),$$

where l is a final particle index drawn from a categorical distribution with weights W_M. Since $\hat{Z}_n^{\theta, \phi}$ is unbiased, we have

$$\mathbb{E}_{q_\phi(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l | \theta)} \left[\hat{Z}_M^{\theta, \phi} \right] = p_\theta(y_{0:M}). \quad (4)$$

3 Variational bounds for state space models using SMC samplers

Variational inference (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017) allows Bayesian inference to scale to large data sets (Hoffman et al., 2013)

and is applicable to a wide range of models (Ranganath et al., 2014; Kucukelbir et al., 2017). It generally postulates a family of approximating distributions with variational parameters that minimize some divergence, most commonly the KL divergence, between the approximating distribution and the posterior. The quality of the approximation hinges on the expressiveness of the variational family.

Let $q_\psi(\theta)$ be a distribution on Θ with variational parameters ψ. We aim to approximate the posterior density $p(\theta, x_{0:M}|y_{0:M})$ in (1) with a variational distribution that results as an appropriate marginal of auxiliary variables arising from an SMC sampler of the form

$$q_{\psi,\phi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l) := q_\psi(\theta)q_\phi(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l|\theta), \tag{5}$$

defined precisely below. Note that sampling from the extended variational distribution (5) just means sampling $\theta \sim q_\psi(\theta)$ and then running a particle filter using the sampled value θ as the static parameter.

We introduce the proposed variational bound first as a lower bound on $\log p(y_{0:M}) - \mathrm{KL}(q_\psi(\theta)||p(\theta|y_{0:M}))$. We then show that optimizing the proposed bound means minimizing the KL-divergence between the extended variational distribution (5) and an extended target density that resembles closely the density targeted in particle MCMC methods.

We can write $p(\theta|y_{0:M}) = p(\theta)p_\theta(y_{0:M})/p(y_{0:M})$. Hence, using the fact that the likelihood estimator is unbiased (4) and due to Jensen's inequality,

$$- \mathrm{KL}(q_\psi(\theta)||p(\theta|y_{0:M})) + \log p(y_{0:M})$$
$$= \mathbb{E}_{q_\psi(\theta)}\left[\log p_\theta(y_{0:M}) + \log p(\theta) - \log q_\psi(\theta)\right]$$
$$= \mathbb{E}_{q_\psi(\theta)}\left[\log \mathbb{E}_{q_\phi(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l|\theta)}\left[\hat{Z}_M^{\theta,\phi}\right] + \log \frac{p(\theta)}{q_\psi(\theta)}\right]$$
$$\geq \mathbb{E}_{q_\psi(\theta)}\left[\mathbb{E}_{q_\phi(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l|\theta)}\left[\log \hat{Z}_M^{\theta,\phi}\right] + \log \frac{p(\theta)}{q_\psi(\theta)}\right]$$
$$=: \mathcal{L}(\psi, \phi).$$

In particular, $\mathcal{L}(\psi, \phi)$ is a lower bound on $p(y_{0:M}) - \mathrm{KL}(q_\psi(\theta)||p(\theta|y_{0:M}))$.

Remark 1 (Inference for multiple independent time series). *Instead of considering one latent process $\{X\}$ and observable process $\{Y\}$, we can also consider S independent latent processes $\{X^s\}_{s=1,...,S}$ with corresponding observable processes $\{Y^s\}_{s=1,...,S}$ described by the same static parameter θ. We obtain a lower bound on $p(y_{0:M}) - KL(q_\psi(\theta)||p(\theta|y_{0:M}^1, ..., y_{0:M}^S))$ given by*

$$\mathbb{E}_{q_\psi(\theta)}\left[\mathbb{E}_{\prod_s q_\phi(x_{0:M}^{s,1:K}, a_{0:M-1}^{s,1:K}, l^s|\theta)}\left[\sum_{s=1}^S \log \hat{Z}_{M,s}^{\theta,\phi}\right]\right.$$
$$\left. + \log p(\theta) - \log q_\psi(\theta)\right],$$

where $\hat{Z}_{M,s}^{\theta,\phi}$ is the estimator of $p_\theta(y_{0:M}^s)$. Note that we can obtain an unbiased estimate of this bound by sampling an element $s \in \{1, ..., S\}$ and using $S \cdot \log \hat{Z}_{M,s}^{\theta,\phi}$ as an estimate of $\sum_{s'=1}^S \log \hat{Z}_{M,s'}^{\theta,\phi}$, thereby allowing our method to scale to a large number of independent time series. For ease of exposition, we formulate our results for a single time series only.

