Detecting Network Intrusion Using a Markov Modulated Nonhomogeneous Poisson Process

Steven L. Scott*

June 14, 2000

Abstract

Network intrusion occurs when a criminal gains access to a customer's telephone, computer, bank, or other type of account. Detecting network intrusion is an important problem that has received little attention in the statistics literature. This article proposes a Markov modulated nonhomogeneous Poisson process (MMNHPP) to monitor transactions on a customer's account for deviations from the customer's established behavior patterns. An important benefit of the MMNHPP is its ability to model the posterior probability of a criminal presence as a function of time. The MMNHPP combines aspects of the Markov modulated Poisson process and the nonhomogeneous Poisson process to model point processes exhibiting both regular patterns and irregular bursts of activity. The need to accommodate both types of behavior is demonstrated using data from two telephone accounts.

MMNHPP parameters are sampled from their posterior distribution given a set of observed event times using an MCMC algorithm. The algorithm switches between drawing missing descriptions of criminal activity given model parameters and sampling model parameters given complete data. An augmented variables scheme is used to render otherwise strongly related elements of the missing data independent in their posterior distribution. Stochastic forward backward recursions for nonstationary hidden Markov models allow the augmenting variables, and thus the entire missing data vector, to be sampled by a single Gibbs step.

*Assistant Professor of Statistics, The Marshall School of Business, University of Southern California, Los Angeles, CA 90089. Email: sls@usc.edu.

Key Words: forward-backward recursions, Gibbs sampler, hidden Markov model, network intrusion, augmented variables

1 Introduction

This article presents a model for screening an account against criminal intrusion. The example considered is a telephone account victimized by a criminal who gains access to the account and generates unauthorized traffic. Network intrusion detection is of great interest to long distance telephone companies in the U.S., who lose an estimated $4 billion per year to this type of fraud. Similar problems are faced by credit card companies, computer system administrators, and increasingly by users of the World Wide Web.

Network intrusion detection involves monitoring the point process describing an account's transactions to detect deviations from a customer's established behavior patterns. Our approach is to model traffic on the account as a superposition of customer and criminal traffic, while considering the likely possibility that no criminal is present. We model the customer's behavior using a parametric nonhomogeneous Poisson process. We suppose a two state Markov process governs the criminal's presence or absence, and that the criminal generates traffic according to a homogeneous Poisson process while he is present. The Markov process introduces bursts of activity typically associated with episodes of criminal contamination. The homogeneous Poisson process represents the most serious type of telephone fraud: when a criminal sells access to an account as if he were a legitimate long distance carrier. Thus the "criminal" is actually a large group whose individual calling habits aggregate to a homogeneous rate. The homogeneous Poisson process is also chosen for its simplicity. Because the observed data are a superposition of customer and criminal processes, we lack reliable information indicating which calls are due to the criminal rather than the customer. Estimating an elaborate model for criminal behavior would spread limited information in the data over a large number of parameters.

A Poisson process whose rate parameter varies according to a Markov process is Markov modulated Poisson process (MMPP). The MMPP is a flexible model for point processes whose event rates vary among different levels at irregular intervals. The MMPP is frequently seen in queuing theory

(Du, 1995; Olivier and Walrand, 1994), but it is rare in statistics. An exception is Davison and Ramesh (1996), who apply an MMPP to a binary time series of precipitation data by numerically optimizing the MMPP likelihood. Hidden Markov models (Scott, 2000; MacDonald and Zucchini, 1997), which are discrete time cousins of the MMPP, are more prevalent in the statistics literature. Turin (1996) derives the MMPP as a limiting case of hidden Markov models and uses the EM algorithm to obtain maximum likelihood estimates of MMPP parameters. Scott (1999) exploits a link between the MMPP and hidden Markov models to produce an efficient Gibbs sampler for the MMPP. None of the above references accommodates the nonhomogeneous Poisson process at the core of our model. Many telephone customers have predictable daily and hourly calling patterns. Section 5 demonstrates the need to incorporate these patterns in the model for customer behavior. We therefore introduce the Markov modulated nonhomogeneous Poisson process (MMNHPP) to model point processes exhibiting both regular patterns and irregular bursts of extra activity.

We employ a Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970) to sample the posterior distribution of MMNHPP parameters. The algorithm alternates between simulating missing descriptions of criminal activity given observed data and model parameters and sampling model parameters given complete data. A key component of the sampling algorithm is an augmented variables scheme (Besag and Green, 1993; Higdon, 1998; Damian *et al.*, 1999) allowing the missing data to be drawn in a single Gibbs step. Given model parameters, the posterior probability of a criminal presence can be computed as a function of time.

The remaining sections proceed as follows. Section 2 defines the model. Section 3 describes the posterior sampling algorithm. Section 4 examines the algorithm's performance on simulated data. Section 5 applies the algorithm to two telephone accounts, one that was contaminated by fraud and one that was not. Section 6 concludes with a discussion of the advantages the MMNHPP possesses over simpler intrusion detection methods. Two appendices develop forward-backward recursions used by the posterior simulation algorithm.

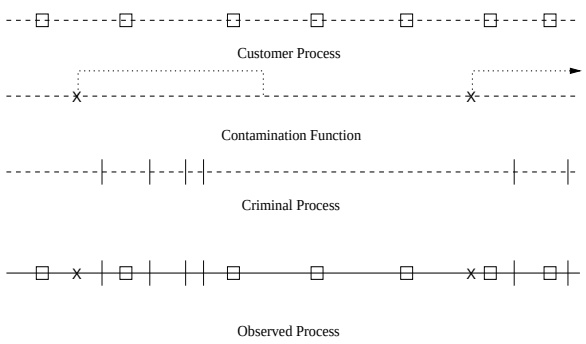


Figure 1: Depiction of an account victimized by fraud. Legitimate transactions (produced by N_0) are marked with boxes. The beginnings of contamination intervals ($C(\cdot)$) are marked with X's, and fraudulent transactions in the interior of a contamination interval (N_1) are marked with vertical lines.

2 The Markov Modulated Nonhomogeneous Poisson Process

Section 2.1 defines the MMNHPP as a superposition of stochastic processes. Section 2.2 expresses the stochastic processes as an equivalent set of random variables suitable for analysis by traditional missing data techniques. Section 2.3 describes our parameterization of the nonhomogeneous Poisson process, and Section 2.4 describes related modeling attempts.

