Bayesian Bivariate Hawkes

Kar Wai Lim
Data61/CSIRO
karwai.lim@anu.edu.au

Young Lee
Data61/CSIRO
young.lee@data61.csiro.au

Cheng Soon Ong
Data61/CSIRO
chengsoon.ong@anu.edu.au

ABSTRACT

This article discusses the fully Bayesian framework for a bivariate time series model that explains the clustered arrival of events. The theoretical properties are worked out and its Bayesian inference procedure is carried through. These extends some results not currently in the literature.

Keywords

Bivariate Hawkes Processes; Branching Representation; Fully Bayesian Inference; Adaptive Rejection Sampling

1. INTRODUCTION

Time series modelling have been the basis for any study of a behaviour or process over a period of time. In this paper, we address the modelling of asynchronous time series using point processes. In particular, we consider a bivariate time series process for which the components excite each other, *i.e.*, the occurrence of an event of the first process tends to trigger an event for the second process, and *vice versa*. We show that this can be an effective tool to capture the evolving data that shows mutual excitations.

Contributions.

We formulate a Bayesian bivariate Hawkes process and utilise a branching structure representation of general point processes to perform inference using Markov chain Monte Carlo (MCMC) methods. This extends the one-dimensional case of Rasmussen (2013) and provide an alternative method to standard parameter inference such as the maximum likelihood estimation (Ozaki, 1979).

2. BIVARIATE HAWKES PROCESSES

A Hawkes process may be *completely* characterised by its underlying intensity function, or it may be formulated as a Poisson cluster process as illustrated by Hawkes and Oakes (1974) and Møller and Rasmussen (2005, 2006). Here, we start with the intensity formulation (for details, see Daley and Vere-Jones, 2003).

Consider the 2-dimensional Hawkes processes with dissimilar exponential decaying intensities, which we call the bivariate Hawkes process, its intensity functions (we assume time t starts at 0) can be written as

$$\lambda_1(t) = \mu_1 + Y_1^1(0)\, e^{-\delta_1^1 t} + Y_1^2(0)\, e^{-\delta_1^2 t} + \sum_{j=1:\, t\geq t_j^1}^{N^1(t)} Y_{1,j}^1\, e^{-\delta_1^1 t} + \sum_{j=1:\, t\geq t_j^2}^{N^2(t)} Y_{1,j}^2\, e^{-\delta_1^2 t}, \quad (1)$$

$$\lambda_2(t) = \mu_2 + Y_2^1(0)\, e^{-\delta_2^1 t} + Y_2^2(0)\, e^{-\delta_2^2 t} + \sum_{j=1:\, t\geq t_j^1}^{N^1(t)} Y_{2,j}^1\, e^{-\delta_2^1 t} + \sum_{j=1:\, t\geq t_j^2}^{N^2(t)} Y_{2,j}^2\, e^{-\delta_2^2 t}, \quad (2)$$

where $\lambda_1(t)$ and $\lambda_2(t)$ are the intensity functions for process 1 and 2, respectively.

Here, $\mu_1 > 0$ and $\mu_2 > 0$ are the *background intensity* for process 1 and 2, that is, the constant rate for which the events in each process are generated. We note that the parameters $Y_1^1(0)$, $Y_1^2(0)$, $Y_2^1(0)$, and $Y_2^2(0)$ capture the *edge effect* (see Møller and Rasmussen, 2005) associated with unseen events before $t = 0$. Put simply, these four variables account for all the added intensity generated by previously unseen events in $t < 0$. To illustrate, $Y_1^2(0)$ corresponds to the added intensity generated by all the previous events from process 2, realised at the start of our observation ($t = 0$).

The amount associated within the summation in Equations (1) and (2) is the *additional* intensity corresponding to either self-excitation (for $Y_{1,j}^1$ and $Y_{2,j}^2$) or external-excitation (for $Y_{1,j}^2$ and $Y_{2,j}^1$). These added intensities are attributed to the events associated to each process. The added intensities are assumed to follow an exponential decay over time. We note that the decay rates δ_1^1, δ_1^2, δ_2^1, and δ_2^2 are not assumed to be the *same* unlike in most models, for example, the decay rates in Dassios and Zhao (2013, Section 5) are constant within each process. Finally, $N^1(t)$ and $N^2(t)$ represent the counting processes of the two-dimensional Hawkes. For instance, $N^1(t)$ is the number of events attributed to process 1 observed at and before time t.

In the remaining of this paper, we will assume that the levels of excitation Y follows i.i.d. gamma distributions, though noting that the distributions can easily be modified or extended. One benefit of choosing the gamma distribution is that their marginal posteriors of Y will then be log-concave, allowing sampling *via* the ARS. The other parameters (*e.g.*,

TSAA '16, December 06 2016, Hobart, TAS, Australia

ISBN 978-1-4503-4820-1/16/12...$15.00

DOI: http://dx.doi.org/10.1145/3014340.3014343

δ, α, β) will be assigned prior distributions outlined in Section 3.1.

We note that these Hawkes processes reduce to the ordinary Hawkes processes with exponential decay when the levels of excitation Y are constant, for instance, see the model of Muni Toke and Pomponio (2012). It is also interesting to note that the multivariate Hawkes processes can be formulated as a shot noise Cox process (Cox and Isham, 1980) by removing the self-excitation bit, easily achieved by setting certain Y to be zero. For a more detailed reviews on the Hawkes processes, we refer the readers to the recent review papers by Bacry et al. (2015). and Laub et al. (2015).