Next, we show that the variational bound can be represented as the difference between the log-evidence and the KL divergence between the variational distribution and an extended target density. More concretely, following Andrieu et al. (2010), we consider a target density on the extended space $\Theta \times \mathcal{X}$, $\mathcal{X} := (\mathbb{R}^{d_x})^{(M+1)K} \times \{1, ..., K\}^{MK+1}$,

$$\tilde{\pi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l) := \frac{\pi(\theta, x_{0:M}^l)}{K^{M+1}}$$
$$\cdot \frac{q_\phi(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l|\theta)}{M_0^\phi(x_0^{b_0^l}|y_0)\prod_{n=1}^M r(b_{n-1}^l|w_{n-1})M_n^\phi(x_n^{b_n^l}|y_n, x_{0:n-1}^{b_{n-1}^l})}.$$

Here, we have defined $b_M^l = l$ and $b_n^l = a_n^{b_{n+1}^l}$ for $n = M-1, ..., 1$, i.e. b_n^l is the index that the ancestor of particle $X_{0:M}^l$ at generation n had. It follows, using $r(b_n^l|w_{n-1}) = w_{n-1}^{b_{n-1}^l}$, that the ratio between the extended target density and the variational distribution is given by

$$\frac{\tilde{\pi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l)}{q_{\phi,\psi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l)}$$
$$= \frac{K^{-(M+1)}p(\theta)p_\theta(x_{0:M}^l, y_{0:M})/p(y_{0:M})}{q_\psi(\theta)W_M^l M_0^\phi(x_0^{b_0^l}|y_0)\prod_{n=1}^M W_{n-1}^{b_{n-1}^l}M_n^\phi(x_n^{b_n^l}|y_n, x_{0:n-1}^{b_{n-1}^l})}. \tag{6}$$

Proposition 2 (KL divergence in extended space). *It holds that*

$$\mathcal{L}(\psi, \phi) = -KL(q_{\phi,\psi}||\tilde{\pi}) + \log p(y_{0:M}).$$

The proof can be found in Appendix B. Recall that we have introduced $\mathcal{L}(\psi, \phi)$ so that its maximisation pushes the variational approximation of the static parameter θ closer to its true posterior as measured by the KL divergence. The above proposition shows that this objective also minimizes the KL divergence between densities on an extended space that includes multiple latent paths. To elucidate further the relation between the variational distribution of a single latent path and its posterior, we need to introduce a further distribution. Consider the density under $\tilde{\pi}$ of the variables generated by a SMC algorithm conditional on a fixed latent path $(x_{0:M}^l, b_{0:M-1}^l)$. This is known as a conditional SMC algorithm (Andrieu et al., 2010), with distribution given by

$$\tilde{\pi}_{\text{CSMC}}(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l, b_{0:M}^l)$$

$$= \frac{q_\phi(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l|\theta)}{W_M^l M_0^\phi(X_0^{b_0^l}|y_0) \prod_{n=1}^{M} r(b_{n-1}^l|W_{n-1}) M_n^\phi(x_n^{b_n^l}|y_n, x_{0:n-1}^{b_{n-1}^l})},$$

where $\neg b_{0:M}^l$ are the indices of all particles that are not equal to $b_{0:M}^l$. We obtain the following corollary proved in Appendix C.

Corollary 3 (Marginal KL divergence and marginal ELBO). *The KL divergence in the extended space is an upper bound on the KL divergence between the marginal variational approximation and the posterior, with the gap between bounds being*

$$KL\left(q_{\psi,\phi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l)||\tilde{\pi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l)\right)$$
$$- KL\left(q_{\psi,\phi}(\theta, x_{0:M})||\pi(\theta, x_{0:M}^l)\right)$$
$$= \mathbb{E}_{q_{\psi,\phi}(\theta, x_{0:M}^l, b_{0:M}^l)} \Bigg[$$
$$KL(q_\phi(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l, b_{0:M}^l)||$$
$$\tilde{\pi}_{CSMC}(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l, b_{0:M}^l))\Bigg].$$

Particularly, $\mathcal{L}$ is a lower bound compared to the standard ELBO using the marginal $q_{\psi,\phi}(\theta, x_{0:M})$ with $x_{0:M}^l = x_{0:M}$ as the variational distribution:

$$\mathcal{L}(\psi, \phi) \leq -KL\left(q_{\psi,\phi}(\theta, x_{0:M})||\pi(\theta, x_{0:M})\right) + \log p(y_{0:M}).$$

The proposed surrogate objective resembles variational bounds with auxiliary variables (Salimans et al., 2015; Maaløe et al., 2016; Ranganath et al., 2016) where the gap between the two bounds is expressed by the KL-divergence between the variational approximation of the auxiliary variable given the latent variable of interest and a so-called reverse model. Here, this reverse model is specified by the conditional SMC algorithm. The above corollary implies that the variational bound is looser than the standard ELBO with the auxiliary variables integrated out. This marginal variational distribution cannot in general be evaluated analytically. However, we can obtain unbiased estimates of it by computing the log-likelihood estimate under a conditional SMC algorithm, resembling a particle Gibbs update. This constitutes an extension of Proposition 1 in Naesseth et al. (2018). We present a proof in Appendix D.

Proposition 4 (Marginal variational distribution). *We have*

$$q_{\psi,\phi}(\theta, x_{0:M}^l, b_{0:M}^l) = q_\psi(\theta)\gamma_\theta(x_{0:M}^l)$$
$$\cdot \mathbb{E}_{\tilde{\pi}_{CSMC}(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l)}\left[\left(\hat{Z}_M^{\theta,\phi}\right)^{-1}\right]$$

and there exists $c(\theta, \phi) < \infty$ so that

$$KL(q_{\psi,\phi}(\theta, x_{0:M})||p(\theta, x_{0:M}|y_{0:M})$$
$$\leq \mathbb{E}_{q_\psi(\theta)}\left[\frac{c(\theta, \phi)}{K}\right] + KL(q_\psi(\theta)||p(\theta|y_{0:M})).$$

The last inequality in Proposition 4 is a straightforward extension of an analogous result in the EM setting (Naesseth et al., 2018). It implies that, for fixed variational parameters ψ and ϕ, the approximation becomes more accurate for increasing K. Sampling from this distribution can be seen as an extension of visualizing the expected importance weighted approximation in Importance Weighted Auto-Encoders (Cremer et al., 2017). Since this distribution can be high-dimensional, the preceding proposition gives an alternative to kernel-density estimation.

