Appendices

A SMC algorithm

Algorithm 1 Sampling from $q_\phi(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l|\theta)$ via an SMC sampler

Input: observations $y_{0:M}$, prior density p_θ, initial density $f_\theta(x_0)$, state transition density $f_\theta(x_{n+1}|x_n, y_n)$, observation density $g_\theta(y_n|x_n)$, proposal densities $M_n^\phi(x_n|y_n, x_{0:n-1})$ and resampling criteria.
Output: $(X_{0:M}^{1:K}, A_{0:M-1}^{1:K}, L) \sim q_\phi(\cdot|\theta)$.
for $k = 1...K$ **do**
  Sample $X_0^k \sim M_0^\phi(\cdot|y_0)$.
  Set $\alpha_0(X_0^k) = \frac{g_\theta(y_0|X_0^k) f_\theta(X_0^k|y_0)}{M_0^\phi(X_0^k)}$.
  Set $w_0(X_{0:n}^k) = \alpha_0(X_{0:n}^k)/K$.
  Set $W_0^k \propto w_0(X_0^k)$.
end for
for $n = 2...M$ **do**
  if resampling criteria satisfied **then**
    for $k = 1...K$ **do**
      Sample $A_{n-1}^k \sim r(\cdot|W_{n-1})$.
    end for
    Set $W_{n-1} = (\frac{1}{K}, ..., \frac{1}{K})$.
  else
    Set $A_{n-1} = (1, ..., K)$.
  end if
  for $k = 1...K$ **do**
    Sample $X_n^k \sim M_n^\phi(\cdot|y_n, X_{0:n-1}^{A_{n-1}^k})$.
    Set $X_{0:n}^k = (X_{0:n-1}^k, X_n^k)$.
    Set $\alpha_n(X_{0:n}^k) = \frac{g_\theta(y_n|X_n^k) f_\theta(X_n^k|X_{n-1}^{A_{n-1}^k}, y_{n-1})}{M_n^\phi(X_n^k|y_n, X_{0:n-1}^{A_{n-1}^k})}$.
    Set $w_n(X_{0:n}^k) = W_{n-1}^k \alpha_n(X_{0:n}^k)$.
    Set $W_n^k \propto w_n(X_{0:n}^k)$.
  end for
  Sample $L = l$ with probability W_M^l
end for

B Proof of Proposition 2

Consider an SMC algorithm with K particles targeting

$$\pi_\theta(x_{0:M}) \coloneqq \gamma(\theta, x_{0:M})/\gamma_M(\theta),$$

where $\gamma(\theta, x_{0:M}) = p(\theta, x_{0:M}, y_{0:M})$ is related to the posterior via $\pi(\theta, x_{0:M}) = \gamma(\theta, x_{0:M})/Z_M$. Z_M is a normalising constant independent of θ that represents the marginal likelihood $Z_M = p(y_{0:M})$. Furthermore, $\gamma_M(\theta) = \int \gamma(\theta, x_{0:M}) dx_{0:M} = p(\theta)p_\theta(y_{0:M})$. We denote the likelihood estimator of this SMC algorithm as $\tilde{Z}_M^{\theta,\phi}$. Following analogous arguments as in Andrieu et al. (2010), we have from the definition of the importance weights

$$\begin{aligned}
&\frac{\tilde{\pi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l)}{q_{\phi,\psi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l)} \\
=&\frac{\pi(\theta, x_{0:M}^l) K^{-(M+1)}}{q_\psi W_M^l M_0^\phi(x_0^{b_0^l}|y_0) \prod_{n=1}^M W_{n-1}^{b_{n-1}^l} M_n^\phi(x_n^{b_n^l}|y_n, x_{0:n-1}^{b_{n-1}^l})} \\
=&\frac{\pi(\theta, x_{0:M}^l) K^{-(M+1)}}{q_\psi(\theta) M_0^\phi(x_0^{b_0^l}|y_0) \prod_{n=1}^M M_n^\phi(x_n^{b_n^l}|y_n, x_{0:n-1}^{b_{n-1}^l})} \\
&\cdot \frac{\prod_{n=0}^M \left(\sum_{k=1}^K w_k(x_{0:M}^k)\right)}{\prod_{n=0}^M w_n(X_{0:M}^{b_n^l})} \\
=&\frac{\pi(\theta, x_{0:M}^l) \tilde{Z}_M^{\theta,\phi}}{q_\psi(\theta) \gamma(\theta, x_{0:M}^l)} \\
=&\frac{\tilde{Z}_M^{\theta,\phi}}{q_\psi(\theta) p(y_{0:M})}.
\end{aligned}$$

