Modeling Correlated Arrival Events with Latent Semi-Markov Processes

Wenzhao Lian WL89@DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA

Vinayak Rao VAR11@STAT.DUKE.EDU
Department of Statistical Science, Duke University, Durham, NC 27708, USA

Brian Eriksson BRIAN.ERIKSSON@TECHNICOLOR.COM
Technicolor Research Center, 735 Emerson Street, Palo Alto, CA 94301, USA

Lawrence Carin LCARIN@DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA

Abstract

The analysis of correlated point process data has wide applications, ranging from biomedical research to network analysis. In this work, we model such data as generated by a latent collection of continuous-time binary semi-Markov processes, corresponding to external events appearing and disappearing. A continuous-time modeling framework is more appropriate for multichannel point process data than a binning approach requiring time discretization, and we show connections between our model and recent ideas from the discrete-time literature. We describe an efficient MCMC algorithm for posterior inference, and apply our ideas to both synthetic data and a real-world biometrics application.

1. Introduction

The study and development of complex dynamical systems has led to the availability of increasingly large datasets, recording events evolving asynchronously and at multiple time-scales. Modeling such data by discretizing time at the resolution of the fastest events is inefficient and inelegant, especially when the relevant time-scales themselves must be inferred. It is much more natural to work directly in continuous-time, and there has been a growth of such applications in the statistics and machine learning communities (*e.g.,* (Nodelman et al., 2002; Scott & Smyth, 2003; Golightly & Wilkinson, 2011; Saeedi & Bouchard-Côté, 2011; Teh et al., 2011)). Nevertheless, the scope of this prior research is limited and much work remains to apply continuous-time models to high-dimensional problems with interesting structure in the latent states and dynamics of the latent process.

The specific application motivating the methodology of this paper is the analysis of biometric galvanic skin response (GSR) data. User arousal (or excitation) events translate to the generation of specific observed GSR waveforms (Silveira et al., 2013). Here, we consider GSR observations taken from a number of users exposed simultaneously to a common video stimulus. The ability to accurately characterize the latent stimulus events generating the observed biometric excitation reactions has applications in the areas of recommendation, market research, advertising, etc. More generally, our ideas are also applicable to point process data from other domains, like neuroscience (*e.g.,* (Yu et al., 2009)), biometrics (*e.g.,* (Barbieri et al., 2005)), network data analysis (*e.g.,* (Ryu & Lowen, 1996)), and ecology (*e.g.,* (Ogata & Tanemura, 1985)).

In all these applications, one observes point process data exhibiting significant variability, or *inhomogeneity*, over time. In the context of GSR observations from movie stimulus, an intense action scene may have many GSR arousal reactions (*i.e.,* arrivals) in a short time, while a quiet dialog scene may elicit very few arrivals over a potentially long interval. In other applications, the rate of webpage requests can vary with external events, while the spiking of a neuron can vary with stimulus. Without explicit knowledge of the external events, we look to extract the latent structure from the observed point process data.

For a single observed point process, an appropriate modeling framework is that of Markov-modulated Poisson processes (MMPP) (*e.g.,* (Scott & Smyth, 2003)). This is a

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

doubly-stochastic Poisson process, where the unknown rate is modeled as a realization of a continuous-time Markov jump process (MJP). In our case, we observe a number of correlated inhomogeneous Poisson processes, which we couple via a common low-dimensional underlying process. Specifically, we assume that the arrival rates are described by a small number of binary switching signals, indicating the presence or absence of a relevant external event. Examples for biometric point processes from movie stimulus could include action scenes, dramatic dialog, comedic events, or a main character appearing. We model this as a collection of continuous-time binary signals linked by a factor-loading matrix to the intensities of the observed processes. Rather than risking under-fitting with a small number of factors, we allow the model to choose from a large number of relevant sources by placing a shrinkage prior on a source loading matrix. Furthermore, rather than modeling the binary sources as memoryless (as in MMPPs), we allow more flexibility in the dynamics general hazard functions. This mirrors recent work in the discrete-time literature (*e.g.,* (Fox et al., 2011; Johnson & Willsky, 2013)) to model state persistence and more flexible distributions over state durations.

We evaluate the performance of our model and inference methodology on both synthetic and real-world biometrics data. For the latter, we apply our methodology to a biometric galvanic skin response dataset taken from users viewing a full feature-length film. We find that the resolved latent structure correlates with explicit feedback (*i.e.,* user ratings) in terms of both excitement during specific scenes in the film and the users' overall impression of the film.

2. Problem and Model formulation

We wish to model a collection of U sequences of events over an interval $[0, T]$. For "user" $u \in \{1, \cdots, U\}$, we denote the set of arrival times as $\{y_{u,j}\}$, with $y_{u,j}$ being the time stamp of the j-th event in stream u. Each sequence $y_{u,\cdot}$ is an inhomogeneous Poisson process with instantaneous rate $\gamma_u(t)$. The latter is expressed as a user-specific base Poisson rate λ_u modulated by a stimulus-determined function over time. The rates of these U Poisson processes share information through K binary latent sources $s_k(t)$ with $k \in \{1, \cdots, K\}$. Below, we use $s_\cdot(t)$ to denote at time t, the column vector consisting of all K sources $s_k(t)$. Calling the loading matrix $\boldsymbol{W}$ (with $\boldsymbol{w}_u \in \mathbb{R}^K$, a row vector specific to user u), we have

$$y_{u,\cdot} \quad \sim \quad \mathcal{PP}(\gamma_u(\cdot)), u = 1, \cdots, U \qquad (1)$$

$$\gamma_u(t) \quad = \quad \lambda_u \text{exp}(\boldsymbol{w}_u s_\cdot(t)), t \in [0, T] \qquad (2)$$

Here, λ_u represents the baseline arrival rate, while the elements of $\boldsymbol{w}_u$ indicate how relevant each of the K sources is to user u. Our goal is to estimate both these user specific parameters as well as the latent sources $s_k(t)$ from the set of arrival times $\{y_{u,j}\}$.