The Branching Structure Interpretation.

An alternative but similar view of the Hawkes process refers to the Poisson cluster process interpretation. This line of interpretation was initiated by Hawkes and Oakes (1974). The event times of a Hawkes process can be separated into two categories, an event time is either an *immigrant* or an *offspring*. The offspring event times are generated by existing event times. The immigrant event times do not have existing parent event points and the numbers of which follow a homogeneous poisson process with the base intensity. Affiliated with each immigrant event is a cluster of offspring event times. For more details, we refer to Daley and Vere-Jones (2003).

3. MODEL LIKELIHOOD

Hawkes processes can also be represented by a Poisson cluster process, as outlined in previous section. For our inference algorithm to learn the parameters, we employ the following 'branching representation' which leads to a fully Gibbs sampler. This representation corresponds to introducing additional random variables called the auxiliary variables (see data augmentation, van Dyk and Meng, 2001). We note that in contrast to the representation described in Veen and Schoenberg (2008); Rasmussen (2013); and Lee et al. (2016), we depict the branching representation for a bivariate Hawkes process.

In this representation, we say an event time t_j^1 (for $j = 1, \ldots, N^1(T)$) is an immigrant if it is generated from the background intensity μ_1, otherwise, we say t_j^1 is an offspring. If t_j^1 is an offspring, it can be from the edge effect, or an offspring of another observed event time (either self excited or externally excited).

For each t_j^1, we introduce an indicator matrix A_j^1 that tells us the source of t_j^1. Collectively, all the A_j^1 tell us the branching structure of the Hawkes processes. We note that each A_j^1 is a special indicator matrix where only one of its element is 1. The particular entry of 1 tells us the type of the event time t_j^1:

1. If t_j^1 is an immigrant, $(A_j^1)_{00} = 1$.
2. If t_j^1 is an offspring due to edge effect from process 1, then $(A_j^1)_{10} = 1$.
3. If t_j^1 is an offspring due to edge effect from process 2, then $(A_j^1)_{20} = 1$.
4. If t_j^1 is a self-excited offspring of another event in process 1, t_k^1, satisfying $t_k^1 < t_j^1$, then $(A_j^1)_{1k} = 1$.
5. If t_j^1 is an externally-excited offspring of another event in process 2, t_k^2, satisfying $t_k^2 < t_j^1$, then $(A_j^1)_{2k} = 1$.

Similarly for t_j^2 (for $j = 1, \ldots, N^2(T)$).

3.1 Priors

We denote $\mathbf{\Theta} = \{\mu, \delta, \alpha, \beta\}$ as the set of all parameters of interest. These parameters are assigned gamma prior either due to conjugacy (such that their posterior is also gamma), or simply for convenience. We note that this choice also allows the non-conjugate variables to exhibit log-concavity, allowing our MCMC algorithm to be fully Gibbs sampler (details in Section 4). The priors are outlined as follows.

$$\mu_1 \sim \text{Gamma}\big(\tau_{\mu_1}, \psi_{\mu_1}\big) \qquad \mu_2 \sim \text{Gamma}\big(\tau_{\mu_2}, \psi_{\mu_2}\big) \tag{3}$$

$$\delta_1^1 \sim \text{Gamma}\big(\tau_{\delta_1^1}, \psi_{\delta_1^1}\big) \qquad \delta_1^2 \sim \text{Gamma}\big(\tau_{\delta_1^2}, \psi_{\delta_1^2}\big) \tag{4}$$

$$\delta_2^1 \sim \text{Gamma}\big(\tau_{\delta_2^1}, \psi_{\delta_2^1}\big) \qquad \delta_2^2 \sim \text{Gamma}\big(\tau_{\delta_2^2}, \psi_{\delta_2^2}\big) \tag{5}$$

$$\alpha_1^1 \sim \text{Gamma}\big(\tau_{\alpha_1^1}, \psi_{\alpha_1^1}\big) \qquad \alpha_1^2 \sim \text{Gamma}\big(\tau_{\alpha_1^2}, \psi_{\alpha_1^2}\big) \tag{6}$$

$$\alpha_2^1 \sim \text{Gamma}\big(\tau_{\alpha_2^1}, \psi_{\alpha_2^1}\big) \qquad \alpha_2^2 \sim \text{Gamma}\big(\tau_{\alpha_2^2}, \psi_{\alpha_2^2}\big) \tag{7}$$

$$\beta_1^1 \sim \text{Gamma}\big(\tau_{\beta_1^1}, \psi_{\beta_1^1}\big) \qquad \beta_1^2 \sim \text{Gamma}\big(\tau_{\beta_1^2}, \psi_{\beta_1^2}\big) \tag{8}$$

$$\beta_2^1 \sim \text{Gamma}\big(\tau_{\beta_2^1}, \psi_{\beta_2^1}\big) \qquad \beta_2^2 \sim \text{Gamma}\big(\tau_{\beta_2^2}, \psi_{\beta_2^2}\big) \tag{9}$$

where the $\tau > 0$ and $\psi > 0$ are the shape and rate parameters for the gamma priors.

Prior Likelihood.