Note that $\tilde{Z}^{\theta,\phi} = p(\theta)\hat{Z}^{\theta,\phi}$, where $\hat{Z}^{\phi,\theta}$ is the SMC likelihood estimator in the main paper targeting a density proportional to $p_\theta(x_{0:M}, y_{0:M})$, whilst $\tilde{Z}^{\theta,\phi}$ targets a density proportional to $p(\theta)p_\theta(x_{0:M}, y_{0:M})$. Consequently,

$$\begin{aligned}
\mathrm{KL}(q_{\psi,\phi}||\tilde{\pi}) &= -\mathbb{E}_{q_{\psi,\phi}}\left[\log \frac{\tilde{Z}_M^{\theta,\phi}}{q_\psi(\theta)}\right] + \log p(y_{0:M}) \\
&= -\mathcal{L}(\psi, \phi) + \log p(y_{0:M}),
\end{aligned}$$

which concludes the proof.

C Proof of Corollary 3

Observe that we can write

$$\begin{aligned}
&\mathrm{KL}\left(q_{\psi,\phi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l)||\tilde{\pi}(\theta, x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l)\right) \\
&= \mathbb{E}_{q_{\psi,\phi}(\theta, x_{0:M}^l, b_{0:M}^l)}\Bigg[\mathbb{E}_{q_\phi(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l})|\theta, x_{0:M}^l, b_{0:M}^l)}\Bigg[\\
&\qquad \log q_{\psi,\phi}(\theta, x_{0:M}^l, b_{0:M}^l) \\
&\qquad + \log q_\phi(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l, b_{0:M}^l)\Bigg] \\
&\qquad - \log \tilde{\pi}(\theta, x_{0:M}^l, b_{0:M}^l) \\
&\qquad - \log \tilde{\pi}_{\mathrm{CSMC}}(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l, b_{0:M}^l)\Bigg] \\
&= \mathrm{KL}(q_{\psi,\phi}(\theta, x_{0:M}^l)||\pi(\theta, x_{0:M}^l)) \\
&\quad + \mathbb{E}_{q_{\psi,\phi}(\theta, x_{0:M}^l, b_{0:M}^l)}\Bigg[\\
&\qquad \mathrm{KL}(q_\phi(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l})|\theta, x_{0:M}^l, b_{0:M}^l)\Big|\Big| \\
&\qquad \tilde{\pi}_{\mathrm{CSMC}}(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l, b_{0:M}^l))\Bigg].
\end{aligned}$$

D Proof of Proposition 4

We can write the extended target distribution as

$$\begin{aligned}&\tilde{\pi}(x_{0:M}^{1:K}, a_{0:M-1}^{1:K}, l)\\ =&\frac{\pi(\theta, x_{0:M}^l)}{K^{M+1}}\tilde{\pi}_{\text{CSMC}}(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l, b_{0:M}^l).\end{aligned}$$

This follows from the fact that $x_{0:M}^l = (x_0^{b_0^l}, ..., x_M^{b_M^l})$ and that $b_{0:M}|x_{0:M}^l, \theta$ is uniformly distributed on $\{1, ..., K\}^{M+1}$. Hence, $\frac{\pi(\theta, x_{0:M}^l)}{K^{-(M+1)}}$ is the marginal density $\tilde{\pi}(\theta, x_{0:M}^l, b_{0:M}^l)$. Moreover, the variational approximation of the static parameter θ and latent states $x_{0:M}^l$, obtained as the marginal of the extended variational distribution, is given by, following similar arguments as in Naesseth et al. (2018),