2.1. Binary semi-Markov Jump processes

In our application, we model each of the K latent sources as a binary signal, switching on and off depending on whether the associated external characteristic is active. While it is simplest to model the latent functions, $s_k(t)$, as Markov jump processes (MJPs), the resulting memoryless property can be inappropriate, allowing unnecessary switching between states. Thus, we model these as binary semi-Markov Jump Processes (bsMJP) (Feller, 1964). Realizations of a bsMJP are right-continuous, piecewise constant binary functions where, unlike an MJP, the rate of state transitions vary with the time since the last transition. This is formalized by a hazard function $h^{(0)}(\nu)$, giving the rate of transition from state 0-to-1, ν time units after entering state 0 ($h^{(1)}(\nu)$ is similarly defined, and we do not allow for self-transitions). For an MJP, $h^{(s)}(\nu)$ is a constant, resulting in a memoryless property; in contrast, by making $h^{(s)}(\nu)$ take small values for small values of ν, we can discourage the system from leaving a state it has just entered. In our applications, we assume h belongs to the Weibull family, with

$$h_k^{(s)}(\nu) \quad = \quad \frac{\beta_k^{(s)}}{\mu_k^{(s)}} \left(\frac{\nu}{\mu_k^{(s)}}\right)^{\beta_k^{(s)}-1} \qquad (3)$$

This corresponds to the the interval length between two state transitions following a Weibull distribution

$$Pr(\nu|\beta_k^{(s)}, \mu_k^{(s)}) = \exp(-\left(\frac{\nu}{\mu_k^{(s)}}\right)^{\beta_k^{(s)}})\frac{\beta_k^{(s)}}{\mu_k^{(s)}}\left(\frac{\nu}{\mu_k^{(s)}}\right)^{\beta_k^{(s)}-1}$$

$\beta_k^{(s)}$ and $\mu_k^{(s)}$ are the shape and scale parameters of the Weibull distribution for state s. Setting $\beta_k^{(s)}$ to 1 recovers the exponential distribution (and our switching process reduces to an MJP).

The hazard functions govern the dynamics of the bsMJP; we also need an initial distribution $\boldsymbol{\pi}_k$ over states. Then, we have

$$s_k(\cdot) \quad \sim \quad \text{bsMJP}(\boldsymbol{\pi}_k, h_k^{(1)}(\cdot), h_k^{(0)}(\cdot)) \qquad (4)$$

2.2. Number of Latent Sources

In most applications, the number of latent sources is unknown, and must be inferred. It is desirable to place a prior on this quantity, and try to infer it from the data. A more flexible approach is a nonparametric solution: allow for an *infinite* number of potential sources, with each user influenced by a finite, and unknown, subset of them. This corresponds to a binary matrix with infinite columns, and with element (u, k) indicating whether or not the bsMJP k is relevant to user u. A popular prior on the resulting binary association matrix is the Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2006), however, despite its elegance, inference of the IBP is complicated and the resulting Markov chain Monte Carlo algorithm mixes poorly as it explores the combinatorial space of binary matrices. This raises a need for alternate approaches to flexibly model the unobserved latent structure.

Here, inspired by Bayesian shrinkage estimation for sparse regression problems (Polson & Scott, 2012), we control complexity of the factor loading matrix using a multiplicative gamma process (MGP) shrinkage prior (Bhattacharya & Dunson, 2011). Rather than placing a spike-and-slab prior on the columns of $\boldsymbol{W}$, we model each row vector of the factor loading matrix ($\boldsymbol{w}_u \in \mathbb{R}^K$) as a collection of Gaussian variables increasingly concentrated around the origin. Specifically,

$$\begin{aligned} \boldsymbol{w}_u &\sim \mathcal{N}(0, \Lambda^{-1}), & \Lambda &= \mathrm{diag}(\tau_1, \tau_2, \cdots) \\ \tau_k &= \prod_{l=1}^{k} \xi_l, & \xi_l &\sim \mathrm{Ga}(\alpha, 1) \end{aligned} \tag{5}$$

By choosing $\alpha > 1$, the $\{\xi_l\}$ are on average larger that 1, encouraging τ_k to increase with k. This in turn causes the amplitudes of the $\boldsymbol{w}_{uk}$ to shrink close to (while not exactly equaling) 0. The construction results in stochastically ordering the latent sources, with early components having large variance and corresponding to sources relevant to most users. As k increases, these weights are shrunk more and more strongly towards zero, while still allowing a few to escape this pull (allowing us to model sources specific to certain users). Thus, we potentially obtain an infinite number of latent sources $\{s_k(t)\}$, stochastically ranked by their contribution to the rate function $\{\gamma_u(t)\}$. In practice, we truncate at a sufficiently large K.

2.3. Miscellaneous variables

Finally, we describe prior distributions over the remaining variables. We model the base Poisson rate λ_u of each user as independently drawn from $\mathrm{Ga}(c, d)$. We construct conjugate hyperpriors for $\beta_k^{(s)}$ and $\mu_k^{(s)}$ after a variable transformation outlined in Section 4.