Lastly, from a different angle, the variational objective can be seen as a sequential variational-autoencoding (VAE) bound. Indeed, as a consequence of Proposition 2 and equation (6), we obtain immediately the following result. We elaborate on it further in the next section.

Corollary 5 (Sequential VAE representation). *The variational bound can be written as*

$$\mathcal{L}(\psi, \phi) = \mathbb{E}_{q_\psi(\theta)}\Bigg[\mathbb{E}_{q_\phi(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l|\theta)}\Bigg[$$
$$\sum_{n=0}^{M} \log g_\theta(y_n|x_n^{b_n^l}) - \log W_n^{b_n^l}$$
$$+ \log \frac{f_\theta(x_n^{b_n^l}|x_{n-1}^{b_{n-1}^l}, y_{n-1})}{M^\phi(x_n^{b_n^l}|y_n, x_{0:n}^{b_{n-1}^l})}\Bigg]\Bigg]$$
$$- (M+1)\log K - KL(q_\psi(\theta)||p(\theta)).$$

4 Related Work

The representation in Corollary 5 allows us to contrast the variational bound to previously considered sequential VAE frameworks (Chung et al., 2015; Archer et al., 2015; Fraccaro et al., 2016; Krishnan et al., 2017; Goyal et al., 2017). The introduced bound contains the cross-entropy between the proposal distribution and the likelihood common to sequential VAE bounds. However, this reconstruction error is only evaluated for surviving particles. Similarly, while a sequential VAE framework includes a KL-divergence between the proposal distribution and the prior transition probability, the log-ratio of these two densities is only evaluated for a surviving path. Most work using sequential VAEs have considered observation and state transition models parametrised by neural networks, and given the high-dimensionality of the static parameters, have confined their analysis to variational EM inferences. This

is also the case for the approaches in Maddison et al. (2017); Naesseth et al. (2018); Le et al. (2018), to which this work is most closely related. They have demonstrated that resampling increases the variational bound compared to a sequential IWAE (Burda et al., 2015) approach. Rainforth et al. (2018) demonstrated that increasing the number of particles leads to a worse signal to noise ratio of the gradient estimate of the proposal parameters in an IWAE setting. Le et al. (2018) suggested to use fewer particles without resampling for calculating the proposal gradient. A possible approach left for future work would be to consider a different resampling threshold for the proposal gradients. Finally, the objective in this work differs from adaptive SMC approaches optimizing the reverse KL-divergence (or χ^2-divergence) between the posterior and the proposal, cf. Cornebise et al. (2008); Gu et al. (2015).

5   Optimization of the variational bound

The gradient of the variational bound is given by

$$
\begin{aligned}
&\nabla_{\psi,\phi}\mathcal{L}(\psi,\phi) \quad\quad\quad (7)\\
&=\nabla_{\psi,\phi}\left(\mathbb{E}_{q_\psi(\theta)}\left[\mathbb{E}_{q_\phi(x_{0:M}^{1:K},a_{0:M-1}^{1:K},l|\theta)}\left[\log\hat{Z}_M^{\theta,\phi}\right]\right]\right)\\
&\quad+\nabla_\psi\left(\mathbb{E}_{q_\psi(\theta)}\left[\log\frac{p(\theta)}{q_\psi(\theta)}\right]\right).
\end{aligned}
$$

We focus on the gradient of the first expectation and note that the gradient of the second expectation can be estimated by standard (black-box) approaches in variational inference, depending of course on the chosen variational approximation. If for instance the variational distribution over the static parameters is continuously reparametrisable, one can use standard low-variance reparametrised gradients (Kingma and Welling, 2014; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014). This is the gradient estimator that we use in our experiments in combination with mean-field variational families. We assume that the proposals $X_n^k\sim M_n^\phi(\cdot|y_n,x_{0:n-1}^{a_{n-1}^k})$ are reparametrisable, i.e. there exists a differentiable deterministic function h_ϕ such that $X_n^k=h_\phi(X_{0:n-1}^{A_{n-1}^k},\epsilon_n^k)$, with $\epsilon_n^k\sim p(\cdot)$ continuous and independent of ϕ. Similarly, we assume that the variational distribution of the static parameters is reparametrisable, i.e. there exists a differentiable deterministic function h_ψ such that $\theta=h_\psi(\eta)$, with $\eta\sim p(\cdot)$ continuous and independent of ψ. We abbreviate $\boldsymbol{\epsilon}=\epsilon_{0:M}^{1:K}$, $\boldsymbol{x}=x_{0:M}^{1:K}$ and $\boldsymbol{a}=a_{0:M-1}^{1:K}$. Using the product rule, observe that the first gradient in (7) is