$$\begin{aligned}&q_{\psi,\phi}(\theta, x_{0:M}^l) = \frac{q_{\psi,\phi}(\theta, x_{0:M}^l, b_{0:M}^l)}{q_{\psi,\phi}(b_{0:M}^l|\theta, x_{0:M}^l)}\\ &= \frac{1}{K^{-(M+1)}}\int q_{\psi,\phi}(\theta, x_{0:M}^l, a_{0:M-1}^l, x_{0:M}^{\neg b^l}, a_{0:M-1}^{\neg b^l})\\ &\qquad d(x_{0:M}^{\neg b^l}, a_{0:M-1}^{\neg b^l})\\ &= K^{M+1}\int q_\psi(\theta)\frac{w_M^l(x_{0:M}^{b^l})}{\sum_{l'} w_M^{l'}(x_{0:M}^{l'})}\prod_{k=1}^K M_0^\phi(x_0^k|y_0)\\ &\qquad\cdot\prod_{n=1}^M\frac{w_{n-1}^k(x_{0:n}^{b_{n-1}^k})}{\sum_{l'} w_{n-1}^{l'}(x_{0:n-1}^{b_{n-1}^{l'}})}M_n^\phi(x_n^k|y_n, x_{0:n-1}^{b_{n-1}^{a_{n-1}^k}})\\ &\qquad d(x_{0:M}^{\neg b^l}, a_{0:M-1}^{\neg b^l})\\ &= \int q_\psi(\theta)\left(\prod_{n=1}^M\frac{\gamma_\theta(x_{0:n}^l)}{\gamma_\theta(x_{0:n-1}^l)\sum_{l'} w_n^{l'}((x_{0:n}^{l'}))}\right)\\ &\quad\cdot\prod_{k:k\neq b_0^l} M_0^\phi(x_0^k|y_0)\\ &\quad\cdot\prod_{n=1}^M\prod_{k:k\neq b_n^l} W_{n-1}^k M_n^\phi(x_n^k|y_n, x_{n-1}^{a_{n-1}^k})d(x_{0:M}^{\neg b^l}, a_{0:M-1}^{\neg b^l})\\ &= q_\psi(\theta)\gamma_\theta(x_{0:M}^l)\\ &\quad\cdot\mathbb{E}_{\tilde{\pi}_{\text{CSMC}}(x_{0:M}^{\neg b_{0:M}^l}, a_{0:M-1}^{\neg b_{0:M-1}^l}|\theta, x_{0:M}^l)}\left[\left(\hat{Z}_M^{\theta,\phi}\right)^{-1}\right]\end{aligned}$$

E Natural gradients

We have also experimented with optimizing the variational distribution over the static parameters using natural gradients (Amari, 1998; Martens, 2014) to take into account the Riemannian geometry of the approximating distributions, as explored previously for variational approximations, see for instance Honkela et al. (2010); Hoffman et al. (2013). Recall that we are optimizing over the space of probability distributions $q_\psi(\cdot)$ with parameter ψ, for which we can consider a possible metric given by the Fisher information

$$\begin{aligned}I(\psi) &= \mathbb{E}_{q_\psi(\theta)}\left[\nabla_\psi \log q_\psi(\theta)\left(\nabla_\psi \log q_\psi(\theta)\right)^T\right]\\ &= -\mathbb{E}_{q_\psi(\theta)}\left[H_{\log q_\psi}(\theta)\right],\end{aligned}$$

The last equation assumes that q_ψ is twice differentiable and $H_{\log q_\psi}(\theta) = \left(\frac{\partial^2 \log q_\psi(\theta)}{\partial\psi_i\partial\psi_j}\right)_{ij}$ denotes the Hessian. This induces an inner product $\langle\psi_1, \psi_2\rangle_{\psi_0} = \psi_1^T F(\psi_0)\psi_2$ locally around ψ_0, hence gives rise to a norm $||\cdot||_{\psi_0}$. The Fisher information matrix is connected to the KL divergence, since the distance in the induced metric is given approximately by the square root of twice the KL-divergence:

$$\begin{aligned}&\text{KL}(q_{\psi_1}||q_{\psi_2})\\ =&\frac{1}{2}(\psi_2-\psi_1)I(\psi_1)(\psi_2-\psi_1)^T + O((\psi_2-\psi_1)^3),\end{aligned}$$