3. Related Work

Early work analyzing arrival data did not exploit correlation across multiple arrival streams (Daley & Vere-Jones, 1998; Kass & Ventura, 2001; Riehle et al., 1997). The work in (Smith & Brown, 2003) introduces a single first-order autoregressive latent state model with a proposed EM-algorithm to estimate parameter values from multi-user arrival data. A more complex hidden Markov Model (HMM) approach using multiple latent states is presented in (Escola et al., 2011). In contrast to the work in this paper, the time evolving stimulus has a simple structure and is explicitly known.

Perhaps most similar to the work in this paper is the Gaussian process factor analysis (GPFA) approach of (Yu et al., 2009). They modeled the intensities as a series of correlated neural spike trains by linearly transforming a small set of latent Gaussian processes. By contrast, the binary switching signals in our model captures specific aspects of the latent structure, and our shrinkage approach allows us to infer the number of sources. Finally, the simpler structure of the bsMJP means our model is scalable to longer observation intervals. Inference in (Yu et al., 2009) required a variational approximation to the complicated posterior, while no such approximations are needed here.

Our modeling also relates to previous work on finite state systems in discrete and continuous time. We generalize the Markov-modulated Poisson process (MMPP) (Scott & Smyth, 2003) by allowing correlated Poisson intensities, and by allowing more flexible (*i.e.,* non-exponential) holding times. This latter point of semi-Markovianity has been a topic of recent interest in the discrete time-series modeling community (Fox et al., 2011; Johnson & Willsky, 2013), although it fits more naturally in the continuous-time setting. While we took a shrinkage approach to coupling the latent sources, our ideas easily extend to genuine, truncation-free nonparametric models based on the IBP. Such ideas are related to the infinite factorial HMM (Van Gael et al., 2009), a discrete-time Markov model with infinite, independent latent sources, and provide alternatives to the Dirichlet process-based infinite state MJP models developed in the continuous-time literature (Teh et al., 2011; Saeedi & Bouchard-Côté, 2011).

4. MCMC for bsMJPs

A challenge for more widespread application of continuous-time models is the problem of posterior inference. For our model, the central problem reduces to sampling from the posterior over bsMJP paths given the Poisson arrival times, $\{y_{u,j}\}$. We are aware of two general approaches to MCMC sampling for such models: the particle MCMC approach of (Andrieu et al., 2010; Golightly & Wilkinson, 2011) and a thinning-based approach of (Rao & Teh, 2012; 2013). Here, we adapt the latter to our problem.

Observe that a sample path s_k of the bsMJP is entirely determined by the set of transition times $\phi_k = \{\phi_{k,1}, \cdots, \phi_{k,n_k}\}$, and the states evaluated at these times $\{s_k(t), t \in \phi_k\}$ (for the binary sMJP, the latter are redundant given the initial state). Also, recall that ν time-units after entering state 0, the rate of transitioning from state 0 to 1 is given by $h_k^{(0)}(\nu)$ (respectively, $h_k^{(1)}(\nu)$ for state 1). Typically self-transitions are not allowed, and have rate equal to zero.

We now define an equivalent continuous-time system but with self-transitions, occurring with constant rates $\Omega_k^{(0)}$ and $\Omega_k^{(1)}$ for states 0 and 1. These self-transitions will serve as auxiliary variables facilitating the easy resampling of new state values. Our approach simplifies (Rao & Teh, 2012; 2013), where a multiplicative bound on the transition rates resulted in self-transition rates depending on the current holding time.

For state s, candidate transition times (whether self-transitions or not) are now drawn from a hazard function $H_k^{(s)}(\nu) = h_k^{(s)}(\nu) + \Omega_k^{(s)}$. We sample these candidate events sequentially. Thus, assume l time units have passed since the last state transition (when we entered state s). Then, we sample the next candidate event-time from $H_k^{(s)}(\cdot)$, conditioning on it being larger than l. Assuming this is $l + \Delta$, we advance time by Δ, and assign an actual transition out of state s with probability $\frac{h_k^{(s)}(l+\Delta)}{H_k^{(s)}(l+\Delta)}$ (otherwise this event is treated as a self-transition). After updating l, we repeat this process until the current time exceeds T. Algorithm 1 in the appendix gives details of this generative process. It is easy to show that discarding the self-state transitions corresponds to a draw from the original bsMJP.

We use the construction above to define a Markov operator over bsMJP paths, with the desired posterior as its stationary distribution. Denote the set of self-transitions as $\tilde{\phi}_k$, with the set of actual transitions given by ϕ_k. From our construction of the previous paragraph, it follows that given the current bsMJP trajectory s_k, the set $\tilde{\phi}_k$ is conditionally an inhomogeneous Poisson with piecewise-constant rate $\Omega_k^{(s_k(t))}(t)$. Thus, when the system is in state s, the Poisson rate is $\Omega_k^{(s)}$, and we can easily reconstruct $\tilde{\phi}_k$.

Denote all the candidate times as $\Phi_k = \phi_k \cup \tilde{\phi}_k$. Having introduced $\tilde{\phi}_k$, we sample a new path, now restricting ourselves to paths that change state at some subset of Φ_k (rather than searching over all paths with all possible transition times). Consequently, sampling a new path conditioned on Φ_k amounts to reassigning labels "transition" or "self-transition" to each of the elements of Φ_k. Our problem now reduces to sampling a trajectory of discrete-time model, and can be carried out efficiently using dynamic programming algorithms such as forward filtering backward sampling (FFBS) (Früwirth-Schnatter, 1994). Note that this step accounts for the Poisson observations. In particular, given two successive candidate transition times, t_1 and t_2 in Φ_k, the likelihood the system remains in state s over this interval equals the probability of the subset of Poisson observations $\{y_{u,j}\}$ falling in $[t_1, t_2)$ under state s. This follows a Poisson distribution and is easy to calculate.