For gamma distributed variable $x \sim \text{Gamma}(\tau, \psi)$, its likelihood can be written as

$$p(x \,|\, \tau, \psi) \propto x^{\tau - 1}\, e^{-\psi x} \tag{10}$$

Thus, for the priors, we explicitly write down their prior likelihood.The joint prior likelihood is thus

$$\begin{aligned} p(\mathbf{\Theta}) \propto\ & {\mu_1}^{\tau_{\mu_1}-1}\, e^{-\psi_{\mu_1}\mu_1} {\mu_2}^{\tau_{\mu_2}-1}\, e^{-\psi_{\mu_2}\mu_2}\, {\delta_1^1}^{\tau_{\delta_1^1}-1}\, e^{-\psi_{\delta_1^1}\delta_1^1} \\ & \times {\delta_1^2}^{\tau_{\delta_1^2}-1}\, e^{-\psi_{\delta_1^2}\delta_1^2}\, {\delta_2^1}^{\tau_{\delta_2^1}-1}\, e^{-\psi_{\delta_2^1}\delta_2^1}\, {\delta_2^2}^{\tau_{\delta_2^2}-1}\, e^{-\psi_{\delta_2^2}\delta_2^2} \\ & \times {\alpha_1^1}^{\tau_{\alpha_1^1}-1}\, e^{-\psi_{\alpha_1^1}\alpha_1^1}\, {\alpha_1^2}^{\tau_{\alpha_1^2}-1}\, e^{-\psi_{\alpha_1^2}\alpha_1^2}\, {\alpha_2^1}^{\tau_{\alpha_2^1}-1}\, e^{-\psi_{\alpha_2^1}\alpha_2^1} \\ & \times {\alpha_2^2}^{\tau_{\alpha_2^2}-1}\, e^{-\psi_{\alpha_2^2}\alpha_2^2}\, {\beta_1^1}^{\tau_{\beta_1^1}-1}\, e^{-\psi_{\beta_1^1}\beta_1^1}\, {\beta_1^2}^{\tau_{\beta_1^2}-1}\, e^{-\psi_{\beta_1^2}\beta_1^2} \\ & \times {\beta_2^1}^{\tau_{\beta_2^1}-1}\, e^{-\psi_{\beta_2^1}\beta_2^1}\, {\beta_2^2}^{\tau_{\beta_2^2}-1}\, e^{-\psi_{\beta_2^2}\beta_2^2} \end{aligned} \tag{11}$$

3.2 Model Joint Likelihood

Branching Structure.

Here, we use the branching representation of Hawkes processes described above. Note that *a priori* all branching structure is equally likely, thus for $\mathbf{A} = \{A_1^1, \ldots, A_{N^1(T)}^1, A_1^2, \ldots, A_{N^2(T)}^2\}$, we have

$$p(\mathbf{A}) = \prod_{j=1}^{N^1(T)} p(A_j^1) \prod_{j=1}^{N^2(T)} p(A_j^2) \propto 1 \tag{12}$$

Levels of Excitation.

Recall that we assumed the levels of excitation Y follow

gamma distributions, outlined below.

$$Y^1_{1,j} \sim \text{i.i.d. Gamma}\big(\alpha^1_1, \beta^1_1\big)\,, \quad \text{for } j = 1, \ldots, N^1(T)\,, \quad (13)$$
$$Y^2_{1,j} \sim \text{i.i.d. Gamma}\big(\alpha^2_1, \beta^2_1\big)\,, \quad \text{for } j = 1, \ldots, N^1(T)\,, \quad (14)$$
$$Y^1_{2,j} \sim \text{i.i.d. Gamma}\big(\alpha^1_2, \beta^1_2\big)\,, \quad \text{for } j = 1, \ldots, N^2(T)\,, \quad (15)$$
$$Y^2_{2,j} \sim \text{i.i.d. Gamma}\big(\alpha^2_2, \beta^2_2\big)\,, \quad \text{for } j = 1, \ldots, N^2(T)\,, \quad (16)$$

For all $\mathbf{Y}$, the joint likelihood is the product of all individual likelihood.

$$\begin{aligned} p(\mathbf{Y} \,|\, \alpha, \beta) = &\prod_{j=1}^{N^1(T)} \frac{(\beta^1_1)^{\alpha^1_1}}{\Gamma(\alpha^1_1)} (Y^1_{1,j})^{\alpha^1_1 - 1}\, e^{-\beta^1_1 Y^1_{1,j}} \\ &\times \frac{(\beta^2_1)^{\alpha^2_1}}{\Gamma(\alpha^2_1)} (Y^2_{1,j})^{\alpha^2_1 - 1}\, e^{-\beta^2_1 Y^2_{1,j}} \\ &\times \prod_{j=1}^{N^2(T)} \frac{(\beta^1_2)^{\alpha^1_2}}{\Gamma(\alpha^1_2)} (Y^1_{2,j})^{\alpha^1_2 - 1}\, e^{-\beta^1_2 Y^1_{2,j}} \\ &\times \frac{(\beta^2_2)^{\alpha^2_2}}{\Gamma(\alpha^2_2)} (Y^2_{2,j})^{\alpha^2_2 - 1}\, e^{-\beta^2_2 Y^2_{2,j}} \end{aligned} \quad (17)$$

Event Times.