2.1 Model Description

Transactions in an interval $(a, b] \subset (t_0, T]$ are generated by an observed point process $N(a, b]$, where $N(t_0, T] = n$. In our example "transactions" are placement times for telephone calls. In other applications transactions could be credit card purchases or requests for computer resources. The process N is a superposition of three component processes N_0, C, and N_1 describing the customer's behavior, the criminal's presence or absence, and the criminal's behavior while he is present (see Figure 1). Both N_0 and N_1 are point processes, and C is a random step function known as the contamination function. Let $C(t) = 1$ if a criminal is present at time t, and $C(t) = 0$ otherwise. Assume N_0 is a Poisson process with a parametric rate function $\lambda_0(t)$ defined in Section 2.3. Model N_1 as a homogeneous Poisson process with rate λ_1, and assume $C(t)$ obeys a Markov process with generator matrix

$$\Gamma = \begin{pmatrix} -\gamma & \gamma \\ \phi & -\phi \end{pmatrix},$$

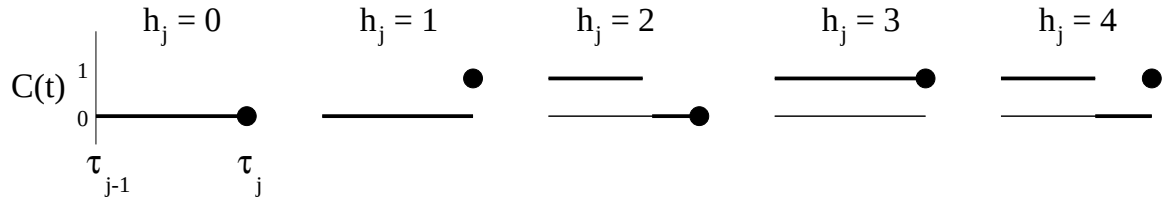


Figure 2: The five possible states describing the behavior of $C(\cdot)$ over $(\tau_{j-1}, \tau_j]$. Heavy lines denote values of $C(t)$. Heavy dots indicate $C(\tau_j)$, which differs from $C(\tau_j - \epsilon)$ when a new contamination episode begins at τ_j. A transition from 1 to 0 is only possible in the interior of the interval, while a transition from 0 to 1 is only possible at the right endpoint.

that is, successive waiting times to arrivals and departures are independent exponential random variables with rates γ and ϕ.

Define a criminal's arrival as the time of his first fraudulent event in a contamination episode, eliminating the nonsensical possibility of empty contamination intervals. An equivalent assumption is that a transition of $C(\cdot)$ from 0 to 1 generates a fraudulent event. Allow contamination episodes containing a single fraudulent event by assuming transitions of $C(\cdot)$ from 1 to 0 generate no traffic.

2.2 The $\boldsymbol{\tau}$, h, w, y Decomposition

Denote transaction times by $\tau_1 < \cdots < \tau_n$, and let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)$. This section discretizes $C(\cdot)$, without information loss, by classifying its behavior over a partition of $(t_0, T]$ determined by $\boldsymbol{\tau}$. For convenience let $\tau_0 = t_0$ and let $I_j = (\tau_{j-1}, \tau_j]$, $j = 1, \ldots, n$. We omit discussion of the final interval $(\tau_n, T]$, an easily accommodated but tedious detail. Let $w_j = \int_{I_j} C(t)\ dt$ represent the amount of time the criminal was present during I_j. Let $y_j = 0$ if N_0 produced event j and $y_j = 1$ otherwise. There is a one-to-one mapping between $(N_0, C(\cdot), N_1)$ and the vectors $\boldsymbol{\tau}$, $\mathbf{w} = (w_1, \ldots, w_n)$, and $\mathbf{y} = (y_1, \ldots, y_n)$.

The definition of criminal arrival in Section 2.1 restricts the behavior of $C(\cdot)$ over I_j to one of five possible states illustrated in Figure 2. Let h_j denote the state best describing $C(\cdot)$ over I_j. The vector $\mathbf{h} = (h_1, \ldots, h_n)$ is useful for estimation, though superfluous as a descriptor. There is a strong relationship between $\mathbf{w}$ and $\mathbf{y}$, reflecting the fact that there can be no crime if there is no criminal. Conditioning on $\mathbf{h}$ decouples the elements of $(\mathbf{w}, \mathbf{y})$, rendering them independent in

their posterior distribution given $\boldsymbol{\tau}$ and model parameters. Thus $\mathbf{h}$ is an example of "augmented variables" in the spirit of Higdon (1998), Besag and Green (1993), and Damian *et al.* (1999). Henceforth, we refer to $\boldsymbol{\tau}$ as the observed data, $(\mathbf{h}, \mathbf{w}, \mathbf{y})$ as the missing data, and $(\boldsymbol{\tau}, \mathbf{h}, \mathbf{w}, \mathbf{y})$ as the complete data.

2.3 The Nonhomogeneous Poisson Process

This section defines the rate function for N_0 mentioned in Section 2.1. Number days of the week $\{0, \ldots, 6\}$, beginning with Saturday. Let $\text{day}(t) = d$ if the continuous time point t occurs during the d'th day of the week. Number hours of the day $\{0, \ldots, 23\}$, and define $\text{hour}(t)$ similarly. Let

$$\lambda_0(t) = \lambda_0 \delta_d \eta_{D(d)h} \tag{1}$$

when $\text{day}(t) = d$ and $\text{hour}(t) = h$. The notation $D(d)$ reduces the number of model parameters by creating equivalence classes for days of the week. In later sections we distinguish between weekend (we) and weekday (wd) calling patterns by $D(d) = we$ if d=0 or 1, and $D(d) = wd$ if $d > 1$. Restrict the parameters so all are non-negative,

$$\sum_{d=0}^{6} \delta_d = 7, \text{ and } \sum_{h=0}^{23} \eta_{D(d)h} = 24. \tag{2}$$

Let $\boldsymbol{\delta} = (\delta_d)$ and $\boldsymbol{\eta} = (\eta_{D(d)h})$. The parameters in (1) are easily interpretable. For instance, λ_0 is the long run average daily calling rate, and $\delta_d = 2$ implies events on day d occur at twice the long run average rate.

2.4 Related Work

The logic of our method is similar to Byers and Raftery (1998), who use a two dimensional Poisson process to detect minefields hidden among clutter. Our MMNHPP differs from the MMPP of Davison and Ramesh (1996), who observe the process at discrete, pre-chosen time points and use a homogeneous Poisson process for N_0. The nonhomogeneous Poisson process fit by Green (1995) can be expressed as a homogeneous MMPP with several levels of "criminal activity." Green's Poisson

process could therefore be fit without his reversible jump MCMC algorithm. Our parameterization of the nonhomogeneous Poisson process is closer to that of Kolaczyk (1999), though we do not take advantage of Kolaczyk's fast pseudo-wavelet transforms. The $\boldsymbol{\tau}$, $\mathbf{h}$, $\mathbf{w}$, $\mathbf{y}$ decomposition is due to Scott (1999). The use of $\mathbf{h}$ as an augmented variable is outside Damian *et al.* (1999)'s scheme because the full conditional distributions of the variables it decouples are not uniform. The assumption of constant λ_0 in (1) is convenient for the examples in Section 5, though perhaps unrealistic in other cases. Guo *et al.* (1999) fit a Gaussian state space model to a hormonal time series by superimposing random bursts on a slowly moving baseline. Guo *et al.*'s approach suggests the constant λ_0 assumption could be relaxed.

3 Posterior Computation

We simulate from the posterior distribution of model parameters using a Markov chain Monte Carlo algorithm alternating between draws of model parameters given complete data and draws of missing data given $\boldsymbol{\tau}$ and model parameters. Section 3.1 discusses the prior used to compute the posterior distribution. Section 3.2 explains the draw of the missing data. Section 3.3 explains how to draw model parameters given complete data.