Overall, sampling proceeds by alternately sampling a set of thinned events $\tilde{\phi}_k$ given the current trajectory, and then a new trajectory given $\Phi_k = \phi_k \cup \tilde{\phi}_k$. We leave the details of this for the appendix.

5. Model inference

We now describe the overall Markov Chain Monte Carlo (MCMC) algorithm. We wish to infer the posterior distribution over the variables $\{\{\boldsymbol{w}_u\}, \{s_k(\cdot)\}, \{\lambda_u\}, \{\xi_k\}, \{\beta_k^{(s)}\}, \{\mu_k^{(s)}\}\}$. Our algorithm is a Gibbs sampler, sampling each variable conditioned on the remaining variables. For ease of notation, we use $p(\cdot| \sim)$ to denote the posterior distribution of one variable conditioning on all other variables.

Inference of Latent Sources ($s_k(\cdot)$): We cycle through all K sources, resampling each bsMJP path s_k conditioned on all other variables. To do so, we use the algorithm of the previous section: for each source, conditionally introduce the Poisson distributed thinned events $\tilde{\phi}_k$, and then conditionally resample the new trajectory using the FFBS algorithm.

Inference of Factor Loadings ($\boldsymbol{w}_u$): The Gaussian prior on $\boldsymbol{w}_u$ and the nonlinear likelihood in Equation (2) result in a nonstandard posterior distribution. While a simple Metropolis-Hastings (MH) approach is to perturb the current value $\boldsymbol{w}_u$, proposing a new value from a normal variable drawn with mean centered at $\boldsymbol{w}_u$, we instead use the following proposal distribution tailored to our problem.

Noting that $p(\boldsymbol{w}_u| \sim)$ is log-concave, we use Newton's method to find the MAP $\hat{\boldsymbol{w}}_u$ with gradient and Hessian

$$g(\boldsymbol{w}_u) = \sum_{j=1}^{N_u} s.(y_{u,j}) - \int_0^T \exp(\boldsymbol{w}_u s.(t)) \lambda_u s.(t) dt - \Lambda \boldsymbol{w}_u^T$$
$$H(\boldsymbol{w}_u) = -\int_0^T \exp(\boldsymbol{w}_u s.(t)) \lambda_u s.(t) s.(t)^T dt - \Lambda$$

respectively. Here, N_u is the number of arrivals observed for user u. Because latent sources, $s_k(t)$, are simple binary functions, the integrals above reduce to finite summations and are easily computable. Thus, at each iteration of the sampler, we propose $\boldsymbol{w}_u^* \sim \mathcal{N}(\hat{\boldsymbol{w}}_u, qH^{-1}(\hat{\boldsymbol{w}}_u))$, where q is a tuning parameter. The new proposal $\boldsymbol{w}_u^*$ is accepted with probability

$$\min\{1, \frac{Pr(\boldsymbol{w}_u^*| \sim)\mathcal{N}(\boldsymbol{w}_u^{(old)}; \hat{\boldsymbol{w}}_u, qH^{-1}(\hat{\boldsymbol{w}}_u))}{Pr(\boldsymbol{w}_u^{(old)}| \sim)\mathcal{N}(\boldsymbol{w}_u^*; \hat{\boldsymbol{w}}_u, qH^{-1}(\hat{\boldsymbol{w}}_u))}\}$$

We set q to 5 in our experiments, and the high acceptance rates suggest this is an efficient approach.

Hyperparameter inference ($\beta_k^{(s)}$, $\mu_k^{(s)}$, ξ_k):
By applying the following transformation: $\beta' = \beta, \mu' =$

$\mu^{-\beta}$, we obtain a new representation of the Weibull distribution with p.d.f. and c.d.f. :

$$f(\nu;\beta',\mu') = \beta'\mu'\nu^{\beta'-1}\text{exp}(-\mu'\nu^{\beta'}) \tag{6}$$

$$F(\nu;\beta',\mu') = 1-\text{exp}(-\mu'\nu^{\beta'}) \tag{7}$$

Using these representations, a gamma prior on μ' is the conjugate, and with hyperparameters (e, f), we can directly sample μ' from the posterior,

$$p(\mu_k'^{(s)}|\sim) \propto \text{Ga}(\mu_k'^{(s)}; e+\sum_i \mathbb{I}(\phi_{k,i}=s), f+\sum_{i:s_k(\phi_{k,i})=s} \Delta_{k,i+1}^{\beta_k'^{(s)}})$$

Here $\Delta_{k,i+1} = \phi_{k,i+1} - \phi_{k,i}$.

The components of the MGP shrinkage prior, $\{\xi_k\}$, can be directly sampled from the conditional posterior:

$$p(\xi_k|\sim) \propto \text{Ga}(\alpha+\frac{U(K-k+1)}{2}, 1+\frac{1}{2}\sum_{u=1}^{U}\sum_{l=k}^{K}\boldsymbol{w}_l^2\prod_{m=1,m\neq k}^{l}\xi_m)$$

We place a uniform prior for $\beta_k'^{(s)}$ on a set of discretized grid points and sample from the candidates according to their posterior mass. Other variables, including $\{\lambda_u\}$, are straightfoward to sample.