$$
\begin{aligned}
&\nabla_{\psi,\phi}\int p(\eta)p(\boldsymbol{\epsilon})q_\phi(\boldsymbol{a}|\theta,\boldsymbol{x})\\
&\quad\cdot\log\hat{Z}_M^{\theta,\phi}d(\eta,\boldsymbol{a},\boldsymbol{\epsilon})\bigg|_{\theta=h_\psi(\eta),\boldsymbol{x}=h_\phi(\boldsymbol{\epsilon})}\\
&=\int p(\eta)p(\boldsymbol{\epsilon})\nabla_{\psi,\phi}q_\phi(\boldsymbol{a}|\theta,\boldsymbol{x})\\
&\quad\cdot\log\hat{Z}_M^{\theta,\phi}d(\eta,\boldsymbol{a},\boldsymbol{\epsilon})\bigg|_{\theta=h_\psi(\eta),\boldsymbol{x}=h_\phi(\boldsymbol{\epsilon})}\\
&=\mathbb{E}_{p(\eta)p(\boldsymbol{\epsilon})q_\phi(\boldsymbol{a}|h_\psi(\eta),h_\phi(\boldsymbol{\epsilon}))}\bigg[\nabla_{\psi,\phi}\log\hat{Z}_M^{h_\psi(\eta),\phi}\\
&\quad+\nabla_{\psi,\phi}\log q_\phi(\boldsymbol{a}|h_\psi(\eta),h_\phi(\boldsymbol{\epsilon}))\log\hat{Z}_M^{h_\psi(\eta),\phi}\bigg].
\end{aligned}
$$

Analogously to Maddison et al. (2017); Le et al. (2018); Naesseth et al. (2018) in a variational EM framework, we have also ignored the second summand in the gradient due to its high variance in our experiments. We take Monte Carlo samples of the expectation above and optimize the bound using Adam (Kingma and Ba, 2014). It is also possible to use natural gradients (Amari, 1998), see Appendix E.

6   Experiments

6.1   Linear Gaussian state space models

Regularisation in a high-dimensional model. We illustrate potential benefits of a fully Bayesian approach in a standard linear Gaussian state space model
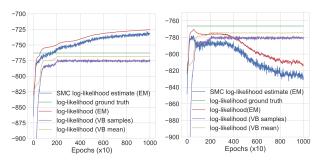
$$
\begin{aligned}
f_\theta(x_n|x_{n-1})&=\mathcal{N}(Ax_{n-1},\Sigma_x), &(8)\\
g_\theta(y_n|x_n)&=\mathcal{N}(Bx_n,\Sigma_y), &(9)
\end{aligned}
$$

with initial state distribution $X_0\sim\mathcal{N}(A^0,\Sigma_x^0)$ and parameters $A,\Sigma_x,\Sigma_x^0\in\mathbb{R}^{d_x\times d_x}$, $A^0\in\mathbb{R}^{d_x}$, $B\in\mathbb{R}^{d_x\times d_y}$, and $C,\Sigma_y\in\mathbb{R}^{d_x\times d_y}$. Naesseth et al. (2018) have shown in a linear Gaussian model that learning the proposal yields a higher variational lower bound compared to proposing from the prior and the variational bound is close to the true log-marginal likelihood for both sparse and dense emission matrices B. However, an EM approach might easily over-fit, unless one employs some regularisation, such as stopping early if the variational bound decreases on some test set. We demonstrate this effect by re-examining one of the experiments in Naesseth et al. (2018), setting $(d_x,d_y)=(10,3)$, $M=10$ and assume that Σ_x, Σ_x^0 and Σ_y are all identity matrices. Furthermore, $A^0=0$ and $(A_{ij})=\alpha^{|i-j|+1}$ with $\alpha=0.42$, and B has randomly generated elements with $B_{ij}\sim\mathcal{N}(0,1)$. We assume that the proposal density is

$$
M_{n+1}^\phi(x_{n+1}|x_n,y_{n+1})=\mathcal{N}(x_{n+1}|A_\phi x_n+B_\phi y_{n+1},\Sigma_\phi),
$$

and $M_0^\phi(x_0|y_0) = \mathcal{N}(x_0|A_\phi^0 + B_\phi y_0, \Sigma_\phi^0)$, with Σ_ϕ and Σ_ϕ^0 diagonal matrices. We perform both a variational EM approach and a fully Bayesian approach over the static parameters using $K = 4$ particles. In the latter case, we place Normal priors $B_{ij} \sim \mathcal{N}(0, 10)$ and $A_{ij} \sim \mathcal{N}(0, 1)$. Furthermore, we suppose that a priori Σ_y is diagonal with variances drawn independently from an Inverse Gamma distribution with shape and scale parameters of 0.01 each. A mean-field approximation for the static parameters is assumed. We suppose that the variational distribution over each element of A and B is a normal distribution and the approximation over the diagonal elements of Σ_y is log-normal. For identifiability reasons, we assume that Σ_x, Σ_x^0 and A^0 are known. We compare the EM and VB approach in terms of log-likelihoods on out-of-sample data assuming training and testing on 10 iid sequences. Figure 1 shows that in contrast to the VB approach, the EM approach attains a higher log-likelihood on the training data with a lower log-likelihood on the test set as the training progresses.
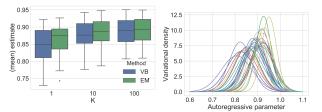

(a) Log-likelihood on train-ing data. (b) Log-likelihood on testing data.

Figure 1: Log-likelihood for linear Gaussian state space models. Log-likelihood values are computed using Kalman filtering. The static parameters used in the VB case are the mean of the variational distribution (VB mean) or the samples from the variational distribution (VB samples) as they are drawn during training.