This follows from a second order Taylor expansion and from using the fact that $\mathbb{E}_{q_\psi}\left[\nabla_\psi \log q_\psi\right] = 0$. Recall that the natural gradient of a function $\mathcal{L}(\psi)$ is defined by

$$\tilde{\nabla}_\psi\mathcal{L}(\psi) = I(\psi)^{-1}\nabla_\psi\mathcal{L}(\psi)$$

and one can show that under mild assumptions (Martens, 2014),

$$\begin{aligned}&\sqrt{2}\frac{\tilde{\nabla}_\psi\mathcal{L}(\psi)}{||\tilde{\nabla}_\psi\mathcal{L}(\psi)||_\psi}\\ =&\lim_{\epsilon\to 0}\frac{1}{\epsilon}\text{argmax}_{d:\text{KL}(q_{\psi+d}||q_\psi)\leq\epsilon^2}\mathcal{L}(\psi+d).\end{aligned}$$

Thus the natural gradient is the steepest ascent direction with the distance measured by the KL-divergence. The natural gradient ascent does not depend on the parametrisation of q_ψ as a consequence of the invariance of the KL-divergence with respect to reparametrisations.

For mean-field approximations, computing the inverse of the Fisher information matrix simplifies, as the Fisher information has a block-diagonal structure in this case. We consider both normal and log-normal factors. For a univariate Gaussian distribution $q_{\mu,v}$ with mean μ and variance $\exp(v)^2$ parametrized by the logarithm of the standard deviation v, we obtain $\nabla_{\mu,v}\log q_{\mu,v}(\theta) = (\text{e}^{-2v}(\theta-\mu), \text{e}^{-2v}(\theta-\mu)^2-1)^T$. Consequently,

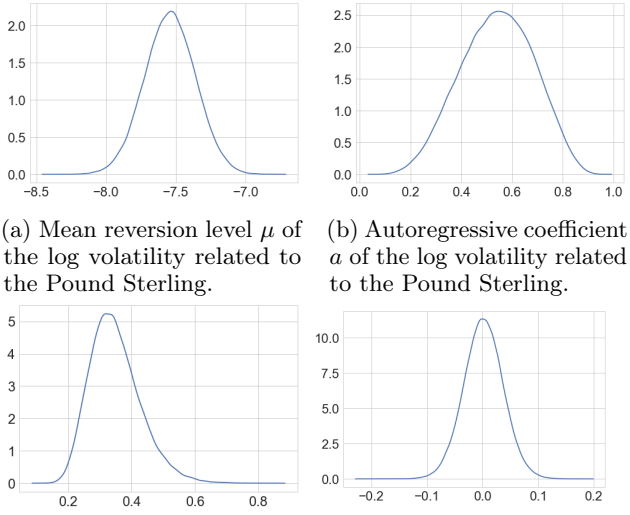
$$I(\mu, v) = \begin{pmatrix}\text{e}^{-2v} & 0\\ 0 & 2\end{pmatrix}.$$

For a log-normal distribution $q_{a,b}(\theta)$, parametrized so that $\log\theta \sim \mathcal{N}(a, \exp(b)^2)$, we have $\nabla_{a,b}\log q_{a,b}(\theta) = (\text{e}^{-2b}(\log(\theta)-a), \text{e}^{-2b}(\log(\theta)-a)^2-1)^T$ and we arrive at the same form for the Fisher information

$$I(a, b) = \begin{pmatrix}\text{e}^{-2b} & 0\\ 0 & 2\end{pmatrix}.$$

F Priors and variational approximations for the stochastic volatility model

Compared to Guarniero et al. (2017), we choose a different structure of Σ_x to guarantee its positive-definiteness, along with slightly different priors. We model Σ_x with its unique Cholesky factorisation (Dellaportas and Pourahmadi, 2012), i.e. $\Sigma_x = LL^T$ with L a lower triangular matrix having positive values on its diagonal. We set Σ_x^0 as the stationary covariance of the latent state. Independent priors are placed for $a_i \sim U(0,1)$ and $\mu_i \sim \mathcal{N}(0,10)$ as well as $L_{ij} \sim \mathcal{N}(0,10)$, for $i < j$ and $\log L_{ii} \sim \mathcal{N}(0,10)$. We assume a mean-field variational approximation with normal factors for μ and for the entries of L below the diagonal and log-normal factors for its diagonal. Furthermore, a_i is assumed to be the sigmoid transform sigm: $x \mapsto 1/(1+\mathrm{e}^{-x})$ of normally distributed variational factors. We initialized the mean of L with a diagonal matrix having entries 0.2 and the mean of μ_i with the logarithm of the standard deviation of the ith component of the time series. Densities of the variational approximation for parameters corresponding to the GBP exchange rate are given in Figure 4.