3.1 Choice of Prior

Assume the elements of $\theta = (\lambda_0, \boldsymbol{\delta}, \boldsymbol{\eta}, \lambda_1, \gamma, \phi)$ to be *a priori* independent with

$$\lambda_1 \sim \Gamma(a_1, b_1), \quad \gamma \sim \Gamma(a_\gamma, b_\gamma), \quad \phi \sim \Gamma(a_\phi, b_\phi), \tag{3}$$

$$\lambda_0 \sim \Gamma(a_0, b_0), \quad \boldsymbol{\delta} \sim \mathcal{D}^7(\boldsymbol{\nu}_d), \qquad \boldsymbol{\eta} \sim \prod_D \mathcal{D}^{24}(\boldsymbol{\nu}_{Dh}), \tag{4}$$

where $\mathcal{D}^n(\cdot)$ represents the n dimensional Dirichlet distribution scaled so its deviates sum to n. The gamma density $\Gamma(\alpha, \beta)$ is parameterized so its mean is α/β and its variance is α/β^2. The product in (4) is over all distinct values of $D(d)$. Independent Dirichlet and gamma densities are chosen for their conjugacy properties and because of their easily interpretable hyperparameters. The a and $\boldsymbol{\nu}$ parameters correspond to prior event counts, and the b parameters are prior observation times.

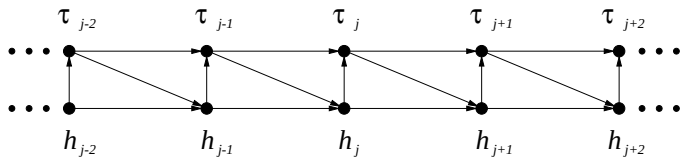


Figure 3: Directed acyclic graph for $(\mathbf{h}, \boldsymbol{\tau})$. The conditional distribution of a node given all its ancestors depends only on its parents.

The posterior distribution of $(\lambda_0, \boldsymbol{\delta}, \boldsymbol{\eta})$ is insensitive to reasonable prior assumptions about hyperparameters and functional forms provided a moderate amount of data is available. The posterior distribution of $(\lambda_1, \gamma, \phi)$ is sensitive to prior assumptions unless some elements of $(\mathbf{h}, \mathbf{w}, \mathbf{y})$ can actually be observed or unless the rate of criminal traffic is much higher than the customer's calling rate. Prior sensitivity is a further argument for (3) and (4) because of the straightforward correspondence between prior information and hypothetical data. Later sections elaborate on the prior sensitivity of $(\lambda_1, \gamma, \phi)$.

3.2 Drawing Missing Data Given Model Parameters

When N_0 is a homogeneous Poisson process, Scott (1999) introduces a method for drawing $(\mathbf{h}, \mathbf{w}, \mathbf{y})$ directly from its conditional distribution given $\boldsymbol{\tau}$ and θ. In other words, the simulation does not involve cycling through and drawing each (h_j, w_j, y_j) given all else. The method observes that $(\boldsymbol{\tau}, \mathbf{h})$ forms the observed and latent data from a hidden Markov model. It then applies stochastic forward-backward recursions for hidden Markov models (Scott, 2000) to simulate $\mathbf{h}$ from $p(\mathbf{h}|\boldsymbol{\tau}, \theta)$. Replacing N_0 with a nonhomogeneous Poisson process replaces the hidden Markov relationship between $\boldsymbol{\tau}$ and $\mathbf{h}$ with

$$Pr(h_j = s|h_1, \ldots, h_{j-1}, \tau_1, \ldots, \tau_{j-1}, \theta) = Pr(h_j = s|h_{j-1}, \tau_{j-1}, \theta), \tag{5}$$

$$p(\tau_j|\tau_1, \ldots, \tau_{j-1}, h_1, \ldots, h_j, \theta) = p(\tau_j|\tau_{j-1}, h_j, \theta), \tag{6}$$

which would describe a hidden Markov relationship if neither equation depended on τ_{j-1}. Figure 3 shows the directed acyclic graph (Whittaker, 1990) describing (5) and (6). Appendix A develops stochastic forward-backward recursions to simulate from $p(\mathbf{h}|\boldsymbol{\tau}, \theta)$ under the nonstationary hidden

Markov model (5) – (6). Appendix B presents specific computational details customizing the recursions for the MMNHPP.

Once $\mathbf{h}$ is drawn, the elements of $(\mathbf{w}, \mathbf{y})$ are conditionally independent of one another given $\mathbf{h}$, $\boldsymbol{\tau}$, and θ. The value of h_j determines y_j unless $h_j = 3$, when

$$Pr(y_j = 1 | h_j = 3, \tau_j, \theta) = \frac{\lambda_1}{\lambda_1 + \lambda_0(\tau_j)}.$$

Likewise, h_j, τ_{j-1}, and τ_j determine w_j unless $h_j = 2$ or 4. Otherwise,

$$p(w_j | \boldsymbol{\tau}, h_j = 2) \propto W_2(w_j), \qquad p(w_j | \boldsymbol{\tau}, h_j = 4) \propto W_4(w_j). \tag{7}$$

Both W_2 and W_4 are convolution integrals defined in Appendix B. Numerical integration of W_2 and W_4 is required when drawing $\mathbf{h}$, so the normalizing constants for (7) are known. Each required w_j can be independently sampled by numerical CDF inversion.

3.3 Drawing Model Parameters Given Complete Data

Given complete data, $\lambda_0(t)$, λ_1, γ, and ϕ are independent in their posterior distribution with

$$p(\gamma | \cdot) = \Gamma(a_\gamma + n_{01}, b_\gamma + T_0), \quad p(\phi | \cdot) = \Gamma(a_\phi + n_{10}, b_\phi + T_1), \tag{8}$$
$$p(\lambda_1 | \cdot) = \Gamma(a_1 + n_1, b_1 + T_1).$$

The complete data sufficient statistics in (8) are $T_1 = \int_{t_0}^{T} C(t)\, dt$ and $T_0 = \int_{t_0}^{T} [1 - C(t)]\, dt$, the amount of time the criminal was present/absent; n_{01} and n_{10}, the number of transitions in $C(\cdot)$ from 0 to 1 and from 1 to 0; and n_1, the number of events produced by N_1.

The arbitrary observation window $(t_0, T]$ prevents λ_0, $\boldsymbol{\delta}$, and $\boldsymbol{\eta}$ from being independently sampled. Denote the expected number of events in $(t_0, T]$ as $\Lambda_0^{t_0}(T) = \int_{t_0}^{T} \lambda_0(u)\, du$. The complete data likelihood for λ_0, $\boldsymbol{\delta}$, $\boldsymbol{\eta}$ is

$$L = \exp\big(-\Lambda_0^{t_0}(T)\big) \prod_j \lambda_0(\tau_j), \tag{9}$$

where the product is over events produced by N_0. If $(t_0, T]$ covers W complete weeks, (1) and (2) imply $\Lambda_0^{t_0}(T) = 7W\lambda_0 + R(\boldsymbol{\delta}, \boldsymbol{\eta}, t_0, T)$. Here $R(\cdot)$ is the contribution to Λ_0 from incomplete weeks at the ends of the observation window, so $R(\cdot) = 0$ if $T - t_0$ is an integer number of weeks. The product in (9) factors as:

$$\prod_j \lambda_0(\tau_j) = \lambda_0^{n_0} \prod_{d=0}^{6} \delta_d^{n_d} \prod_{h=0}^{23} \eta_{D(d)h}^{n_{dh}}, \tag{10}$$

where n_d is the number of events occurring on day d, and n_{dh} is the number of events taking place during hour h of day d. Thus, the full conditional for λ_0 is

$$p(\lambda_0|\cdot) = \Gamma(a_0 + n_0, b_0 + B^*), \tag{11}$$

where $B^* \equiv \Lambda_0^{t_0}(T)/\lambda_0$ is the operational time (Cox and Isham, 1980) the process has been active. As the length of $(t_0, T]$ increases the relative effect of $(\boldsymbol{\delta}, \boldsymbol{\eta})$ on B^* vanishes, implying λ_0 is asymptotically independent of $(\boldsymbol{\delta}, \boldsymbol{\eta})$.