Approximation bias in a low-dimensional model. Variational approximations for the latent path can yield biased estimates of the static parameters, see Turner and Sahani (2011). We illustrate that this bias decreases for increasing K in a two-dimensional linear Gaussian model, both in an EM and VB setting. We therefore consider inference in a linear Gaussian state space model (8-9) with two-dimensional latent states and one-dimensional observations. The state transition matrix is assumed to be determined by the autoregressive parameter λ with $A = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$. We consider inference over λ as the static parameter and

fix $B = (1, 1)$ with Σ_x and Σ_y being identity matrices. We simulate 30 realisations of length $M = 100$ each using $\lambda = 0.9$. Inference is performed with different initialisations and learning rates over the simulated datasets. It has been documented in such a linear Gaussian model, see Turner and Sahani (2011), that Gaussian variational approximations of the latent path that factorise over the state components underestimate λ. We observe the same effect in Figure 2a when using just $K = 1$ particle. However, increasing the number of particles used during inference reduces this bias. Furthermore, we find that point estimates of the static parameters show some variation over different simulations, while an approximate Bayesian approach can be argued to better account for this uncertainty. The variational distributions for θ for each of the simulations using $K = 100$ particles is shown in Figure 2b, confirming that they all put significant mass on the ground truth. Let us remark that these experiments also complement those in Le et al. (2018), where it is illustrated that increasing K improves learning point estimates of the static parameters in a Gaussian model with a one-dimensional latent state. Indeed, as shown next, the marginal variational distribution allows not just for dependencies in the latent states across time, but also across different state dimensions, even if they are independent under the proposal.

(a) Point estimate of the autoregressive parameter λ in the EM case or the variational mean in the VB case over 30 simulations for $K \in \{1, 10, 100\}$ particles. (b) Variational distribution of the autoregressive parameter λ using $K = 100$ particles for each of the 30 simulations.

Figure 2: Inference on the autoregressive parameter λ over 30 simulations of length $M = 100$. Ground truth values are $\lambda = 0.9$.
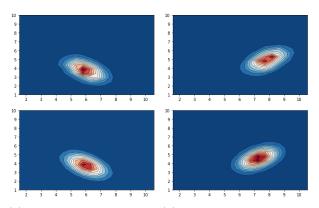
Marginal variational distribution in a low-dimensional model. In an additional experiment, we evaluate if the variational approximation from Proposition 4 of the latent path matches the distribution of its true posterior. We consider the above state space model over 2 time steps as in Turner and Sahani (2011). Note that for given static parameters, the posterior is Gaussian. Indeed, for $\boldsymbol{x} = (x_0^{(0)}, x_0^{(1)}, x_1^{(0)}, x_1^{(1)})$, where $x_n^{(i)}$ denotes dimen-

sion i of x_n, we have $p(\boldsymbol{x}|y_{0:1}, \lambda) = \mathcal{N}(\mu_{x|y}, \Sigma_{x|y})$ with

$$\Sigma_{x|y}^{-1} = \begin{pmatrix} 2 & 1 & -\lambda & 0 \\ 1 & 2 & 0 & -\lambda \\ -\lambda & 0 & 2 & 1 \\ 0 & -\lambda & 1 & 2 \end{pmatrix}, \mu_{x|y} = \Sigma_{x|y} \begin{pmatrix} y_0 \\ y_0 \\ y_1 \\ y_1 \end{pmatrix},$$

assuming $X_0 \sim \mathcal{N}(0, \frac{1}{1-\lambda^2}I)$ is drawn from its stationary distribution. We visualise the posterior distribution along with the marginal variational distribution

$$q_\phi(x_{0:M}^l|\theta)$$

$$= \gamma_\theta(x_{0:M}^l) \mathbb{E}_{\tilde{\pi}_{\text{CSMC}}(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l)} \left[\left(\hat{Z}_M^{\theta,\phi}\right)^{-1} \right]$$

in Figure 3 using $K = 100$ particles and 50 samples for the expectation. We find that the approximation mirrors the true posterior. In particular, it accounts for explaining-away between different dimensions of the latent state, although we have used isotropic proposals.

(a) Joint distribution of the latent states at the second time step. Top: variational approximation, bottom: true posterior.

(b) Joint distribution of the first state component at the first and second time step. Top: variational approximation, bottom: true posterior.

Figure 3: Two-dimensional contour plots of the distribution of the latent path over two time steps and two state components. Function arguments are set to the ground truth state values as simulated if they are not shown.

6.2 Stochastic volatility models

To show that our method allows inference of latent states and static parameters of higher dimensions, we consider a multivariate stochastic volatility model,

$$f_\theta(x_n|x_{n-1}) = \mathcal{N}(\mu + \text{diag}(a)(x_{n-1} - \mu), \Sigma_x),$$
$$g_\theta(y_n|x_n) = \mathcal{N}(0, \exp(\text{diag}(x_n))),$$

where $X_0 \sim \mathcal{N}(\mu, \Sigma_x^0)$ with $x_n, y_n, \mu, a \in \mathbb{R}^D$, and covariance matrix $\Sigma_x \in \mathbb{R}^{D \times D}$, $\theta = (\mu, a, \Sigma_x, \Sigma_x^0)$. This