6. Experiments

We consider two synthetic datasets of arrival data to study our model and sampler, before considering a real-world biometric application. For the latter, we assess the model's ability to capture and explain observed explicit user feedback (*i.e.,* user ratings). As a baseline, we used the kernel methods analyzed in (Zhang & Kou, 2010) to estimate the inhomogeneous Poisson rates, and thus the underlying model parameters (*e.g.,* instantaneous rate, factor loadings, etc.). We found that with a carefully chosen kernel width, the kernel shape does not matter much, and for simplicity, we used a uniform kernel (giving a time-discretized binning method). An oracle choice of the best bin size was used after testing multiple sizes.

6.1. Synthetic Experiments

For our first dataset, we generated $K = 3$ bsMJP paths over an interval of length T with initial state distribution $\boldsymbol{\pi}_k = [0.5, 0.5]^T$ and Weibull hazard rates. Weibull shape $\beta_k^{(s)}$ and scale parameters $\mu_k^{(s)}$ were uniformly drawn from $[1, 5]$ and $[50, 100]$, while each row of the loading matrix $\boldsymbol{W}$ was independently generated from a standard normal, $\mathcal{N}(\boldsymbol{0}_K, \boldsymbol{I}_K)$, where $\boldsymbol{0}_K$ is a K-dimensional vector of zeros, and $\boldsymbol{I}_K$ is the $K \times K$ identity matrix. The columns of $\boldsymbol{W}$ were reordered, so that the energy contained in $\boldsymbol{w}_{\cdot k}$ decreased monotonically. Our observations consisted of $U = 12$ sequences of event arrivals, each with a base Poisson rate λ_u drawn from a truncated Gaussian distribution centered at λ_0 with small variance ($\ll \lambda_0$). As explained below, we varied both the average base rate, λ_0, and the observation time length, T. For inference, the fixed hyperparameters of the sampler were set as: $\alpha = 3, c = d = e = f = 10^{-3}$, and $\boldsymbol{\pi}_k = [0.5, 0.5]^T$. We ran 5000 MCMC iterations of our MCMC sampler, discarding the first 2000 as burn-in, with posterior samples collected every 5 iterations. The running time of a typical trial (with $T = 1000$ and about 120 event arrivals for each user) was about 3000 seconds with unoptimized Matlab code on a computer with 2.2GHz CPU and 8GB RAM. To evaluate the mixing behavior of the sampler, we use R-coda (Plummer et al., 2006) to compute the effective sample size (ESS), as well as Markov chain autocorrelation functions (ACF) of various model parameters. Table 1 shows these statistics, with the ACF shown for the parameters λ_1, $\boldsymbol{w}_{11}$ and n_1, the number of transitions of the first latent source. These numbers are typical of MCMC samplers, and show the sampler mixes well [1].

Table 1. Effective sample size (ESS) and autocorrelation function (ACF) of the base rate (λ_u), number of transitions of source $s_k(t)$ (n_k), and the factor loadings ($\boldsymbol{w}_{uk}$)

PARAMETERS	λ_u	n_k	$\boldsymbol{w}_{uk}$
ESS/ITERATION	0.046	0.278	0.161
ESS/SECOND	0.076	0.467	0.266
ACF (LAG 5)	0.609	0.101	0.225
ACF (LAG 10)	0.415	0.049	0.100
ACF (LAG 50)	-0.047	-0.016	-0.039

Instantaneous Rate Estimation: One of the main advantages to the proposed model is the ability to exploit correlations across streams of observed arrival data. This is especially important when the base arrival rate of each user is quite low. In this experiment, we examine the ability to accurately recover $\{\gamma_u(t)\}$, the instantaneous arrival rate of each user, choosing the mean of base rates λ_0 from values in $\{0.01, 0.02, 0.05, 0.10, 0.20\}$. We keep the observation time length constant at $T = 2000$.

We measured the estimation error of the instantaneous rate by discretizing the interval $[0, T]$ using $N = 1000$ evenly spaced grid points. We compute the posterior mean estimation error at each grid point, normalizing with respect to that point's true rate, and record the average normalized error for 15 repeats. Rate inference is performed using both the proposed model, and the binning approach of (Zhang & Kou, 2010) using bandwidth values of $1, 3, 10$ times the inverse of mean arrival rate. The results are shown in Figure 1.

For very low base arrival rates (*i.e.,* $\lambda_0 = 0.01$) all methods perform similarly. As λ_0 increases, our model performs

[1] Code available at `http://people.duke.edu/~wl89/`

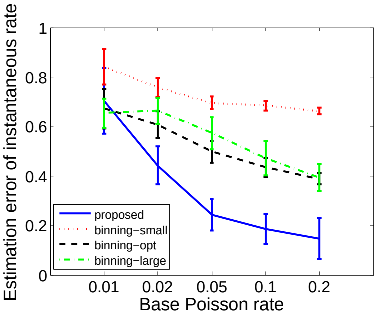

Figure 1. Normalized estimation error of instantaneous Poisson rate for proposed method and binning methods

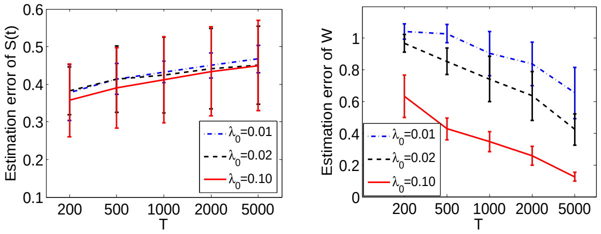

Figure 2. Estimation error of latent sources and factor loadings with different observation lengths and base rates

significantly better than the competing binning method. For a mean base rate of $\lambda_0 = 0.05$, we find that the error rate is less than half the error rate of the best choice of binning bandwidth.