The full conditionals for $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are

$$p(\boldsymbol{\delta}|\cdot) \propto \exp\left[-R(\boldsymbol{\delta}, \boldsymbol{\eta}, t_0, T)\right] \mathcal{D}^7(\boldsymbol{\nu}_d + \mathbf{n}_d), \tag{12}$$

$$p(\boldsymbol{\eta}|\cdot) \propto \exp\left[-R(\boldsymbol{\delta}, \boldsymbol{\eta}, t_0, T)\right] \prod_D \mathcal{D}^{24}(\boldsymbol{\nu}_{Dh} + \mathbf{n}_{Dh}), \tag{13}$$

where $\mathbf{n}_d = (n_d)$ and $\mathbf{n}_{Dh} = (\sum_{D(d)=D} n_{dh})$. The Dirichlet densities in (12) and (13) provide natural jumping kernels for the Metropolis-Hastings algorithm, while $R(\cdot)$ determines the probability of accepting a Hastings candidate. The role of $R(\cdot)$ is to downweight elements of $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ from partially observed intervals at the ends of $(t_0, T]$. For example, if $(t_0, T]$ contains more Tuesdays than Wednesdays, and if Tuesdays and Wednesdays produce the same number of events, then the "Wednesday effect" is larger than the "Tuesday effect," because Wednesday produced its events in a shorter time period.

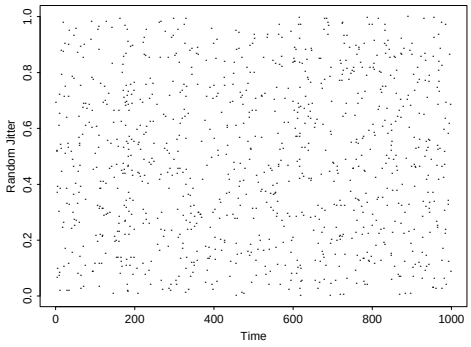


Figure 4: Jittered dotplot showing event times simulated from a Markov modulated (homogeneous) Poisson process.

4 A Simulation Study

The posterior sampling algorithm was tested on data generated from a homogeneous MMPP over the interval $(0, 1000]$. The parameters used in the simulation were $\lambda_0 = 1.0$, $\lambda_1 = 1.0$, $\gamma = 0.001$, $\phi = 0.1$, and $\delta_d = \eta_{D(d)h} = 1.0$ for all d and h. Figure 4 plots the 1,035 events produced by the simulation, which also produced two contamination episodes. Our prior distribution fixes $p(\lambda_0) = p(\lambda_1) = \Gamma(1, 1)$, corresponding to one day of criminal activity in which we observe no events from either the criminal or the customer. The Dirichlet priors for $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ have unit parameters corresponding to the uniform density. Finally we assume $p(\phi) = \Gamma(1, 1)$ and $p(\gamma) = \Gamma(1, 100)$.

The informative prior on γ, which is equivalent to observing 100 days during which the account was free of contamination, eliminates an identifiability issue associated with the homogeneous MMPP. When fraud is rare, when $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ are uniform, and with no prior information to the contrary, the model cannot distinguish between customer traffic and criminal traffic produced by a succession of very short criminal invasions (Scott, 1999). The prior on γ communicates genuine prior knowledge that criminal intrusion is rare.

Figure 5 shows posterior density estimates for θ given the event times in Figure 4. The density estimates in Figure 5 are from an MCMC run of 1,000 iterations. The sampler converges very rapidly because of the efficient way in which the missing data are drawn. Nearly every boxplot in Figure 5 covers the unit line, which reflects the absence of systematic time effects in Figure 4. The long run event rate λ_0 is precisely estimated. The criminal parameters λ_1, γ, and ϕ are less

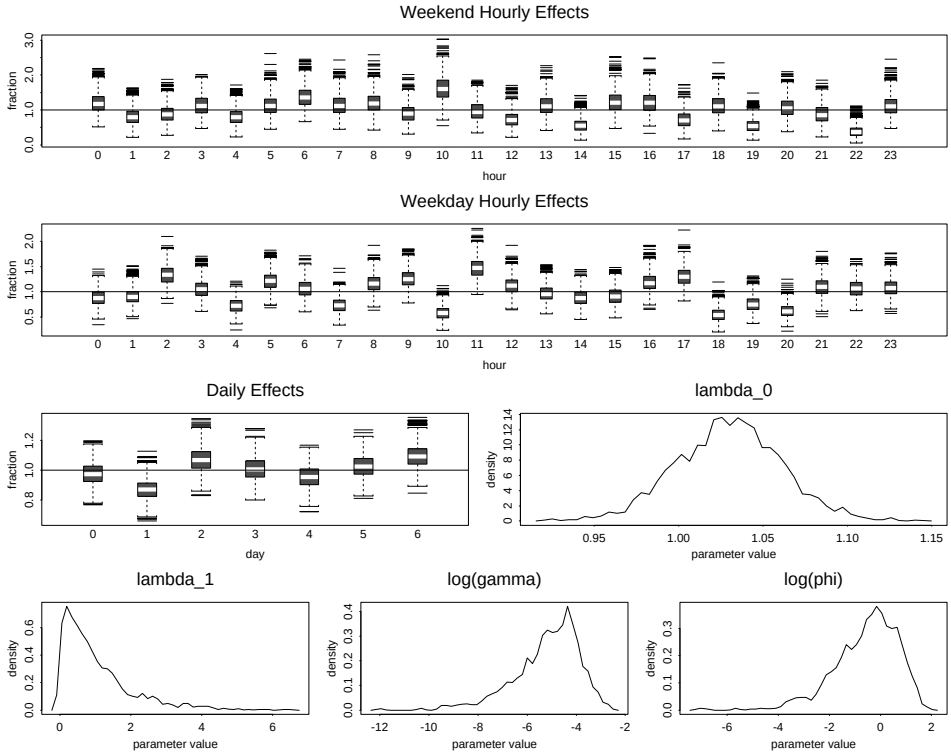
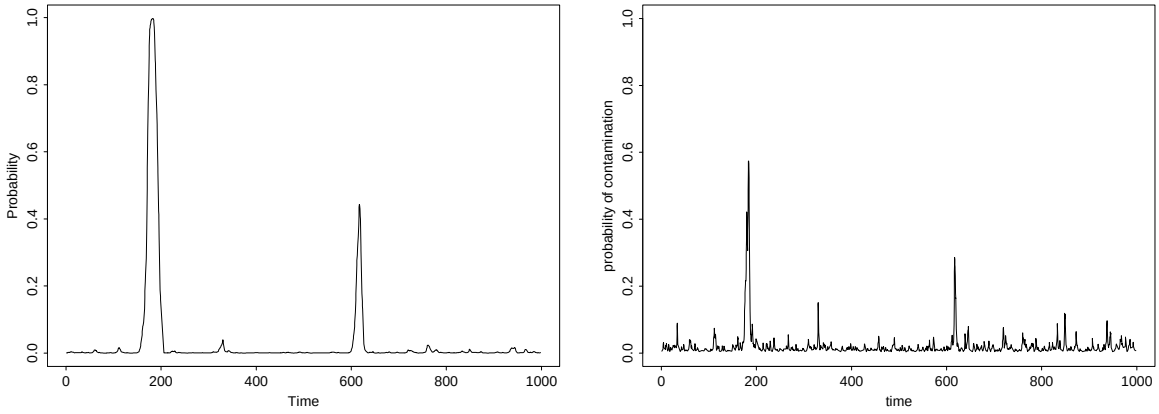