(a) Mean reversion level μ of the log volatility related to the Pound Sterling.

(b) Autoregressive coefficient a of the log volatility related to the Pound Sterling.

(c) Variance part of Σ_x for the error term of the log volatility related to the Pound Sterling.

(d) Covariance part of Σ_x for the error term of the log volatilities related to the Pound Sterling and Euro.

Figure 4: Density estimates for the parameters related to the Pound Sterling in the multivariate stochastic volatility model.

G Hawkes point processes and state space models

In contrast to linear Hawkes processes (Hawkes, 1971a,b), we also allow for negative excitations, as explored previously for instance in Brémaud and Massoulié (1996); Bowsher et al. (2007); Duarte et al. (2016). The values of A^b and β^b are commonly assumed to be fixed through time, while time-varying μ have been considered in various settings. Stochastic time-varying excitations have been analysed in a probabilistic setting in Brémaud and Massoulié (2002); Dassios and Zhao (2011). Moreover, Ricci (2014) considered frequentist inference of the excitation model parameters from a matrix-valued categorical distribution, while Lee et al. (2016) performed MCMC with excitations evolving according to an Ito process in the one-dimensional case. However, scalable Bayesian inference for non-linear stochastic Hawkes processes has been missing, with previous variational inference schemes (Linderman and Adams, 2015) having been restricted to linear Hawkes processes due to their resilience on the branching structure of linear Hawkes processes. SMC methods for shot-noise Cox processes has been considered in Whiteley et al. (2011); Martin et al. (2013) for on-line filtering and Finke et al. (2014) for static-parameter inference. While we expect such methods to scale poorly to models with many parameters and observations, we borrow their idea of describing the dynamics of the point process using piecewise-deterministic processes (Davis, 1984), which enables us to employ the proposed inference approach for discrete-time state space models.
More concretely, since Ξ_t^b follows deterministic dynamics between two events, we can write $\Xi_t^b = F_b(t, T_n, \Xi_{T_n}^b)$ for $t \in [T_n, T_{n+1})$ with the deterministic function $F_b(t,s,z^b) = \mathrm{e}^{-\beta_b(t-s)}z^b$. Whenever an event of type C_n occurs at time T_n, the process Ξ^b jumps with size $\Delta\Xi_{T_n}^b = \beta_b A_n^b$. The process $Z_n^b = \Xi_{T_n}^b$, $n > 0$, satisfies $\Xi_t^b = F_b(t, T_n, Z_n^b)$ for $t \in [T_n, T_{n+1})$. Note that we scale each A_n^b with the diagonal matrix β_b. This ensures that the triggering kernel functions $s \mapsto \beta^b \mathrm{e}^{-\beta^b s}$ have L_0 norm of one for any b.

H Inference and predictions details for Hawkes process models

We place the following priors for the dynamics of A: For any $d \in \{1,...,D\}$, $\alpha_d \sim \otimes_{i=1}^{DB}\mathcal{N}(0,10)$ and consider mean-field variational approximations having the same forms. Furthermore, a priori, suppose that $\mu \sim \otimes_{i=1}^{D}\mathrm{Ga}(0.01,0.01)$, $\mathrm{diag}(\sigma_d^2) \sim \otimes_{i=1}^{DB}\mathrm{Ga}(0.01,0.01)$ and $\beta_b - \beta_{b-1} \sim \mathcal{LN}(0,1)$, $b \in \{1,...,B\}, \beta_0 = 0$, all with a log-normal variational approximation. Eventually, for the softmax scale parameter, a priori $\nu \sim U(0,1)$ with a variational approximation as the sigmoid

transform of a normal factor. The proposal function used is

$$\begin{aligned} &M_\phi(a_n, z_n | a_{n-1}, z_{n-1}, t_{n+1}, c_{n+1}, t_n, c_n) \\ =&h_\phi(a_n|c_n) f_\theta(z_n|z_{n-1}, a_{n-1}, t_n, c_n), \end{aligned} \tag{10}$$

with $h_\phi(a_n|c_n) = \mathcal{N}(\sum_d \tilde{\alpha}_d \delta_{c_n d}, \sum_d \tilde{\sigma}_d^2 \delta_{c_n d})$, $\tilde{\alpha_d} \in \mathbb{R}^{BD}$, $\tilde{\sigma_d}$ positive diagonal matrices and wheref_θ describes the determinsitic decay of Z_n according to the prior transition density.