Factor Loading Matrix Inference: As Figure 1 suggests, at low arrival rates there is not enough information to recover the instantaneous Poisson rates (and thus the state of the latent sources). This is shown in the left plot of Figure 2: for base rates $\lambda_0 = \{0.01, 0.02, 0.10\}$, as T increases (taking values in $T_{cand} = \{200, 500, 1000, 2000, 5000\}$), the error in the latent sources (estimated over a grid as in the previous section) increases slightly: this is because the 'number of variables' in a bsMJP path increases with T.

In such situations, it is still of interest of estimate parameters like the factor loading matrix $\boldsymbol{W}$: even if we cannot reconstruct exactly what a user had in mind at any previous time, we would still like to characterize their behaviour to make predictions in the future. The right plot shows that the estimation error of $\boldsymbol{W}$ decreases monotically as the observation interval increases, implying that these parameters can be recovered even if the posterior distribution over latent sources never concentrates. Here, for each posterior sample, $\hat{\boldsymbol{W}}$, the estimation error with respect to the true factor loadings, $\boldsymbol{W}$, is computed as $\frac{||\hat{\boldsymbol{W}} - \boldsymbol{W}||_2}{||\boldsymbol{W}||_2}$.

Deviation from Markovianity: Setting the Weibull shape

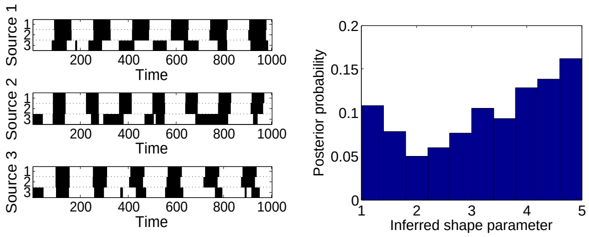

Figure 3. Left: inferred sources using bsMJP and bMJP models. For each subplot, the first row shows the truth, the second the inferred source using bsMJP, and the third, that using bMJP. Right: posterior distribution of a shape parameter inferred using bsMJP

parameter β to 1 recovers the exponential distribution, reducing the latent sources to memoryless binary MJPs (bMJPs). To show the flexibility afforded by this parameter, we consider latent sources that are square waves, switching between '1' and '0' at a fixed frequency. Figure 3 compares inferences over a latent source using a bMJP and a bsMJP prior. We see that the state intervals inferred by bMJP are more irregular, showing unnecessary switching between states. For the bsMJP, we placed a uniform prior on $[1, 5]$ on the shape parameter, allowing the model to estimate the state persistence. Figure 3 shows the posterior over this parameter places significant mass away from 1, forcing more regular state holding times.

Latent Factor Number Estimation: In real world applications, the number of latent sources is usually unknown *a priori*. Our MGP shrinkage prior from Section 2 allows us to infer the number of dominant sources. Again, we vary the observation length from $T_{cand} = \{500, 2000, 5000\}$, set the base arrival rate $\lambda_0 = 0.10$, and set the true number of latent sources as $K = 3$. When doing inference, we truncate the number of sources, picking a large enough number $K_+ = 10$ to avoid under-fitting. Though the sampler is not sensitive to α, a setting $\alpha \to 1$ leads to a higher chance of sampling precision sequences which are not monotonically increasing. As mentioned before, α is set as 3 in our experiments. For each posterior sample, we identify the relevant sources by thresholding the associated weight-vector $\boldsymbol{w}_{\cdot k}$, picking the smallest collection of weights containing 90 percent of total energy of $\boldsymbol{W}$. Figure 4 demonstrates the behavior of the posterior distribution with respect to the inferred number of latent sources. We find that as an increasing number of observations are available, the posterior mass quickly concentrates around the true value $K = 3$.

6.2. Skin Conductance Biometrics

Finally, we apply our model to a dataset from a real-world biometrics application. This dataset was collected with 10 volunteers (users) watching a feature-length film while wearing Galvanic Skin Response (GSR) sensors (the Af-

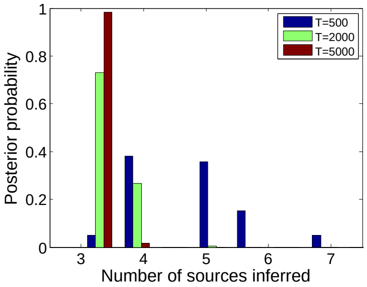

Figure 4. Posterior distribution of number of latent sources

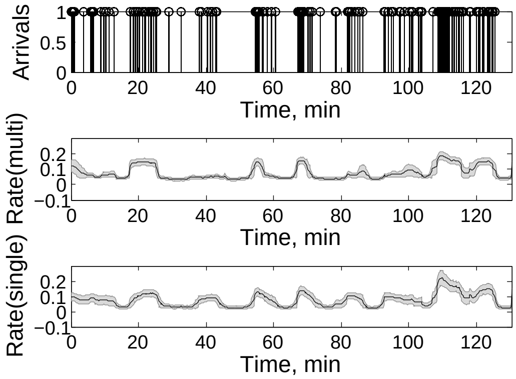

Figure 5. Extracted arousal events from one user. Top: event arrivals; middle: jointly estimated rate. Bottom: rate inferred from single user's trace

fectiva Q-Sensor, (Affectiva, 2012)) measuring skin conductance. Using a current state-of-the-art decomposition approach from (Silveira et al., 2013), we extract user arousal responses from the observed skin conductance signals to obtain arrival data for each user across the two-hour, eighteen minute film. As shown in Figure 5, a typical user has about 150 event arrivals during the recording, similar to the $\lambda_0 = 0.02$ scenario in the synthetic experiments.