Figure 5: Posterior density estimates of MMNHPP parameters given simulated data from Figure 4. The true parameter values generating the data are $\lambda_0 = 1.0$, $\lambda_1 = 1.0$, $\log\gamma = -6.91$, $\log\phi = -2.30$, and $\delta_d = \eta_{D(d)h} = 1.0$ for all d and h.



(a) (b)

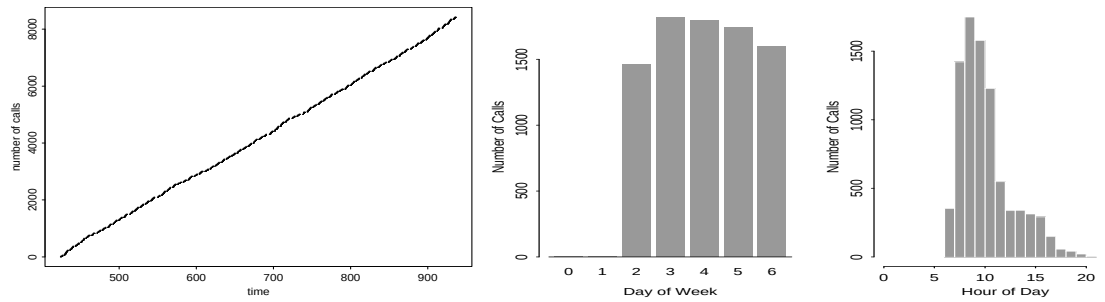
Figure 6: Probability of contamination as a function of time for the simulated data in Figure 4 assuming (a) the true parameter values and (b) parameters estimated from their posterior means in Figure 5.

precisely estimated because the data contain only two relatively short criminal intrusions.

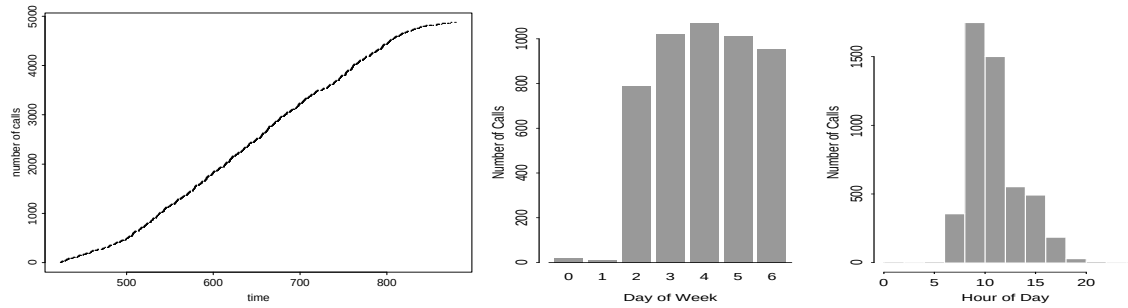
A central quantity in network intrusion detection is the probability of a criminal presence at a given time point, which can be calculated using a non-stochastic version of the forward-backward recursions (Baum *et al.*, 1970). For the data in Figure 4, Figure 6 compares $Pr(C(t) = 1|\boldsymbol{\tau}, \theta)$ given true model parameters to the same calculation with parameters estimated by their posterior means. Despite the relative noise from fitting 53 unnecessary parameters in Figure 6(b), both plots successfully uncover the two contamination intervals from the simulation.

5 International Telephone Traffic

This section analyzes international telephone traffic from two accounts investigated by AT&T fraud experts. Figure 7(a) shows the counting process and daily and hourly calling patterns for an account that did not suffer criminal intrusion. Figure 7(b) shows corresponding summaries for an account investigators found to have been victimized. The counting process in Figure 7(a) is well approximated by a straight line, so the account's long term calling rate is roughly constant. The account rarely generates calls on weekends, before 7 a.m., or after 10 p.m. Heavy traffic before



(a)



(b)

Figure 7: Counting process and daily and hourly patterns for international telephone traffic from (a) an uncontaminated account (b) a contaminated account. Time is measured continuously in days since January 1, 1994. Days of the week are numbered starting with Saturday.

noon suggests the account belongs to a business trading with European partners that must be called early in the day due to time differences. The hourly and daily calling patterns for Figure 7(b) are similar to those in Figure 7(a), although the second account's overall calling rate is slightly less. Again the long run calling rate is roughly constant, though it dips slightly at either end of the observation window.

Figure 8 presents posterior density estimates for θ given the data in Figure 7(a), assuming the prior density from Section 4. The posterior density estimates for $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ reflect the daily and hourly patterns seen in Figure 7(a). The long run calling rate λ_0 is about 16 calls per day, which agrees with a rough calculation from the counting process. The weekend hourly effects $\eta_{we,h}$ are poorly estimated because few calls were made on weekends. Uncertainty about $\eta_{we,h}$ is a self-