The joint conditional likelihood for *all* event times $\mathbf{t} = \{t^1_1, \ldots, t^1_{N^1(T)}, t^2_1, \ldots, t^2_{N^2(T)}\}$ can be derived, using Theorem 2.4.VI (superposition theory) and Proposition 7.2.III (likelihood) in Daley and Vere-Jones (2003), conditioned on the levels of excitations $\mathbf{Y}$, the branching structure $\mathbf{A}$, and the parameters $\mathbf{\Theta}$:

$$\begin{aligned} p(\mathbf{t} \,|\, \mathbf{Y}, \mathbf{\Theta}, \mathbf{A}) = &\left(\prod_{j=1}^{N^1(T)} \lambda_1(t^1_j) \right) \exp\big(-\Lambda_1(T) \big) \\ &\times \left(\prod_{j=1}^{N^2(T)} \lambda_2(t^2_j) \right) \exp\big(-\Lambda_2(T) \big) \end{aligned} \quad (18)$$

where $\lambda_1(t^1_j)$ is the intensity (conditional on $\mathbf{A}$) that generated event time t^1_j, similarly for $\lambda_2(t^2_j)$. More explicitly, for $\lambda_1(t^1_j)$, we have

$$\begin{aligned} \lambda_1(t^1_j) = (\mu_1)^{(A^1_j)_{00}} &\Big(Y^1_1(0)\, e^{-\delta^1_1 t^1_j} \Big)^{(A^1_j)_{10}} \\ &\times \prod_{k=1}^{N^1(T)} \Big(Y^1_{1,k}\, e^{-\delta^1_1 (t^1_j - t^1_k)} \Big)^{(A^1_j)_{1k}} \\ &\times \Big(Y^2_1(0)\, e^{-\delta^2_1 t^1_j} \Big)^{(A^1_j)_{20}} \\ &\times \prod_{k=1}^{N^2(T)} \Big(Y^2_{1,k}\, e^{-\delta^2_1 (t^1_j - t^2_k)} \Big)^{(A^1_j)_{2k}} \end{aligned} \quad (19)$$

while similar can be written for $\lambda_2(t^2_j)$. On the other hand, $\Lambda_1(t)$ and $\Lambda_2(t)$ are the compensator for process 1 and process 2 respectively, which are also known as the *integrated intensity function*, computed as the integral of the intensity function defined by Equation (1) and (2). Here, $\Lambda_1(t)$

$$\begin{aligned} &= \int_0^t \lambda_1(s)\, \mathrm{d}s \\ &= \mu_1 t + \frac{Y^1_1(0)}{\delta^1_1} \left(1 - e^{-\delta^1_1 t} \right) + \sum_{j=1}^{N^1(t)} \frac{Y^1_{1,j}}{\delta^1_1} \left(1 - e^{-\delta^1_1 (t - t^1_j)} \right) \\ &\qquad + \frac{Y^2_1(0)}{\delta^2_1} \left(1 - e^{-\delta^2_1 t} \right) + \sum_{j=1}^{N^2(t)} \frac{Y^2_{1,j}}{\delta^2_1} \left(1 - e^{-\delta^2_1 (t - t^2_j)} \right) \end{aligned} \quad (20)$$

noting that similar can be derived for $\Lambda_2(t)$.

Full Joint Likelihood.

The full joint likelihood for the Hawkes process can thus be written as follows:

$$p(\mathbf{t}, \mathbf{Y}, \mathbf{\Theta}, \mathbf{A}) \propto p(\mathbf{\Theta})\, p(\mathbf{A})\, p(\mathbf{Y} \,|\, \alpha, \beta)\, p(\mathbf{t} \,|\, \mathbf{Y}, \mathbf{\Theta}, \mathbf{A})\,. \quad (21)$$

We note that we can recover the full joint likelihood of the marked multivariate Hawkes process without branching structure by marginalising out the branching structure $\mathbf{A}$. That is, by summing out all possibilities of $\mathbf{A}$:

$$p(\mathbf{t}, \mathbf{Y}, \mathbf{\Theta}) = \sum_{\mathbf{A}} p(\mathbf{t}, \mathbf{Y}, \mathbf{\Theta}, \mathbf{A}) \quad (22)$$

4. FULLY BAYESIAN INFERENCE

A hybrid of MCMC algorithms that updates the parameters one at a time, either by direct draws using Gibbs sampling or through the Metropolis Hastings (MH) algorithm. For ease of readability, we restate the framework as in Lee et al. (2016) which we briefly describe:

Let θ_A and θ_B be parameters of interest. Assume that the posterior $p(\theta_B \,|\, \theta_A)$ is of a known distribution, we can perform inference directly utilising the Gibbs sampler. On the other hand, suppose $p(\theta_A \,|\, \theta_B)$ can only be evaluated but not directly sampled; thus resort to the use of an MH algorithm to update θ_A given θ_B. For the MH step, the candidate θ'_A is drawn from $q(\theta'_A \,|\, \theta^{(k)}_A, \theta^{(k)}_B)$, which indicates that the current step can depend on the past draw of θ_A. The Metropolis step samples from $q(\theta'_A \,|\, \theta^{(k)}_A, \theta^{(k)}_B)$ which implies that we draw $\theta^{(k+1)}_A \sim q(\theta'_A \,|\, \theta^{(k)}_A, \theta^{(k)}_B)$ and that the criteria to accept or reject the proposal candidate is based on the acceptance probability, denoted by

$$\min\left(1, \frac{p(\theta'_A \,|\, \theta_B)\, q(\theta^{(k)}_A \,|\, \theta'_A, \theta^{(k)}_B)}{p(\theta^k_A \,|\, \theta_B)\, q(\theta'_A \,|\, \theta^{(k)}_A, \theta^{(k)}_B)} \right). \quad (23)$$

The hybrid algorithm is as follows: given $\big(\theta^{(0)}_A, \theta^{(0)}_B\big)$, for $k = 0, 1, ..., K$,

1. Sample $\theta^{(k+1)}_A \sim q(\theta'_A \,|\, \theta^{(k)}_A, \theta^{(k)}_B)$ and *accept* or *reject* $\theta^{(k+1)}_A$ based on equation (23).
2. Sample $\theta^{(k+1)}_B \sim p(\theta_B \,|\, \theta^{(k+1)}_A)$ with Gibbs sampling.