Let us also mention that the observation density contains a one-dimenisonal intractable integral. We apply Gaussian quadrature to evaluate the integral after transforming the quadrature points to better cover the interval immediately after an event where the intensity function is varying more quickly, see Appendix I for details. We initialised the variational parameters so that the variational distribution of α is largely concentrated around the maximum likelihood estimates in a linear Hawkes model and the variational distribtuion of ν concentrated around 0. The values of β_b are commonly fixed in a maximum likelihood estimation setting to guarantee concavity of the log-likelihood. We have chosen $B = 5$ with $(\log \beta_1, \log(\beta_2 - \beta_1), ..., \log(\beta_5 - \beta_4)) = (-1, 1, 3, 5, 7)$ fixed. This allows event interactions across various time scales, ranging from $\beta_1 \approx 0.36$ to $\beta_5 \approx 1268$.

We have also split the events in subsamples of length $M = 100$ each and used the particles from the previous event-batch as the initial particles for the subsequent event-batch. We used $K = 20$ particles and performed optimisation with Adam (Kingma and Ba, 2014) and step size 0.0001. Similar performance was observed either using standard or natural gradients for the considered hyperparameters and reported results correspond to optimsiaton with standard gradients only.

Regarding inference for the benchmark models, maximum likelihood estimation for the linear Hawkes model was performed using the tick library (Bacry et al., 2017), with the fixed time scales $\beta_1, ..., \beta_5$ given above. Parameters for the non-linear Hawkes model were estimated using a limiting case of the generative model with very small σ_d, $K = 1$, and proposing the single particle according to the generative model, hence particularly with small variances σ_d. Concretly, we consider

$$\begin{aligned} &f_\theta(a_n|a_{n-1}, z_{n-1}, c_n) \\ =&h_\phi(a_n|c_n) = \mathcal{N}\left(\sum_d \alpha_d \delta_{c_n d}, \sum_d \sigma_d \delta_{c_n d}\right), \end{aligned}$$

recalling h_ϕ from the definition (10) of the proposal function and where for all $d \in \{1, ..., D\}$,

$$\sigma_d = \epsilon \begin{pmatrix} \beta_1^{-1} & & & & & & \\ & \ddots & & & & & \\ & & \beta_1^{-1} & & & & \\ & & & \ddots & & & \\ & & & & \beta_B^{-1} & & \\ & & & & & \ddots & \\ & & & & & & \beta_B^{-1} \end{pmatrix},$$

$\epsilon = 0.0001$. Stochastic gradient descent then yields point estimates over $\alpha_1, ..., \alpha_D$, decay parameters $\beta_1, ..., \beta_B$, softmax scale parameter ν and the background intensity parameter μ. Initial parameters have similary been set to the maximum likelihood estimates from the linear Hawkes model. We used Adam (Kingma and Ba, 2014) with step sizes 0.0001 and 0.0005, with the reported result corresponding to the best performing step size for the considered metric in Table 2.

For the prediction of the next mark c_{m+1} given the observations $t_{1:m}, c_{1:m}$, we can sample $\theta_1, ..., \theta_S \sim q_\psi(\theta)$ and run a particle filter that yields

$$\sum_{k=1}^{K} W_m^{k,s} \delta_{(Z_{0:m-1}^{k,s}, A_{0:m-1}^{k,s})}(z_{0:m-1}^s, a_{0:m-1}^s)$$

as an approximation of $p_{\theta_s}(z_{0:m-1}^s, \alpha_{0:m-1}^s | t_{1:m}, c_{1:m})$. Set

$$\hat{Z}_m^{b,k,s} = \mathrm{e}^{-\beta_b (t_m - t_{m-1})} Z_{m-1}^{b,k,s} + A_m^{b,k,s},$$

with $A_m^{k,s} \sim f_{\theta_s}(\cdot|c_m)$ sampled from the prior transition density. We then sample 10 realisations