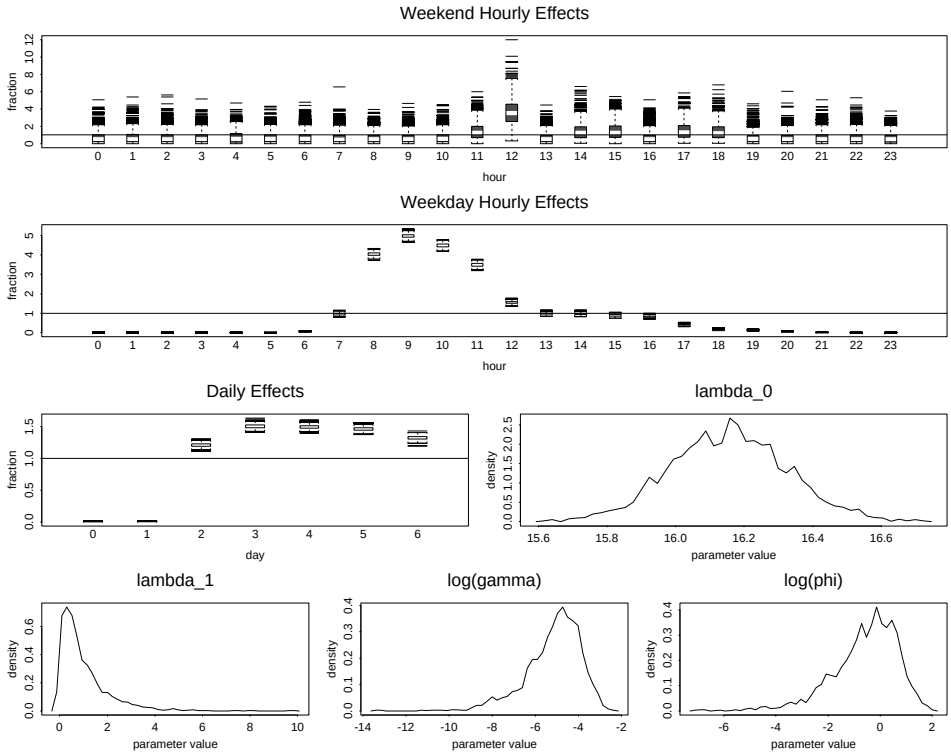
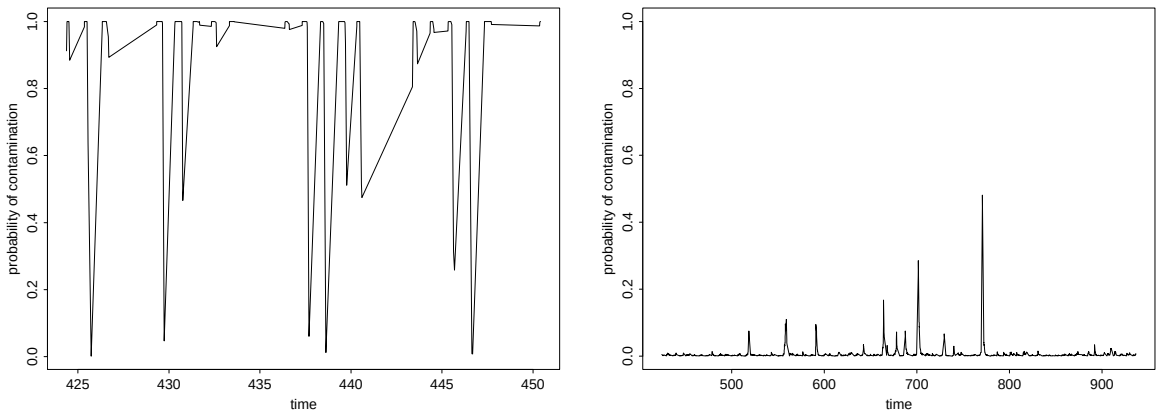


Figure 8: Posterior density estimates of θ based on the uncontaminated account in Figure 7(a).

correcting problem because $\eta_{we,h}$ is multiplied by small values of δ_d, ensuring that $\lambda_0(t)$ is small on the weekends. The posterior distributions for λ_1, γ and ϕ are relatively unchanged from the prior. It is difficult to learn about λ_1, γ, and ϕ from a single account unless the data exhibit an obvious spike in the calling rate. The dependence of λ_1, γ, and ϕ on prior assumptions could be reduced by modeling several accounts with a hierarchical model allowing the accounts to pool information about criminal behavior.

A key feature of the MMNHPP is its ability to distinguish predictable elements of the customer's behavior from truly unusual spikes in the calling rate, a quality not shared by the homogeneous MMPP. Figure 9(a) computes the expected contamination function for the data in Figure 7(a) assuming N_0 is a homogeneous Poisson process. The frequent oscillations between "criminal" presence and absence in Figure 9(a) actually correspond to whether the firm is open or closed for business. Figure 9(b) plots the probability of criminal intrusion when the customer's usual daily



(a) (b)

Figure 9: Probability of criminal activity as a function of time for the telephone account in Figure 7(a) assuming N_0 follows (a) a homogeneous Poisson and (b) a nonhomogeneous Poisson process. Figure 9(a) is shown with a restricted time scale to prevent overplotting. Time is measured in days since January 1, 1994.

and hourly calling patterns are modeled using the MMNHPP. The spikes in Figure 9(b) occur on rare instances when the account generates night or weekend traffic, and they quickly disappear once normal business activity resumes. The MMNHPP highlights deviations from the account's traditional calling pattern, as it is designed to do.

Figure 10 plots the expected contamination function for the contaminated account in Figure 7(b), assuming an MMNHPP with parameters estimated by their posterior means. The spikes in Figure 10 reach higher levels than those in Figure 9(b). More importantly, they cover wider time intervals. The widest spike in Figure 9(b) begins in day 701, a Sunday. The spike persists through eleven calls early Monday morning and falls below 0.10 by 10:40 a.m. Otherwise, no spike in Figure 9(b) covers more than two calls at a height greater than 0.10. By contrast, Figure 10 shows a period of roughly two months in which the expected contamination function consistently exceeds levels of background noise.

Figure 11 examines the expected contamination functions for the two accounts on a smaller time scale, allowing us to locate individual events and to distinguish between weekends and week

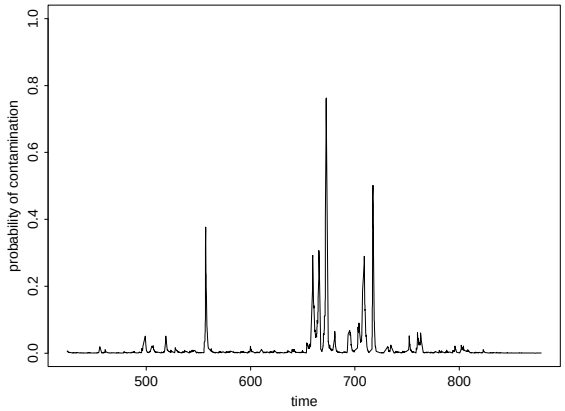


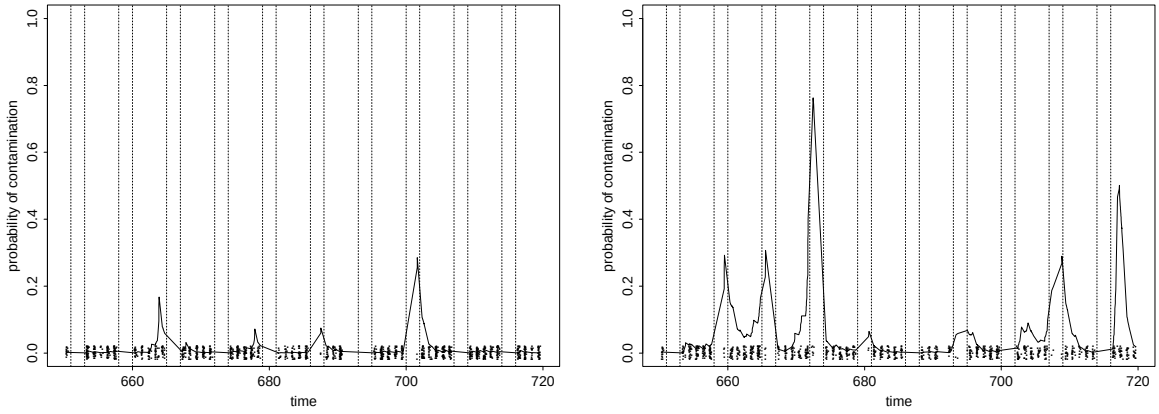
Figure 10: Probability of contamination for the data presented in Figure 7(b) given parameters estimated by their posterior means. Compare to Figure 9(b).

days. Figure 11(a) shows that very few calls in the uncontaminated account took place during periods of suspected criminal activity. The same cannot be said for Figure 11(b). In particular, the weeks beginning on days 660 and 716 both see a large number of calls in periods of risk.