Table 1: Average p-step predictive log-likelihoods per observation for the stochastic volatility model with different number of particles K and number of samples S from the variational distribution. In the EM case, we run S particle filters with the same optimal static values. Mean estimates with standard deviation in parentheses based on 100 replicates.

| | $(S, K) = (4, 50)$ | |
Method	$p = 1$	$p = 2$
EM	9.697 (0.008)	9.716 (0.008)
VB	9.707 (0.011)	9.728 (0.015)
	$(S, K) = (20, 100)$	
Method	$p = 1$	$p = 2$
EM	9.690 (0.003)	9.713 (0.003)
VB	9.701 (0.004)	9.727 (0.005)

model has been considered in Guarniero et al. (2017) using particle MCMC methods under the restriction that Σ_x is band-diagonal to reduce the number of parameters. It is also more general than that entertained in Naesseth et al. (2018) with Σ_x assumed diagonal, see also Chib et al. (2009) for a review on stochastic volatility models. We consider a fully Bayesian treatment as in Guarniero et al. (2017), applied to the same data set of 90 monthly returns (9/2008 to 2/2016) of 20 exchange rates with respect to the US dollar as reported by the Federal Reserve System. The specification of the prior and variational forms of the static parameters are explained in Appendix F. We consider proposals of the form

$$M_\phi(x_{n+1}|y_{n+1}, x_n) = \mathcal{N}(\mu + \text{diag}(a)(x_n - \mu), \Sigma^\phi),$$

where Σ^ϕ is diagonal and using $K = 50$ particles. Densities of the variational approximation that correspond to the GBP exchange rate can be found in Appendix F, Figure 4, which are largely similar to those obtained in (Guarniero et al., 2017). Furthermore, we approximate the one- and two-step predictive distributions

$$p(y_{m+p}|y_{0:m}) \approx \frac{1}{S} \sum_{s=1}^{S} \sum_{k=1}^{K} W_m^{k,s} \delta_{X_{m+p}^{k,s}} p_{\theta_s}(y_{m+p}|X_{m+p}^{k,s})$$

for $p \in \{1, 2\}$, where $\theta_1, ..., \theta_S \sim q_\psi(\theta)$, $\sum_{k=1}^{K} W_m^{k,s} \delta_{X_m^k}$ is the approximation of $p_{\theta_s}(x_m|y_{0:m})$ by the particle filter and $X_n^s \sim p_{\theta_s}(x_n^{k,s}|X_{n-1}^s, Y_{n-1}^{k,s})$ with $Y_n^s \sim p_{\theta_s}(y_n^{k,s}|X_n^{k,s})$ for $n = m+1, ..., m+p$ simulated from the generative model. The predictive distributions are evaluated using a log scoring rule (Gneiting and Raftery, 2007; Geweke and Amisano, 2010) to arrive at the predictive log-likelihoods in Table 1. The full variational approach attains higher predictive log-likelihoods.

6.3 Non-linear stochastic Hawkes processes

There has been an increasing interest in modelling asynchronous sequential data using point processes in various domains, including social networks (Linderman and Adams, 2014; Wang et al., 2017), finance (Bacry et al., 2015), and electronic health (Lian et al., 2015). Recent work (Du et al., 2016; Mei and Eisner, 2017; Xiao et al., 2017b,a) have advocated the use of neural networks in a black-box treatment of point process dynamics.

We illustrate that our approach allows scalable probabilistic inference for continuous-time event data $\{T_n, C_n\}_{n>0}$, $T_n < T_{n+1}$, where T_n is the time when the n-th event occurs and $C_n \in \{1, ..., D\}$ is an additional discrete mark associated with the event. We consider describing such a realisation as a D-variate point process with intensities $\lambda_t = h_\theta(\mu + \sum_{b=1}^B \Xi_t^b)$, driven by B continuous time processes

$$\Xi_t^b = \sum_{n \geq 1} \beta_b A_n^b \, \mathrm{e}^{-\beta_b(t-T_n)} 1_{[0,t)}(T_n), \quad t > 0,$$

and a non-negative monotone function h_θ. Moreover, $\mu, A_n \in \mathbb{R}^D$ and $\beta^b > 0$. Importantly, we allow A_n^b to depend on C_n, and the i-th component of A_n^b describes by how much the n-th event excites, if $(A_n^b)^i > 0$, or inhibits, if $(A_n^b)^i < 0$, subsequent events of type i. It is possible to view the dynamics as a discrete-time SSM; the essential idea being that Ξ^b is piecewise-deterministic between events, see Appendix G for details along with related work on Hawkes point processes (Hawkes, 1971a). Let us define the discrete-time latent process $X_{n+1} = (Z_n, A_n)$ with $Z_n = \Xi_{T_n}$, $A_n = \mathrm{vec}(A_n^1, ..., A_n^B)$. Standard theory about point processes, see Daley and Vere-Jones (2003), implies that the observation density is given by $g_\theta(t_n, c_n | z_{n-1}) = \lambda_{t_n}^{c_n} \exp\left(-\sum_{i=1}^D \int_{t_{n-1}}^{t_n} \lambda_s^i ds\right)$, where our model specification yields λ_s as a deterministic function between T_{n-1} and T_n given Z_{n-1}. Similar to Mei and Eisner (2017), we set $h_\theta(y) = \nu \, \mathrm{softplus}(y/\nu) = \nu \log(1 + \exp(y/\nu))$ as a scaled softplus function with ν a static parameter. Next, we specify the dynamics of A_n. We take the arguable most simple model, assuming $f_\theta(a_n | a_{n-1}, z_{n-1}, c_n) = \mathcal{N}(\sum_d \alpha_d \delta_{c_n d}, \sum_d \sigma_d^2 \delta_{c_n d})$ with $\alpha_1, ..., \alpha_D \in \mathbb{R}^{BD}$ and $\sigma_1^2, ..., \sigma_D^2$ positive diagonal matrices, while remarking in passing that our approach allows readily for extensions that could include temporal dynamics between successive intensity jumps or intensity jumps instantaneously correlated across different marks and time scales. Due to the piecewise deterministic decay of Ξ, note that $Z_n^b | Z_{n-1}^b, A_n^b = \mathrm{e}^{-\beta^b(T_n-T_{n-1})} Z_{n-1}^b + \beta^b A_n^b$, so the state transition of the process X is fully specified.
We apply our model to 20 days of high-frequency financial data for the BUND futures contract. The data is