Below, we analyze users' arousal level based on their GSR signals. We also analyze the similarity between users and explore dominant latent events in the movie using our model. We validate our analyses using two types of explicit feedback obtained from this experiment. First, all the users were asked to recall 10 scenes from the film and rate their arousal intensity from 1 to 10 for each scene ("1" being "tedious" and "10" representing "exciting/tense"). Second, the general rating of the movie from 1 to 5 was also collected, indicating the user's overall opinion of the film ({"Poor", "Fair","Good","Very Good","Excellent"}).

We estimate the rate function, factor loading matrix, and latent sources using MCMC posterior samples. As discussed in Section 4, a non-informative prior $\text{Ga}(10^{-3}, 10^{-3})$ is put on λ_u and $\mu_k^{(s)}$, and uniform prior on $[1, 5]$ is put on $\beta_k^{(s)}$. The maximum number of latent sources is set as $K_+ = 10$. Hyperprior $\alpha = 3$, results in factor loading matrices $\boldsymbol{W}$ where on average about $K = 3$ columns contain 95% of total energy. Producing 10000 samples took about 12000 seconds.

Instantaneous Rate Inference: To qualitatively show the benefit of sharing information across users, we ran our model using both multiple traces ($U = 10$) and a single trace ($U = 1$) (both with a maximum of $K = 10$ latent sources). The middle and bottom plots of Figure 5 show posterior mean of the estimated rate function (along with an uncertainty interval equal to one posterior standard deviation) for the shared and separate cases. We see that using multiple trace information gives estimates of the rate function that are piecewise stable, unlike the single trace case where the estimated rate function tends to follow the empirical Poisson rate. The inferred hazard function tends to be more parsimonious in terms of state-changes as well, as it is contrained to explain more data. Such information is useful to tease apart responses to common stimulus from user-specific responses, and we look at this in more detail at the end of this section.

To quantitatively measure our rate estimation, we correlated our estimated rate values with the explicit arousal feedback for the 10 scenes. We transformed the recalled arousal intensity (1-10) into binary labels by comparing each to the average arousal intensity of each user (such that each user has 5 scenes with a "0" class label and five scenes with a "1" class label). Using these binary labels as ground truth, we compare the instantaneous rate from the posterior mode estimate of our proposed method against that derived from the binning approach and a Markovian version of our methodology using bMJP. For all approaches, we evaluate the inferred rate function at time points corresponding to the 10 recalled scenes for all users and then inferred the binary labels at those time points by thresholding. Varying the threshold, we plot the ROC curves in the left plot of Figure 6.

As shown in the figure, the instantaneous rate inferred by the proposed model conforms to the user explicit feedback better than the rate estimated via the binning approach and the simplified model with bMJP. Specifically, our proposed algorithm is able to correctly classify almost 40% of the user explicit scene ratings with no false alarms, while the binning approach only classifies 10% of the scenes with no false alarms.

Factor Loading Matrix Inference: Each user has their own factor loading vector which can be used to calculate the distance between pairs of users. Thus, we compute the pairwise Euclidean distance between users using the posterior mode estimate of the factor loading matrix. We then test how well this user-similarity metric predicts the user ratings. Using all 45 possible pairs of users, we plot two sets of ROC curves: in the first, we compare pairs of users with the same rating (*e.g.,* both users rate the film at "4") versus users that differ by a single rating point (*e.g.,* "3" and "4"); in the second, we compare pairs of users with the same rating versus users that differ by two ratings points (*e.g.,* "3" and "5"). As illustrated in the right plot of Figure 6, the proposed method does well with predicting user rating similarity, with the ability to classify over 55% of the users with the same rating from the set of users two ratings apart, with no false alarms.

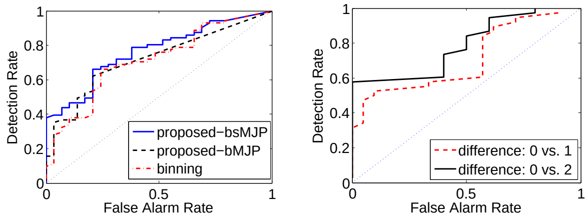

Figure 6. ROC curves predicting (left) user arousal intensity, (right) user rating similarity from loading vectors $\boldsymbol{w}_u$

Latent Source Inference: Finally, we analyze the dominant latent sources underlying the observed arousal responses returned by our inferences. The left plot of Figure 7 shows the posterior distribution over the possible number of dominant sources, defined as the minimum number of columns of the factor loading matrix containing 90 percent of total energy. The posterior mass concentrates around 4 and 5, and we plot the 5 dominant latent sources in the right plot of Figure 7. The first source is an offset of the baseline Poisson rate. We found the following 4 sources had fairly clear interpretations. For the second source, the elements in the corresponding column of the factor loading matrix are all positive, indicating this factor enhances arousal intensity. Specifically, this is activated at around 20 minutes for scenes about a plane crash, around 55 minutes and 65 minutes for key turning points of the plot, and around 115 minutes and 125 minutes for a climax and a surprising denouement respectively. Taking the third source as another example, both positive and negative factor loadings exist among all users, indicating this factor enhances arousal intensity for some of users but suppresses it for others. This is activated for the scene when the main actor first meets the main actress, and for the occurrence of their last dialogue. Such information can be used along with user information to better understand users, the stimulus, and the interaction between the two.