$$t_{m+1}^{k,s,j}, c_{m+1}^{k,s,j} \sim g_{\theta^s}(t_{m+1}, c_{m+1} | \hat{Z}_m^{k,s}), \quad j = 1, ..., 10,$$

using the standard thinning algorithm for point processes, see for instance Ogata (1981); Daley and Vere-Jones (2003); Bowsher et al. (2007). In the stochastic Hawkes process model, we have chosen $S = 4$ and $K = 20$. To account for a similar computational budget for the benchmark models, we sample $10 \cdot 4 \cdot 20$ event realisations in these cases instead. For predicting the next mark c_{m+1}, we use the sampled mark that occurred most often within $\{c_{m+1}^{k,s,j}\}_{k,s,j}$, where the count associated with $c_{m+1}^{k,s,j}$ is weighted by $W_m^{k,s}$. Notice that we do not condition on the observed t_{m+1} for predicting c_{m+1} and the dependence of $c_{m+1}^{k,s,j}$ on $t_{m+1}^{k,s,j}$ is accounted for via the thinning procedure. In the stochastic Hawkes process model, we have also run predictions using $K = 80$ particles, using the same model trained with $K = 20$ particles.

In order to show how the different models generalize if less data is available, we have trained the different

models on either the first 100 or 1000 events of one day and evaluated how well the model performs on predicting the first 10000 events on another day. We have repeated this procedure for 10 days and found that a fully Bayesian treatment is beneficial when trained on 100 events. The fully variational approach has an error rate of 65%, whilst the same stochastic Hawkes process model using a point estimate of the static parameters has an error rate of 70%. The two approaches yield similar results when trained on 1000 events with an error rate of below 50%, whereas a benchmark non-linear Hawkes model without latent intensity dynamics has an error rate of 65%. Although a fully Bayesian treatment might not be necessary if one imposes a parsimonious model for the evolution of the latent intensity, we hope that this example encourages further point process models that allow for online Bayesian updating as we feel that intensity excitations with latent dynamics have been underexplored for Hawkes process models.

I Gaussian quadrature of the intensity function

We approximate the integral of the intensity function with Gaussian quadrature, see for instance Süli and Mayers (2003) for details. Let $p_1, ..., p_n$ be orthogonal polynomials in $L^2[a, b]$ equipped with the scalar product $\langle f, g\rangle = \int_a^b f(t)g(t)dt$, $f, g \in L^2[a, b]$ with p_k having degree k. Note that p_k can be constructed recursively by Gram-Schmidt-orthogonalization. Furthermore, let $t_1, ..., t_n$ be the roots of p_n and consider the Lagrange polynomials for $i = 1, ..., n$,

$$L_i(t) = \prod_{j=1, j\neq i}^{n} \frac{t - t_j}{t_i - t_j},$$

which satisfy $L_i(t_k) = \delta_{ik}, k = 1, ..., n$. Define

$$w_i = \int_a^b L_i(t)dt$$

as well as the Gaussian quadrature

$$I_n(f) = \sum_{i=1}^{n} w_i f(t_i).$$

Then $I_n(p) = \int_a^b p(t)dt$ for polynomials p of degree up to $2n-1$. We are interested in evaluating $\int_{T_{min}}^{T_{max}} \lambda^i(t)dt$ for fixed T_{min} and T_{max}. Here, T_{max} is the time of the next event and we have fixed T_{min} to the previous event plus one microsecond. The lowest resolution of the event timestamps for the considered dataset is one microsecond. Assume there is a function g such that $\lambda(t) = g(\mathrm{e}^t)$. We can write

$$\int_{T_{min}}^{T_{max}} \lambda(t)dt = \int_{\log T_{min}}^{\log T_{max}} g(\mathrm{e}^{\tilde{t}})\, \mathrm{e}^{\tilde{t}}\, d\tilde{t}.$$

This motivates the following change of variables that has also been considered in Bacry et al. (2016) for solving an integral equation involving the kernel function of a Hawkes process. Suppose that $t_1...t_n$ are the quadrature point with weights $w_1, ...w_n$ on $[\log T_{min}, \log T_{max}]$. The transformed quadrature scheme is then

$$(\tilde{t}_n, \tilde{w}_n) = (\mathrm{e}^{t_n}, w_n\, \mathrm{e}^{t_n}).$$

We used 50 quadrature points in our experiments.