Table 2: Prediction metric for different Hawkes process models on the test set of around 206k events. The stochastic Hawkes model is trained with 20 particles and uses $K \in \{20, 80\}$ particles during testing.

Method	Error rate next mark
Linear Hawkes	43.3 %
Non-linear Hawkes	40.9 %
Non-linear stochastic Hawkes ($K = 20$)	40.0%
Non-linear stochastic Hawkes ($K = 80$)	39.3%

available as part of the tick library (Bacry et al., 2017) with 4 event types: (i) mid-price up moves, (ii) mid-price down moves, (iii) buyer-initiated trades leaving the mid-price unchanged and (iv) seller-initiated trades not changing the mid. We train our model on 15 days and evaluate how well it predicts the type of the next event on out of sample data from the remaining 5 days. Table 2 reports better predictive performance of the proposed model in comparison with two benchmark models. First, a linear Hawkes process model estimated using maximum likelihood. Second, to illustrate that improved predictions might not be just explained due to inhibitory effects, we also compare against a non-linear Hawkes model. The latter can be seen, and has been implemented, as a limiting case of our generative model letting $\sigma_d^2 \to 0$, with inference thus performed using stochastic gradient descent of the negative log-likelihood. Predictions are Monte Carlo samples of the next event realisation from the generative model. Further details including assumptions on the variational distributions and the predictive performance using a smaller training set are given in Appendix H.

7 Conclusion

This paper has explored an inference approach that merges the scalability of variational methods with SMC sampling. We would like to emphasize that our approach is completely complementary to many recent advances in variational inference that can be used to parametrize $q_\psi(\theta)$. For instance, one can consider more expressive variational families (Rezende and Mohamed, 2015; Kingma et al., 2016; Salimans et al., 2015; Maaløe et al., 2016; Ranganath et al., 2016). Similarly, our Bayesian approach naturally allows us to incorporate prior knowledge. For instance, one could place sparsity-inducing priors and impose corresponding variational approximations (Ingraham and Marks, 2017; Ghosh and Doshi-Velez, 2017; Louizos et al., 2017). Applying such variational approximations to more expressive autoregressive models would be an interesting avenue to explore in future work.

References

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. (2015). Black box variational inference for state space models. *arXiv preprint arXiv:1511.07367*.

Bacry, E., Bompaire, M., Gaïffas, S., and Poulsen, S. (2017). tick: a python library for statistical learning, with a particular emphasis on time-dependent modeling. *arXiv preprint arXiv:1707.03003*.

Bacry, E., Jaisson, T., and Muzy, J.-F. (2016). Estimation of slowly decreasing hawkes kernels: application to high-frequency order book dynamics. *Quantitative Finance*, pages 1–23.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bowsher, C. G. et al. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912.

Brémaud, P. and Massoulié, L. (1996). Stability of non-linear hawkes processes. *The Annals of Probability*, pages 1563–1588.

Brémaud, P. and Massoulié, L. (2002). Power spectra of general shot noises and hawkes point processes with a random excitation. *Advances in Applied Probability*, 34(01):205–222.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Chib, S., Omori, Y., and Asai, M. (2009). Multivariate stochastic volatility. In *Handbook of Financial Time Series*, pages 365–400. Springer.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. (2015). A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988.

Cornebise, J., Moulines, É., and Olsson, J. (2008). Adaptive methods for sequential importance sampling with application to state space models. *Statistics and Computing*, 18(4):461–480.

Cremer, C., Morris, Q., and Duvenaud, D. (2017). Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*.

Daley, D. J. and Vere-Jones, D. (2003). An introduction to the theory of point processes volume i: Elementary theory and methods.

Dassios, A. and Zhao, H. (2011). A dynamic contagion process. *Advances in applied probability*, 43(03):814–846.

Davis, M. H. (1984). Piecewise-deterministic markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 353–388.

Del Moral, P. (1996). Non-linear filtering: interacting particle resolution. *Markov processes and related fields*, 2(4):555–581.

Dellaportas, P. and Pourahmadi, M. (2012). Cholesky-garch models with applications to finance. *Statistics and Computing*, 22(4):849–855.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208.

Doucet, A. and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., and Song, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564. ACM.

Duarte, A., Löcherbach, E., and Ost, G. (2016). Stability and perfect simulation of non-linear hawkes processes with erlang kernels. *arXiv preprint arXiv:1610.03300*.

Finke, A., Johansen, A. M., and Spanò, D. (2014). Static-parameter estimation in piecewise deterministic processes using particle gibbs samplers. *Annals of the Institute of Statistical Mathematics*, 66(3):577–609.

Fraccaro, M., Sonderby, S. K., Paquet, U., and Winther, O. (2016). Sequential neural models with stochastic

layers. In *Advances in Neural Information Processing Systems*, pages 2199–2207.

Geweke, J. and Amisano, G. (2010). Comparing and evaluating bayesian predictive distributions of asset returns. *International Journal of Forecasting*, 26(2):216–230.

Ghosh, S. and Doshi-Velez, F. (2017). Model selection in bayesian neural networks via horseshoe priors. *arXiv preprint arXiv:1705.10388*.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goyal, A., Sordoni, A., Côté, M.-A., Ke, N. R., and Bengio, Y. (2017). Z-forcing: Training stochastic recurrent networks. In *Advances in Neural Information Processing Systems*.

Gu, S., Ghahramani, Z., and Turner, R. E. (2015). Neural adaptive sequential monte carlo. In *Advances in Neural Information Processing Systems*, pages 2629–2637.

Guarniero, P., Johansen, A. M., and Lee, A. (2017). The iterated auxiliary particle filter. *Journal of the American Statistical Association*, pages 1–12.

Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 438–443.

Hawkes, A. G. (1971b). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., and Karhunen, J. (2010). Approximate riemannian conjugate gradient learning for fixed-form variational bayes. *Journal of Machine Learning Research*, 11(Nov):3235–3268.

Ingraham, J. and Marks, D. (2017). Variational inference for sparse and undirected models. In *International Conference on Machine Learning*, pages 1607–1616.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., Chopin, N., et al. (2015). On particle methods for parameter estimation in state-space models. *Statistical science*, 30(3):328–351.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pages 4743–4751.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Krishnan, R. G., Shalit, U., and Sontag, D. (2017). Structured inference networks for nonlinear state space models. In *AAAI*, pages 2101–2109.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Le, T. A., Igl, M., Jin, T., Rainforth, T., and Wood, F. (2018). Auto-encoding sequential monte carlo. In *ICLR*.

Lee, Y., Lim, K. W., and Ong, C. S. (2016). Hawkes processes with stochastic excitations. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, pages 79–88.

Lian, W., Henao, R., Rao, V., Lucas, J., and Carin, L. (2015). A multitask point process predictive model. In *Proceedings of the 32st International Conference on Machine Learning (ICML-15), JMLR Workshop and Conference Proceedings*.

Linderman, S. W. and Adams, R. P. (2014). Discovering latent network structure in point process data. In *ICML*, pages 1413–1421.

Linderman, S. W. and Adams, R. P. (2015). Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*.

Louizos, C., Ullrich, K., and Welling, M. (2017). Bayesian compression for deep learning. In *Advances in Neural Information Processing Systems*, pages 3290–3300.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. (2016). Auxiliary deep generative models. In *International Conference on Machine Learning*, pages 1445–1453.

Maddison, C. J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. (2017). Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pages 6576–6586.

Martens, J. (2014). New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*.

Martin, J. S., Jasra, A., and McCoy, E. (2013). Inference for a class of partially observed point process models. *Annals of the Institute of Statistical Mathematics*, 65(3):413–437.

Mei, H. and Eisner, J. M. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, pages 6757–6767.

Naesseth, C. A., Linderman, S. W., Ranganath, R., and Blei, D. M. (2018). Variational sequential monte carlo. In *Proceedings of the 21st International Conference on Artificial Intelligence (AISTATS)*.

Ogata, Y. (1981). On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. *arXiv preprint arXiv:1802.04537*.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *AISTATS*, pages 814–822.

Ranganath, R., Tran, D., and Blei, D. M. (2016). Hierarchical variational models. In *International Conference on Machine Learning*.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1530–1538.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1278–1286.

Ricci, J. (2014). *Applied Stochastic Control in High Frequency and Algorithmic Trading*. PhD thesis, University of Toronto.

Salimans, T., Kingma, D. P., Welling, M., et al. (2015). Markov chain monte carlo and variational inference: Bridging the gap. In *ICML*, volume 37, pages 1218–1226.

Süli, E. and Mayers, D. F. (2003). *An introduction to numerical analysis*. Cambridge university press.

Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979.

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. *Bayesian Time series models*, pages 115–138.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Wang, Y., Ye, X., Zhou, H., Zha, H., and Song, L. (2017). Linking micro event history to macro prediction in point process models. In *Artificial Intelligence and Statistics*, pages 1375–1384.

Whiteley, N., Johansen, A. M., and Godsill, S. (2011). Monte carlo filtering of piecewise deterministic processes. *Journal of Computational and Graphical Statistics*, 20(1):119–139.

Xiao, S., Farajtabar, M., Ye, X., Yan, J., Song, L., and Zha, H. (2017a). Wasserstein learning of deep generative point process models. In *Advances in Neural Information Processing Systems*, pages 3247–3257.

Xiao, S., Yan, J., Farajtabar, M., Song, L., Yang, X., and Zha, H. (2017b). Joint modeling of event sequence and time series with attentional twin recurrent neural networks. *arXiv preprint arXiv:1703.08524*.