In this paper, however, our method consists fully of Gibbs samplers by employing adaptive rejection sampling (ARS, Gilks and Wild, 1992), a method for fast sampling from densities which are log-concave. This allows our sampling to be free of wastage present in MH algorithm, thus more efficient.

4.1 Posterior Likelihoods and Gibbs Sampling

We derive Gibbs samplers for learning the parameters. For the variables $\mathbf{A}$, $\mathbf{Y}$, μ, and β, we sample from their posteriors directly (since their posteriors follow known distributions), while for the parameters δ, and α, we adopt the ARS to sample from their posteriors. In the following, we first derive the posterior distributions for the parameters of interest.

Due to space, some derivations are omitted. The full derivations will be made available in an online appendix.

4.1.1 Gibbs Sampler for A_j^1 and A_j^2

We first note that the posterior of $\mathbf{A}$ can be derived as

$$p(\mathbf{A} \mid \mathbf{t}, \mathbf{Y}, \mathbf{\Theta}) \propto \prod_{j=1}^{N^1(T)} \left[(\mu_1)^{(A_j^1)_{00}} \prod_{i=1}^{M} \left(Y_1^i(0)\, e^{-\delta_1^i t_j^1} \right)^{(A_j^1)_{i0}} \times \prod_{k=1}^{N^i(T)} \left(Y_{1,k}^i\, e^{-\delta_1^i (t_j^1 - t_k^i)} \right)^{(A_j^1)_{ik}} \right] \times \prod_{j=1}^{N^2(T)} \left[(\mu_2)^{(A_j^2)_{00}} \prod_{i=1}^{M} \left(Y_2^i(0)\, e^{-\delta_2^i t_j^2} \right)^{(A_j^2)_{i0}} \times \prod_{k=1}^{N^i(T)} \left(Y_{2,k}^i\, e^{-\delta_2^i (t_j^2 - t_k^i)} \right)^{(A_j^2)_{ik}} \right] \quad (24)$$

From this, we can see that each A_j^1 and A_j^2 follows a Multinomial posterior:

$$A_j^1 \mid \mathbf{t}, \mathbf{Y}, \mathbf{\Theta} \sim \text{Multinomial}(1, Q_j^1) \quad (25)$$

$$A_j^2 \mid \mathbf{t}, \mathbf{Y}, \mathbf{\Theta} \sim \text{Multinomial}(1, Q_j^2) \quad (26)$$

where Q_j^1 and Q_j^2 are probability matrices:

$$Q_j^1 \coloneqq \begin{bmatrix} (Q_j^1)_{00} & 0 & \cdots & 0 \\ (Q_j^1)_{10} & \cdots & (Q_j^1)_{1k} & \cdots \\ (Q_j^1)_{20} & \cdots & (Q_j^1)_{2k} & \cdots \end{bmatrix}, \quad (27)$$

$$Q_j^2 \coloneqq \begin{bmatrix} (Q_j^2)_{00} & 0 & \cdots & 0 \\ (Q_j^2)_{10} & \cdots & (Q_j^2)_{1k} & \cdots \\ (Q_j^2)_{20} & \cdots & (Q_j^2)_{2k} & \cdots \end{bmatrix}. \quad (28)$$

The entries for Q_j^1 are given as follows.

$$(Q_j^1)_{00} = \mu_1 / \lambda_1(t_j^1), \quad (29)$$

$$(Q_j^1)_{10} = \left(Y_1^1(0)\, e^{-\delta_1^1 t_j^1} \right) / \lambda_1(t_j^1), \quad (30)$$

$$(Q_j^1)_{1k} = \left(Y_{1,k}^1\, e^{-\delta_1^1 (t_j^1 - t_k^1)} \right) / \lambda_1(t_j^1), \quad (31)$$

$$(Q_j^1)_{20} = \left(Y_1^2(0)\, e^{-\delta_1^2 t_j^1} \right) / \lambda_1(t_j^1), \quad (32)$$

$$(Q_j^1)_{2k} = \left(Y_{1,k}^2\, e^{-\delta_1^2 (t_j^1 - t_k^2)} \right) / \lambda_1(t_j^1). \quad (33)$$

Similarly for Q_j^2. In the Gibbs sampler, we sample new A_j^1 and A_j^2 directly from their posterior.

4.1.2 Gibbs Sampler for $Y_{1,j}^1$, $Y_{1,j}^2$, $Y_{2,j}^1$, and $Y_{2,j}^2$

With a conjugate Gamma prior on $Y_{1,j}^1$, $Y_{1,j}^2$, $Y_{2,j}^1$, and $Y_{2,j}^2$, the posterior for $Y_{1,j}^1$, $Y_{1,j}^2$, $Y_{2,j}^1$, and $Y_{2,j}^2$ follows a Gamma distribution:

$$Y_{1,j}^1 \mid \mathbf{t}, \mathbf{A}, \mathbf{\Theta} \sim \text{Gamma}\big((\alpha_1^1)^*, (\beta_1^1)^*\big) \quad (34)$$

$$Y_{1,j}^2 \mid \mathbf{t}, \mathbf{A}, \mathbf{\Theta} \sim \text{Gamma}\big((\alpha_1^2)^*, (\beta_1^2)^*\big) \quad (35)$$

$$Y_{2,j}^1 \mid \mathbf{t}, \mathbf{A}, \mathbf{\Theta} \sim \text{Gamma}\big((\alpha_2^1)^*, (\beta_2^1)^*\big) \quad (36)$$

$$Y_{2,j}^2 \mid \mathbf{t}, \mathbf{A}, \mathbf{\Theta} \sim \text{Gamma}\big((\alpha_2^2)^*, (\beta_2^2)^*\big) \quad (37)$$

where

$$(\alpha_1^1)^* = \alpha_1^1 + \sum_{k=1}^{N^1(T)} (A_k^1)_{1j} \quad (38)$$

$$(\beta_1^1)^* = \beta_1^1 + \frac{1}{\delta_1^1}\left(1 - e^{-\delta_1^1 (T - t_k^1)}\right) \quad (39)$$

and the other α^* and β^* are similar.

4.1.3 Gibbs Sampler for $Y_1^1(0)$, $Y_1^2(0)$, $Y_2^1(0)$, and $Y_2^2(0)$

The Gibbs sampler for $Y_1^1(0)$, $Y_1^2(0)$, $Y_2^1(0)$, and $Y_2^2(0)$ can similarly be derived.

$$Y_1^1(0) \mid \mathbf{A}, \mathbf{\Theta}, \tau, \psi \sim \text{Gamma}\left(\tau^*_{Y_1^1(0)}, \psi^*_{Y_1^1(0)}\right) \quad (40)$$

$$Y_1^2(0) \mid \mathbf{A}, \mathbf{\Theta}, \tau, \psi \sim \text{Gamma}\left(\tau^*_{Y_1^2(0)}, \psi^*_{Y_1^2(0)}\right) \quad (41)$$

$$Y_2^1(0) \mid \mathbf{A}, \mathbf{\Theta}, \tau, \psi \sim \text{Gamma}\left(\tau^*_{Y_2^1(0)}, \psi^*_{Y_2^1(0)}\right) \quad (42)$$

$$Y_2^2(0) \mid \mathbf{A}, \mathbf{\Theta}, \tau, \psi \sim \text{Gamma}\left(\tau^*_{Y_2^2(0)}, \psi^*_{Y_2^2(0)}\right) \quad (43)$$

where

$$\tau^*_{Y_1^1(0)} = \tau_{Y_1^1(0)} + \sum_{j=1}^{N^1(T)} (A_j^1)_{10} \quad (44)$$

$$\psi^*_{Y_1^1(0)} = \psi_{Y_1^1(0)} + \frac{1}{\delta_1^1}\left(1 - e^{-\delta_1^1 T}\right). \quad (45)$$

The other τ^* amd ψ^* can similarly be written down.

4.1.4 Gibbs Sampler for μ_1 and μ_2

The posterior for μ_m can be derived as:

$$\mu_1 \mid \ldots \sim \text{Gamma}\left(\tau_{\mu_1} + \sum_{j=1}^{N^1(T)} (A_j^1)_{00}, \psi_{\mu_1} + T\right) \quad (46)$$

$$\mu_2 \mid \ldots \sim \text{Gamma}\left(\tau_{\mu_2} + \sum_{j=1}^{N^2(T)} (A_j^2)_{00}, \psi_{\mu_2} + T\right) \quad (47)$$

4.1.5 Gibbs sampler for β_1^1, β_1^2, β_2^1, and β_2^2

We note that Gamma priors on the β_1^1, β_1^2, β_2^1, and β_2^2 give Gamma posteriors.

$$\beta_1^1 \mid \ldots \sim \text{Gamma}\left(\tau_{\beta_1^1} + N^1(T)\,\alpha_1^1,\ \psi_{\beta_1^1} + \sum_{i=1}^{N^1(T)} Y_{1,j}^1\right) \quad (48)$$

$$\beta_1^2 \mid \ldots \sim \text{Gamma}\left(\tau_{\beta_1^2} + N^2(T)\,\alpha_1^2,\ \psi_{\beta_1^2} + \sum_{i=1}^{N^2(T)} Y_{1,j}^2\right) \quad (49)$$

$$\beta_2^1 \mid \ldots \sim \text{Gamma}\left(\tau_{\beta_2^1} + N^1(T)\,\alpha_2^1,\ \psi_{\beta_2^1} + \sum_{i=1}^{N^1(T)} Y_{2,j}^1\right) \quad (50)$$

$$\beta_2^2 \mid \ldots \sim \text{Gamma}\left(\tau_{\beta_2^2} + N^2(T)\,\alpha_2^2,\ \psi_{\beta_2^2} + \sum_{i=1}^{N^2(T)} Y_{2,j}^2\right) \quad (51)$$

4.1.6 *Posterior for δ_1^1, δ_1^2, δ_2^1, and δ_2^2*

The posterior for δ_1^1, δ_1^2, δ_2^1, and δ_2^2 can be derived as

$$p(\delta_1^1 \,|\, \mathbf{t}, \mathbf{Y}, \mathbf{A}, \tau, \psi)$$
$$\propto \exp\left[-\delta_1^1 \left(\psi_{\delta_1^1} + \sum_{j=1}^{N^1(T)} \sum_{k=0:\, t_k^1 < t_j^1}^{N^1(T)} (A_j^1)_{1k} (t_j^1 - t_k^1) \right) - \sum_{k=0}^{N^1(T)} \frac{Y_{1,k}^1}{\delta_1^1} \left(1 - e^{-\delta_1^1 (T - t_k^1)} \right) \right] (\delta_1^1)^{\tau_{\delta_1^1} - 1} \quad (52)$$

$$p(\delta_1^2 \,|\, \mathbf{t}, \mathbf{Y}, \mathbf{A}, \tau, \psi)$$
$$\propto \exp\left[-\delta_1^2 \left(\psi_{\delta_1^2} + \sum_{j=1}^{N^1(T)} \sum_{k=0:\, t_k^2 < t_j^1}^{N^2(T)} (A_j^1)_{2k} (t_j^1 - t_k^2) \right) - \sum_{k=0}^{N^2(T)} \frac{Y_{1,k}^2}{\delta_1^2} \left(1 - e^{-\delta_1^2 (T - t_k^2)} \right) \right] (\delta_1^2)^{\tau_{\delta_1^2} - 1} \quad (53)$$

$$p(\delta_2^1 \,|\, \mathbf{t}, \mathbf{Y}, \mathbf{A}, \tau, \psi)$$
$$\propto \exp\left[-\delta_2^1 \left(\psi_{\delta_2^1} + \sum_{j=1}^{N^2(T)} \sum_{k=0:\, t_k^1 < t_j^2}^{N^1(T)} (A_j^2)_{1k} (t_j^2 - t_k^1) \right) - \sum_{k=0}^{N^1(T)} \frac{Y_{2,k}^1}{\delta_2^1} \left(1 - e^{-\delta_2^1 (T - t_k^1)} \right) \right] (\delta_2^1)^{\tau_{\delta_2^1} - 1} \quad (54)$$

$$p(\delta_2^2 \,|\, \mathbf{t}, \mathbf{Y}, \mathbf{A}, \tau, \psi)$$
$$\propto \exp\left[-\delta_2^2 \left(\psi_{\delta_2^2} + \sum_{j=1}^{N^2(T)} \sum_{k=0:\, t_k^2 < t_j^2}^{N^2(T)} (A_j^2)_{2k} (t_j^2 - t_k^2) \right) - \sum_{k=0}^{N^2(T)} \frac{Y_{2,k}^2}{\delta_2^2} \left(1 - e^{-\delta_2^2 (T - t_k^2)} \right) \right] (\delta_2^2)^{\tau_{\delta_2^2} - 1} \quad (55)$$

where we have introduced $Y_{1,0}^1$, $Y_{1,0}^2$, $Y_{2,0}^1$, and $Y_{2,0}^2$ as a short hand for $Y_1^1(0)$, $Y_1^2(0)$, $Y_2^1(0)$, and $Y_2^2(0)$, and $t_0 = 0$ to simplify the posterior. We note that this posterior is log-concave when $\tau_{\delta_1^1} \geq 1$.

In Figure 1, we present a plot showing several log posterior likelihood function for the parameter δ obtained during inference.

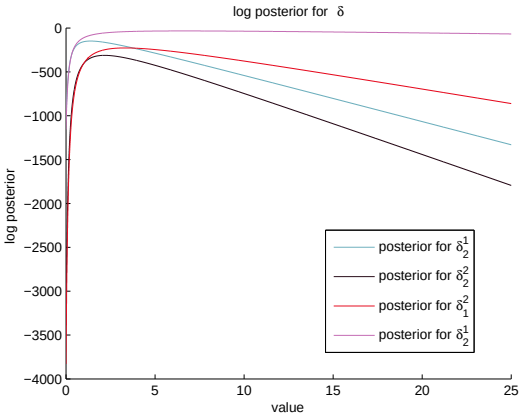

Figure 1: Log posterior of δ during inference.

4.1.7 *Posterior for α_1^1, α_1^2, α_2^1, and α_2^2*

The posteriors for α_1^1, α_1^2, α_2^1, and α_2^2 can be derived as

$$p(\alpha_1^1 \,|\, \mathbf{Y}, \beta, \tau, \psi) \propto (\alpha_1^1)^{\tau_{\alpha_1^1} - 1} [\Gamma(\alpha_1^1)]^{-N^1(T)} \times \left((\beta_1^1)^{N^1(T)} e^{-\psi_{\alpha_1^1}} \prod_{k=1}^{N^1(T)} Y_{1,k}^1 \right)^{\alpha_1^1} \quad (56)$$

$$p(\alpha_1^2 \,|\, \mathbf{Y}, \beta, \tau, \psi) \propto (\alpha_1^2)^{\tau_{\alpha_1^2} - 1} [\Gamma(\alpha_1^2)]^{-N^2(T)} \times \left((\beta_1^2)^{N^2(T)} e^{-\psi_{\alpha_1^2}} \prod_{k=1}^{N^2(T)} Y_{1,k}^2 \right)^{\alpha_1^2} \quad (57)$$

$$p(\alpha_2^1 \,|\, \mathbf{Y}, \beta, \tau, \psi) \propto (\alpha_2^1)^{\tau_{\alpha_2^1} - 1} [\Gamma(\alpha_2^1)]^{-N^1(T)} \times \left((\beta_2^1)^{N^1(T)} e^{-\psi_{\alpha_2^1}} \prod_{k=1}^{N^1(T)} Y_{2,k}^1 \right)^{\alpha_2^1} \quad (58)$$

$$p(\alpha_2^2 \,|\, \mathbf{Y}, \beta, \tau, \psi) \propto (\alpha_2^2)^{\tau_{\alpha_2^2} - 1} [\Gamma(\alpha_2^2)]^{-N^2(T)} \times \left((\beta_2^2)^{N^2(T)} e^{-\psi_{\alpha_2^2}} \prod_{k=1}^{N^2(T)} Y_{2,k}^2 \right)^{\alpha_2^2} \quad (59)$$

We note that these posteriors are log-concave when $\tau_{\alpha_1^1} > 1$. Figure 2 graphs several log posterior for the parameters α.

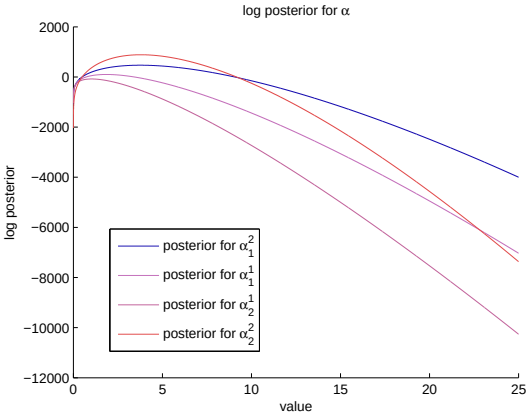

Figure 2: Log posterior of α during inference.

5. SYNTHETIC VALIDATION

In this section, we perform experiments to test the ability of our proposed inference algorithm in learning the ground truth parameters used in generating synthetic data. Testing on the synthetic data allows us to gauge how well the MCMC algorithm performs. The ground truth parameters are displayed in Table 1. The time to maturity is set to $T = 20$.

We repeat the experiments for 1,000 times, with a different simulated samples of bivariate Hawkes each time (refer to Lim et al. (2016) for simulation of Hawkes processes). We obtain the learned parameters after performing the inference algorithm for 250 iterations (include 50 iterations of burn-in) for each synthetic dataset. The MCMC algorithm is found to outperform the maximum likelihood estimation (MLE) in terms of mean square error on parameters estimation. This result is presented in Table 1.

Table 1: Comparison of the parameters learned by MCMC against the MLE on a synthetic dataset. These parameters, learned from 1,000 simulated samples of bivariate Hawkes, are quite close to the ground truth parameters. Overall, the MCMC estimates achieve a lower mean square error compared to the MLE.

		Process $m = 1$			Process $m = 2$		
Name	Var.	True	MLE	MCMC	True	MLE	MCMC
Background intensity	μ_m	1.0000	1.0542	0.5618	2.0000	2.5641	2.1212
Initial added intensity	$Y^1_{m,0}$	1.0000	1.8160	1.4421	2.0000	2.1771	1.1321
	$Y^2_{m,0}$	2.0000	1.6143	1.3724	1.0000	1.5932	1.1001
Decay rates	δ^1_m	5.0000	6.2410	5.4114	3.0000	4.9071	2.8089
	δ^2_m	4.0000	4.2095	3.9237	2.0000	2.8092	2.7452
Shape parameters	α^1_m	2.0000	2.0329	2.0221	1.0000	1.0149	1.0102
	α^2_m	3.0000	3.0453	3.0208	4.0000	4.0729	4.0279
Rate parameters	β^1_m	1.0000	1.0082	1.0161	2.0000	2.0176	2.0449
	β^2_m	2.0000	2.0151	2.0155	3.0000	3.0265	3.0276
Mean square error	MSE	-	0.2672	**0.1064**	-	0.5555	**0.1526**

6. CONCLUSION

For the bivariate Hawkes process, a novel inference procedure based on fully Bayesian Gibbs samplers through adaptive rejection sampling is presented. We exploited the inherent branching structures for the bivariate Hawkes processes and augment the parameter space. The conditions for log-concavity tied to the posterior is spelled out. Experiments show that our inference algorithm performs well in learning the ground truth parameters compared to the MLE.

References

Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(1):1–59.

Cox, D. R. and Isham, V. (1980). *Point Processes*. Chapman & Hall.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, volume I: Elementary Theory and Methods. Springer, 2nd edition.

Dassios, A. and Zhao, H. (2013). Exact simulation of Hawkes process with exponentially decaying intensity. *Electronic Communications in Probability*, 18:1–13.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503.

Laub, P. J., Taimre, T., and Pollett, P. K. (2015). Hawkes processes. *ArXiv e-prints*, 1507.02822:1–28.

Lee, Y., Lim, K. W., and Ong, C. S. (2016). Hawkes processes with stochastic excitations. In *International Conference in Machine Learning*, pages 79–88.

Lim, K. W., Lee, Y., Hanlen, L., and Zhao, H. (2016). Simulation and calibration of a fully Bayesian marked multidimensional Hawkes process with dissimilar decays. In *Asian Conference in Machine Learning*, pages 1–16.

Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of Hawkes processes. *Advances in Applied Probability*, 37(3):629–646.

Møller, J. and Rasmussen, J. G. (2006). Approximate simulation of Hawkes processes. *Methodology and Computing in Applied Probability*, 8(1):53–64.

Muni Toke, I. and Pomponio, F. (2012). Modelling trades-through in a limit order book using Hawkes processes. *Economics: The Open-Access, Open-Assessment E-Journal*, 6(2012-22).

Ozaki, T. (1979). Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155.

Rasmussen, J. G. (2013). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

Veen, A. and Schoenberg, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *Journal of the American Statistical Association*, 103(482):614–624.