7. Discussion

There are a number of variations and extensions to our modeling choices worth exploring. While we placed a multiplicative gamma shrinkage prior on the factor loading matrix, an alternative is to construct similar priors using the framework of Lévy processes (Applebaum, 2004; Polson & Scott, 2012). Similarly, we can allow added structure to the loading matrix $\boldsymbol{W}$ (e.g., clustering its rows, to infer clusters of users), or allow $\boldsymbol{W}$ to vary with time (modeling the evolution of a users tastes). Another important extension incorporates user-covariates like age, sex, or profession. Ultimately, the goal is not just to understand users and stimuli, but to use models like ours to adaptively modify the stimulus by monitoring user response. This is central to applications ranging from human prosthetics and brain-computer interface to recommendation and targeted advertising.

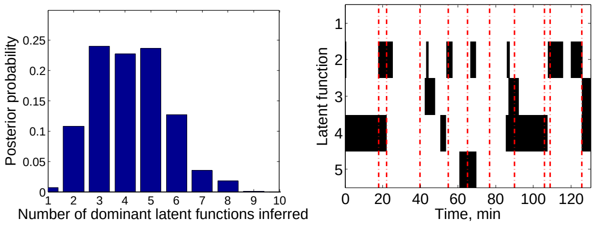

Figure 7. Posterior over number of dominant sources with a sample of sources (black represents 1 and red dash-dot lines delineate scenes the users were asked to recall)

Acknowledgements: This work was supported in part by ARO, DARPA, NGA and ONR.

References

Affectiva, Inc. Liberate yourself from the lab: Q Sensor measures EDA in the wild. *Affectiva White Paper*, 2012.

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

Applebaum, D. *Lévy Processes and Stochastic Calculus*. Cambridge University Press, 2004.

Barbieri, R., Matten, E., Alab, A., and Brown, E. Estimating a state-space model from point process observations. *American Journal of Physiology-Heart and Circulatory Physiology*, 288: 424–435, 2005.

Bhattacharya, A. and Dunson, D. Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011.

Daley, D. and Vere-Jones, D. *Introduction to the Theory of Point Processes*. Springer-Verlag, 1998.

Escola, S., Fontanini, A., Katz, D., and Paniski, L. Hidden Markov models for the simulus-response relationships of multistate neural systems. *Neural Computation*, 23:1071–1132, 2011.

Feller, W. On semi-Markov processes. *Proceedings of the National Academy of Sciences of the United States of America*, 51 (4):653, 1964.

Fox, E.B., Sudderth, E.B., Jordan, M.I., and Willsky, A.S. A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

Früwirth-Schnatter, S. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15:183–202, 1994.

Golightly, A. and Wilkinson, D. J. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820, December 2011.

Griffiths, T. L. and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. In *NIPS Conference*, volume 18, 2006.

Johnson, Matthew J. and Willsky, Alan S. Bayesian nonparametric hidden semi-Markov models. *Journal of Machine Learning Research*, 14:673–701, February 2013.

Kass, R. and Ventura, V. A spike-train probability model. *Neural Computation*, 13:1713–1720, 2001.

Nodelman, U., Shelton, C.R., and Koller, D. Continuous time Bayesian networks. In *UAI Conference*, pp. 378–387, 2002.

Ogata, Y. and Tanemura, M. Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method. *Biometrics*, 41:421–433, 1985.

Plummer, M., Best, N., Cowles, K., and Vines, K. Coda: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1): 7–11, 2006.

Polson, N. G. and Scott, J. G. Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311, 2012.

Rao, V. and Teh, Y. MCMC for continuous-time discrete-state systems. In *NIPS Conference*, Lake Tahoe, NV, 2012.

Rao, V. and Teh, Y. Fast MCMC sampling for Markov jump processes and extensions. *Journal of Machine Learning Research*, 14:3207–3232, 2013.

Riehle, A., Grun, S., Diesmann, M., and Aertsen, A. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science*, 278:1950–1953, 1997.

Ryu, B. and Lowen, S. Point process approaches to the modeling and analysis of self-similar traffic. I. Model construction. In *IEEE Infocom Conference*, San Francisco, CA, 1996.

Saeedi, A. and Bouchard-Côté, A. Priors over recurrent continuous time processes. In *NIPS Conference*, volume 24, 2011.

Scott, S. L. and Smyth, P. The Markov modulated Poisson process and Markov Poisson cascade with applications to web traffic modeling. *Bayesian Statistics*, 7:1–10, 2003.

Silveira, F., Eriksson, B., Sheth, A., and Sheppard, A. A biometrics based approach to characterize viewer responses to movies. In *ACM Ubicomp Conference*, Zurich, Switzerland, 2013.

Smith, A. and Brown, E. Estimating a state-space model from point process observations. *Neural Computation*, 15:965–991, 2003.

Teh, Y. W., Blundell, C., and Elliott, L. T. Modelling genetic variations with fragmentation-coagulation processes. In *NIPS Conference*, 2011.

Van Gael, J., Teh, Y. W., and Ghahramani, Z. The infinite factorial hidden Markov model. In *NIPS Conference*, volume 21, 2009.

Yu, M Byron., Cunningham, J. P, Santhanam, G., Ryu, S. I, Shenoy, K. V, and Sahani, M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102(1):614–635, 2009.

Zhang, T. and Kou, S. C. Nonparametric inference of doubly stochastic Poisson process data via the kernel method. *The Annals of Applied Statistics*, 4(4):1913, 2010.