6 Discussion

This article introduces the MMNHPP as a flexible model for point processes exhibiting both regular structure and irregular bursts of activity. We develop an MCMC algorithm for sampling MMNHPP parameters from their posterior distribution given $\boldsymbol{\tau}$. The data augmentation portion of the algorithm is efficient in the sense that it draws the missing quantities $(\mathbf{h}, \mathbf{w}, \mathbf{y})$ directly from their conditional distribution given $(\boldsymbol{\tau}, \theta)$ without relying on unnecessary Gibbs steps.

We conclude by noting four advantages the MMNHPP possesses over more naive intrusion detection methods, such as plotting histograms of event times and looking for spikes, or regressing the number of calls per hour on some function of time and looking for outliers. The most obvious advantage is that the MMNHPP expresses evidence of contamination on the probability scale. Histogram and outlier based methods face ad-hoc decisions about how to interpret the size of an outlier or histogram spike. When paired with information about the cost of each transaction, MMNHPP contamination probabilities translate into quantities like expected dollars-at-risk, which



(a) Honest account from Figure 9(b)

(b) Contaminated account from Figure 10

Figure 11: Closer looks at the contamination predictions for the two telephone accounts. Event times are shown in the jittered rugplot. Vertical dotted lines indicate weekends (00:01 Saturday to 23:59 Sunday).

are easily explained to decision makers.

Second, we have ignored the possibility of transaction level data like the length of a telephone call or the type of item purchased with a credit card. The MMNHPP readily accommodates such data by acting as a mixing distribution in a manner analogous to discrete-time hidden Markov models. The forward-backward recursions used to draw $\mathbf{h}$ immediately adapt to transaction-level data whose distribution obeys (5) and (6). The ability to accommodate covariates allows the MMNHPP to be paired with "scoring" systems many companies have developed to describe how well individual transactions match a customer's historical behavior.

A third advantage is that the MMNHPP rationally combines evidence of contamination over time. Outlier based schemes face difficult questions about how many outliers must be present, and over what period of time, before criminal intrusion is declared.

Finally, the MMNHPP naturally extends to model a network of accounts by embedding individual parameters in a hierarchical model. A hierarchical model can help eliminate the need for informative prior information discussed in Sections 3, 4, and 5. The strong prior on γ can be replaced with observed information from accounts that were investigated and found to be free of criminal contamination.

Appendix A: Forward-Backward Recursions

For vectors $\mathbf{x}$ write x_j^k to mean $(x_j, \ldots, x_k)$, let $P_s(\tau_j|\tau_{j-1}) = p(\tau_j|\tau_{j-1}, h_j = s, \theta)$, and let $q(r, s|\tau_{j-1}) = Pr(h_j = s|h_{j-1} = r, \tau_{j-1}, \theta)$. The forward recursion develops a sequence of matrices $\mathbf{P}_j = (p_{jrs})$, $j = 2, \ldots, n$, $r, s \in \{0, \ldots, 4\}$, representing the joint distribution of (h_{j-1}, h_j) given τ_1^j and θ. Write $\pi_j(s) = Pr(h_j = s|\tau_1^j, \theta)$ for the corresponding marginal density of h_j. Equations (5) and (6) imply

$$\begin{aligned} p_{jrs} &\propto Pr(h_j = s, h_{j-1} = r, \tau_j|\tau_1^{j-1}, \theta) \\ &= P_s(\tau_j|\tau_{j-1}) q(r, s|\tau_{j-1}) \pi_{j-1}(r), \end{aligned} \tag{14}$$

where the proportionality constant is resolved by $\sum_r \sum_s p_{jrs} = 1$. Compute $\pi_j(s) = \sum_r p_{jrs}$ once $\mathbf{P}_j$ is known to set up the next step in the recursion.

Upon completing the forward recursion, factor $p(\mathbf{h}|\boldsymbol{\tau}, \theta)$ as

$$p(\mathbf{h}|\boldsymbol{\tau}, \theta) = p(h_n|\boldsymbol{\tau}, \theta) \prod_{j=1}^{n-1} p(h_{n-j}|h_{n-j+1}^n, \boldsymbol{\tau}, \theta), \tag{15}$$

and observe that for $k > j + 1$

$$\begin{aligned} p(h_j|h_{j+1}^k, \tau_1^k, \theta) &\propto p(h_j^k, \tau_1^k, \theta) \\ &= P_{h_k}(\tau_k|\tau_{k-1}) q(h_{k-1}, h_k|\tau_{k-1}) p(h_j^{k-1}, \tau_1^{k-1}, \theta) \\ &\propto p(h_j|h_{j+1}^{k-1}, \tau_1^{k-1}, \theta). \end{aligned} \tag{16}$$

Conclude from (16) that

$$p(h_j|h_{j+1}^n, \boldsymbol{\tau}, \theta) = p(h_j|h_{j+1}, \tau_1^{j+1}, \theta). \tag{17}$$

Equations (15) and (17) specify the backward recursion. Sample (h_{n-1}, h_n) from $\mathbf{P}_n$, and draw h_j from column h_{j+1} of $\mathbf{P}_{j+1}$ for $j = n - 2, \ldots, 1$ to produce a draw of $\mathbf{h}$ from $p(\mathbf{h}|\boldsymbol{\tau}, \theta)$.

Appendix B: Computation

The recursions in Appendix A hold for general non-stationary hidden Markov models. We can streamline the forward recursion using four facts about the MMNHPP. First, many states do not communicate, so $q(r,s|\tau_{j-1}) = 0$ if $r \in \{0,2\}$ and $s \in \{2,3,4\}$ or if $r \in \{1,3,4\}$ and $s \in \{0,1\}$. Second, all the information in h_{j-1} relevant to h_j is contained in $C(\tau_{j-1})$, so we can write $q(r,s|\tau_{j-1}) = q_s(\tau_{j-1})$, where

$$q_s(\tau_{j-1}) = \begin{cases} Pr(h_j = s|C(\tau_{j-1}) = 0) & s \in \{0,1\} \\ Pr(h_j = s|C(\tau_{j-1}) = 1) & s \in \{2,3,4\}. \end{cases}$$

Each $q_s(\tau_{j-1})$ represents the probability the first event after τ_{j-1} is produced by the Poisson process corresponding to $h_j = s$ (Cox and Isham, 1980). For example, $q_0(\tau_{j-1})$ is the probability that the first event after τ_{j-1} is produced by N_0 rather than $C(\cdot)$. Third, the forward recursion only requires the products $q_s(\tau_{j-1})P_s(\tau_j|\tau_{j-1})$, rather than individual transition probabilities and conditional densities. Fourth, the forward recursion only determines $\mathbf{P}_j$ up to proportionality, so $q_s(\tau_{j-1})P_s(\tau_j|\tau_{j-1})$ may be rescaled without altering the algorithm. A convenient scaling divides each q_sP_s by $\exp\big(-\Lambda_0^{\tau_{j-1}}(\tau_j) - \gamma(\tau_j - \tau_{j-1})\big)$. The required quantities simplify to

$$\begin{aligned} q_0(\tau_{j-1})P_0(\tau_j|\tau_{j-1}) &= \lambda_0(\tau_j) \\ q_1(\tau_{j-1})P_1(\tau_j|\tau_{j-1}) &= \gamma \\ q_2(\tau_{j-1})P_2(\tau_j|\tau_{j-1}) &= \frac{\phi q_2(\tau_{j-1})\lambda_0(\tau_j)}{1 - q_3(\tau_{j-1})} \int_{\tau_{j-1}}^{\tau_j} W_2(u)\, du \\ q_3(\tau_{j-1})P_3(\tau_j|\tau_{j-1}) &= (\lambda_0(\tau_j) + \lambda_1)\exp\big(-(\lambda_1 + \phi - \gamma)(\tau_j - \tau_{j-1})\big) \\ q_4(\tau_{j-1})P_4(\tau_j|\tau_{j-1}) &= \frac{\phi q_4(\tau_{j-1})}{1 - q_3(\tau_{j-1})} \int_{\tau_{j-1}}^{\tau_j} W_4(u)\, du. \end{aligned} \tag{18}$$

The remainder of this section defines q_2, q_3, q_4, $W_2(u)$, and $W_4(u)$. If $h_j = 2$ then $C(\cdot)$ jumps from 1 to 0 in the interior of I_j, producing a convolution of the form $q_2(\tau_{j-1}) = \int_{\tau_{j-1}}^{\infty} Pr_{\tau_{j-1}}(1 \to 0 \text{ at time } t) Pr_t(0 \to 0)\, dt$.

Specifically,

$$q_2(\tau_{j-1}) = \int_{\tau_{j-1}}^{\infty} \phi \exp\big(- \Lambda_0^{\tau_{j-1}}(t) - (\lambda_1 + \phi)(t - \tau_{j-1})\big) \int_t^{\infty} \lambda_0(u) \exp\big(- \Lambda_0^t(u) - \gamma(u - t)\big)\, du\, dt. \quad (19)$$

Because $q_2 + q_3 + q_4 = 1$, one need not evaluate q_4 if the simpler q_3 is obtained instead.

$$q_3(\tau_{j-1}) = \int_{\tau_{j-1}}^{\infty} (\lambda_0(t) + \lambda_1) \exp\big(- \Lambda_0^{\tau_{j-1}}(t) - (\lambda_1 + \phi)(t - \tau_{j-1})\big)\, dt. \quad (20)$$

Both (19) and (20) have closed form solutions if $\lambda_0(t)$ obeys (1). See Scott (1998) for details. If $s = 2$ or 4 then $P_s(\tau_j|\tau_{j-1})$ is a convolution of the form $\int_{\tau_{j-1}}^{\tau_j} p_{\tau_{j-1}}(w_j) p_{w_j}(\tau_j)\, dw_j$, where $\tau_{j-1} + w_j$ is the time of a transition of $C(\cdot)$ from 1 to 0. For $s = 2$,

$$\begin{aligned} P_2(\tau_j|\tau_{j-1}) &= \int_{\tau_{j-1}}^{\tau_j} \frac{\phi \exp\big(- \Lambda_0^{\tau_{j-1}}(u) - (\lambda_1 + \phi)(u - \tau_{j-1})\big)}{1 - q_3(\tau_{j-1})} \\ &\qquad \frac{\lambda_0(\tau_j) \exp\big(- \Lambda_0^u(\tau_j) - \gamma(\tau_j - u)\big)}{\int_u^{\infty} \gamma \exp\big(- \Lambda_0^u(t) - \gamma(t - u)\big)\, dt}\, du \\ &= \frac{\lambda_0(\tau_j) \phi \exp\big(- \Lambda_0^{\tau_{j-1}}(\tau_j) - \gamma(\tau_j - \tau_{j-1})\big)}{1 - q_3(\tau_{j-1})} \int_{\tau_{j-1}}^{\tau_j} W_2(u)\, du, \end{aligned} \quad (21)$$

where

$$W_2(u) = \frac{\exp\big(- (\lambda_1 + \phi - \gamma)(u - \tau_{j-1})\big)}{\int_u^{\infty} \lambda_0(t) \exp\big(- \Lambda_0^u(t) - \gamma(t - u)\big)\, dt}. \quad (22)$$

Equation (21) relies on the substitution

$$\frac{1 - q_3(\tau_{j-1})}{\phi} = \int_{\tau_{j-1}}^{\infty} \exp\big(- \Lambda_0^{\tau_{j-1}}(t) - (\lambda_1 + \phi)(t - \tau_{j-1})\big)\, dt.$$

Similarly,

$$P_4(\tau_j|\tau_{j-1}) = \frac{\phi \exp\big(- \Lambda_0^{\tau_{j-1}}(\tau_j) - \gamma(\tau_j - \tau_{j-1})\big)}{1 - q_3(\tau_{j-1})} \int_{\tau_{j-1}}^{\tau_j} W_4(u)\, du.$$

where

$$W_4(u) = \frac{\exp\big(-(\lambda_1 + \phi - \gamma)(u - \tau_{j-1})\big)}{\int_u^\infty \exp\big(-\Lambda_0^u(t) - \gamma(t-u)\big)\, dt}.$$

Evaluating $\int_{\tau_{j-1}}^{\tau_j} W_2(u)\, du$ and $\int_{\tau_{j-1}}^{\tau_j} W_4(u)\, du$ requires numerical integration.

References

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* **41**, 164–171.

Besag, J. and Green, P. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 1, 25–37.

Byers, S. and Raftery, A. E. (1998). Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association* **93**, 422, 577–584.

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

Damian, P., Wakefield, J., and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society, Series B, Methodological* **61**, 331–344.

Davison, A. C. and Ramesh, N. I. (1996). Some models for discretized series of events. *Journal of the American Statistical Association* **91**, 601–609.

Du, Q. (1995). A monotonicity result for a single-server queue subject to a Markov-modulated Poisson process. *Journal of Applied Probability* **32**, 1103–1111.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Guo, W., Wang, Y., and Brown, M. B. (1999). A signal extraction approach to modeling hormone time series with pulses and a changing baseline. *Journal of the American Statistical Association* **94**, 447, 746–756.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Higdon, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association* **93**, 422, 585–595.

Kolaczyk, E. D. (1999). Bayesian multi-scale models for Poisson processes. *Journal of the American Statistical Association* **94**, 447, 920–933.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series.* Chapman & Hall.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Olivier, C. and Walrand, J. (1994). On the existence of finite-dimensional filters for Markov-modulated traffic. *Journal of Applied Probability* **31**, 515–525.

Scott, S. L. (1998). *Bayesian Methods and Extensions to the Two State Markov Modulated Poisson Process.* Ph.D. thesis, Harvard University.

Scott, S. L. (1999). Bayesian analysis of a two state Markov modulated Poisson process. *Journal of Computational and Graphical Statistics* **8**, 3, 662–670.

Scott, S. L. (2000). Bayesian methods for hidden Markov models. *Journal of the American Statistical Association* (under second review).

Turin, W. (1996). Fitting probabilistic automata via the EM algorithm. *Communications in Statistics – Stochastic Models* **12**, 405–424.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley.