Efficient MCMC for Continuous Time Discrete State Systems

Vinayak Rao and Yee Whye Teh

Gatsby Computational Neuroscience Unit,
University College London

Overview

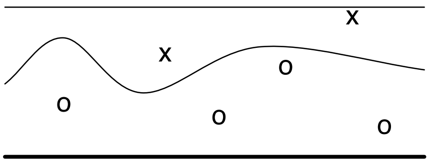
- Continuous time discrete state systems: applications in physics, chemistry, genetics, ecology, neuroscience etc.
- The simplest example: the Poisson process on the real line.
- Generalizations: renewal processes, Markov jump processes, continuous time Bayesian networks etc.
- These relate back to the basic Poisson process via the idea of *uniformization*.
- We use this connection to develop tractable models and efficient MCMC sampling algorithms.

Thinning

Uniformization generalizes the idea of 'thinning'.

Thinning: to sample from a Poisson process with rate $\lambda(t)$.

- Sample from a Poisson process with rate $\Omega > \lambda(t)\ \forall t$.
- Thin or reject each point with probability $1 - \frac{\lambda(t)}{\Omega}$.



Follows from the *complete randomness* of the Poisson process.

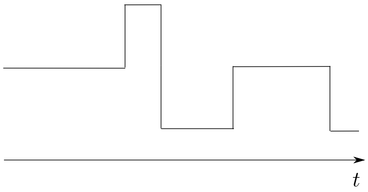
Markov jump processes or renewal processes are *not* completely random: *Uniformization*—thin points by running a *Markov chain*.

Uniformization (at a high level)

- Draw from a Poisson process with rate Ω.
- Ω is larger than the fastest rate at which 'events occur'.
- Construct a Markov chain with transition times given by the drawn point set.
- The Markov chain is *subordinated* to the Poisson process.
- Keep a point t with probability $\lambda(t|state)/\Omega$.

Markov jump processes (MJPs)

An MJP $\mathbf{S}(t),\ t \in \mathbb{R}_+$ is a right-continuous piecewise-constant stochastic process taking values in some finite space. $\mathcal{S} = \{1, 2, ...n\}$. It is parametrized by an *initial distribution* π and a *rate matrix* A.



$$\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix}$$

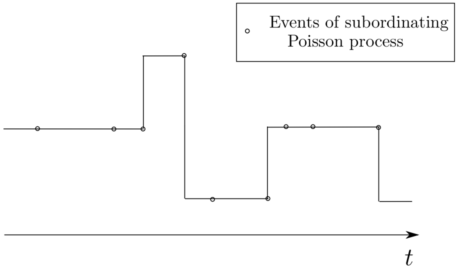
A_{ij} : rate of leaving state i for j

$$A_{ii} = -\sum_{j=1, j\neq i}^{n} A_{ij}$$

$|A_{ii}|$: rate of leaving state i

Uniformization for MJPs

- Alternative to Gillespie's algorithm.
- Sample a set of times from a Poisson process with rate $\Omega \geq \max_i |A_{ii}|$ on the interval $[t_{start}, t_{end}]$.
- Run a discrete time Markov chain with initial distribution π and transition matrix $B = (I + \frac{1}{\Omega} A)$ on these times.



The matrix B allows self-transitions.
[Jensen, 1953]

Uniformization for MJPs [Jensen, 1953]

Lemma

For any $\Omega \geq \max_i |A_{ii}|$, the (continuous time) sequence of states obtained by the uniformized process is a sample from a MJP with initial distribution π and rate matrix A.

Auxiliary variable Gibbs sampler

Given noisy observations of an MJP, obtain samples from the posterior.

Observations can include:

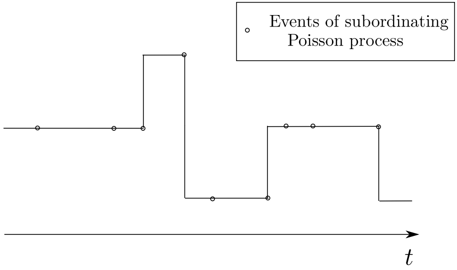
- State values at the end points of an interval.
- Observations $x(t) \sim F(\mathbf{S}(t))$ at a finite set of times t.
- More complicated likelihood functions that depend on the entire trajectory, e.g. Markov modulated Poisson processes and continuous time Bayesian networks (see later).

State space of Gibbs sampler consist of:

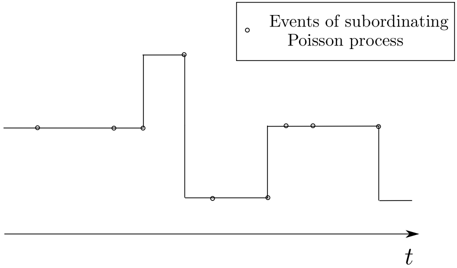
- Trajectory of MJP $\mathbf{S}(t)$.
- Auxiliary set of points rejected via self-transitions.

[Rao and Teh, 2011a]

Auxiliary variable Gibbs sampler

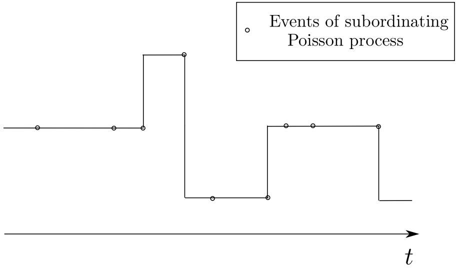


Auxiliary variable Gibbs sampler



- Given current MJP path, we need to resample the set of rejected points. Conditioned on the path, these are:
  - *independent of the observations*,
  - produced by 'thinning' a rate Ω Poisson process with probability $1 + \frac{A_{\mathsf{S}(t)\mathsf{S}(t)}}{\Omega}$,
  - thus, distributed according to a inhomogeneous Poisson process with piecewise constant rate $(\Omega + A_{\mathsf{S}(t)\mathsf{S}(t)})$.

Auxiliary variable Gibbs sampler



- Given all potential transition points, the MJP trajectory is resampled using the forward-filtering backward-sampling algorithm.
- The likelihood of the state between 2 successive points must include all observations in that interval.

Comments

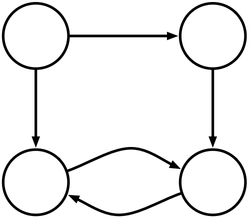
- Complexity: $O(n^2 P)$, where P is the (random) number of points.
- Can take advantage of sparsity in transition rate matrix A.
- Only dependence between successive samples is via the transition times of the trajectory.
- Increasing Ω reduces this dependence, but increases computational cost.
- Sampler is ergodic for any $\Omega > \max_i |A_{ii}|$.

Existing approaches to sampling

[Fearnhead and Sherlock, 2006, Hobolth and Stone, 2009] produce *independent* posterior samples, marginalizing over the infinitely many MJP paths using matrix exponentiation.

- scale as $O(n^3 + n^2 P)$.
- any structure, e.g. sparsity, in the rate matrix A cannot be exploited in matrix exponentiation.
- cannot be easily extended to complicated likelihood functions (e.g. Markov modulated Poisson processes, continuous time Bayesian networks).

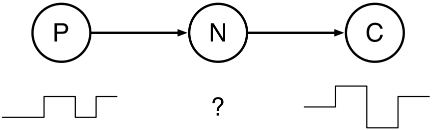
Continuous-time Bayesian networks (CTBNs)



- Compact representations of large state space MJPs with structured rate matrices.
- Applications include ecology, chemistry , network intrusion detection, human computer interaction etc.
- The rate matrix of a node at time is determined by the configuration of its parents at that time.

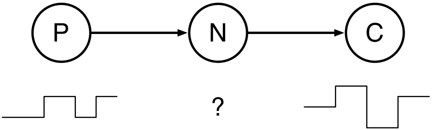
[Nodelman et al., 2002]

Gibbs sampling CTBNs via uniformization



- The trajectories of all nodes are piecewise constant.
- In a segment of constant parent (P) values, the dynamics of N are controlled by a fixed rate matrix A^P.
- Each child (C) trajectory is effectively a *continuous-time* observation.

Gibbs sampling CTBNs via uniformization



- Sample candidate transition times from a Poisson process with rate $\Omega > A^P_{ii}$.
- Between two successive Poisson events, N remains in a constant state.
  - This state must account for the likelihood of children nodes' states.
  - The state must also explain relevant observations.
- With the resulting 'likelihood' function and transition matrix $B = (I + \frac{1}{\Omega}A^P)$, sample new trajectory using forward-filtering backward-sampling.

Existing approaches to inference

[El-Hay et al., 2008] describe a Gibbs sampler involving time discretization, which is expensive and approximate.

[Fan and Shelton, 2008] uses particle filtering which can be inaccurate for long time intervals.

[Nodelman et al., 2002, Nodelman et al., 2005, Opper and Sanguinetti, 2007, Cohn et al., 2010] use deterministic approximations (mean-field and expectation propagation) which are biased and can be inaccurate.

Experiments

- We compare our uniformization-based sampler with a state-of-the-art CTBN Gibbs sampler of [El-Hay et al., 2008]. search on the time interval.
- When comparing running times, we measured times required to produce same effective sample sizes.

Experiments

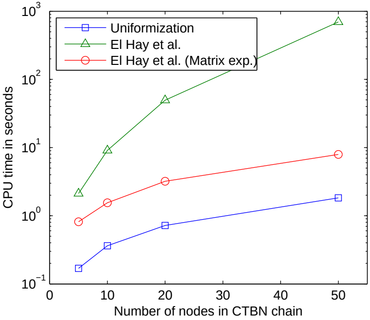


Figure: CPU time vs length of CTBN chain.

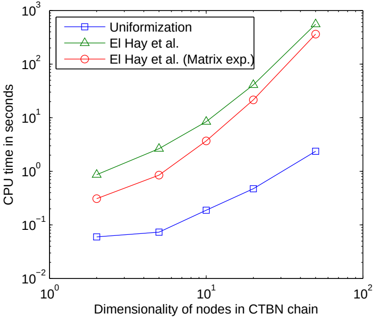


Figure: CPU time vs number of states of CTBN nodes.

The plots above were produced for a CTBN with a chain topology, increasing the number of nodes in the chain (left) and the number of states of each node (right).

Experiments

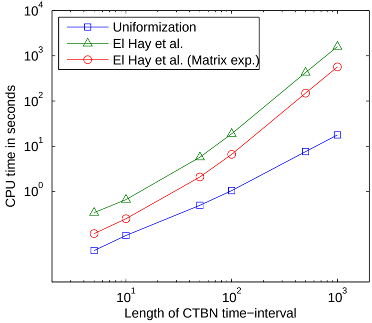


Figure: CPU time vs time interval of CTBN paths.

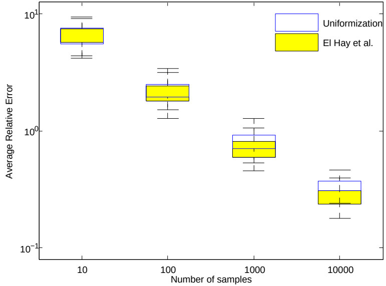
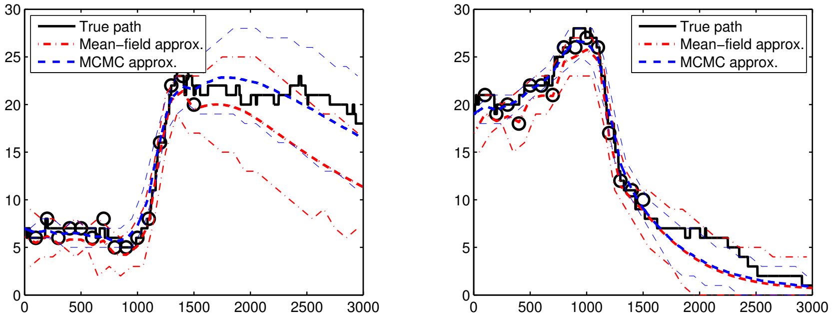


Figure: Average relative error vs number of samples

Produced for the standard 'drug network'.
Left: required CPU time as length of the time interval increases.
Right: (normalized) absolute error in estimated parameters of the network as the (absolute) number of samples increases.

Experiments

Compared against the mean-field approximation of [Opper and Sanguinetti, 2007], for the predator-prey model, a CTBN describing the Lotka-Volterra equations.



Posterior (mean and 90% confidence intervals) over predator paths (observations (circles) only until 1500).

Renewal processes

- Renewal processes: point processes on the real line ('time').
- Inter-event times drawn i.i.d. from some *renewal density*.
- Homogeneous Poisson process: exponential renewal density.
- Can capture burstiness or refractoriness.

Our contribution: modulated renewal processes:

- Nonstationarity: allow external time-varying factors to modulate the inter-event distribution.
- We place a (transformed) Gaussian process prior on the intensity function.

[Rao and Teh, 2011b]

Modulated renewal processes

- Associated with the renewal density g is a *hazard function* h.
- For an infinitesimal Δ, $h(\tau)\Delta$ is the probability of the inter-event interval being in $[\tau, \tau + \Delta]$ conditioned on it being at least τ:

$$h(\tau) = \frac{g(\tau)}{1 - \int_0^\tau g(u)du}$$

- Modulate the hazard function by some time-varying intensity function $\lambda(t)$:

$$h(\tau, t) \equiv m(h(\tau), \lambda(t))$$

- $m(\cdot, \cdot)$ is some *interaction function*.
- We use multiplicative interactions, $h(\tau, t) = h(\tau)\lambda(t)$.
- Another interaction function is additive $h(\tau, t) = h(\tau) + \lambda(t)$.

Modulated renewal processes (continued)

- We place a Gaussian Process prior on the intensity function $\lambda(t)$, transformed via a sigmoidal link function.
- We use a gamma family for the hazard function:

$$h(\tau) = \frac{x^{\gamma-1}e^{-x}}{\int_x^{\infty} u^{\gamma-1}e^{-u}du}$$

where γ is the shape parameter. The generative process is:

$$\begin{aligned} l(\cdot) &\sim \mathcal{GP}(\mu, K) \\ \lambda(\cdot) &= \hat{\lambda}\sigma(l(\cdot)) \\ G &\sim \mathscr{R}(\lambda(\cdot), h(\cdot)) \end{aligned}$$

- We place hyperpriors on $\hat{\lambda}, \gamma$ and the GP hyperparameters

Direct sampling from prior

The modulated renewal density is:

$$g(\tau|t_{prev}) = \lambda(t_{prev} + \tau)h(\tau)\exp\left(-\int_0^\tau \lambda(t_{prev} + u)h(u)du\right)$$

where t_{prev} is the previous event time.

Naïvely, need to numerically evaluate integrals to generate samples.

- can be time consuming and introduce approximation errors.

Sampling via uniformization

- Assume the intensity function $\lambda(t)$ and the hazard function $h(\tau)$ are bounded

$$\exists\,\Omega \geq \max_{t,\tau} h(\tau)\lambda(t)$$

- Sample $E = \{E_0 = 0, E_1, E_2, \ldots\}$ from a Poisson process with rate Ω.
- Let $\{Y_0 = 0, Y_1, Y_2, \ldots\}$ be an integer-valued Markov chain on the times in E, where each Y_i either equals Y_{i-1} or i.
    - $Y_i = Y_{i-1} \rightarrow$ reject E_i,
    - $Y_i = i \rightarrow$ keep E_i.
- $E_i - E_{Y_i}$: time since the last accepted event. For $i > j \geq 0$, define

$$p(Y_i = i | Y_{i-1} = j) = \frac{h(E_i - E_j)\lambda(E_j)}{\Omega}$$

- Define $G = \{E_i \in E \text{ s.t. } Y_i = i\}$.

Sampling via uniformization

Lemma

For any $\Omega \geq \max_{t,\tau} h(\tau)\lambda(t)$*, G is a sample from a modulated renewal process with hazard $h(\cdot)$ and modulating intensity $\lambda(\cdot)$.*

Sampling via uniformization

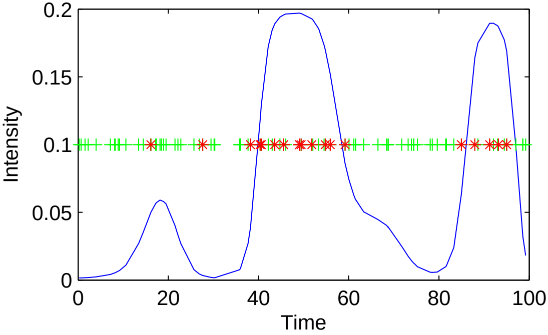


Figure: Green: rejected events, Red: sample for a Gamma(3) modulated renewal process.

Reduction to thinning of Poisson processes

For a Poisson process, the hazard function is a constant:

$$h(\tau) = h$$

Then, the transition probabilities of the Markov chain becomes:

$$p(Y_i = i | Y_{i-1} = j) = \frac{h\lambda(E_j)}{\Omega}$$

This reduces to independent thinning [Adams et al., 2009].

Inference

Given a set of event times G, obtain sample from the modulating function $\lambda(\cdot)$ (and hyperparameters).

As before, directly sampling from the GP posterior is impossible.

Introduce the rejected events as auxiliary variables and proceed by alternately sampling the rejected events given G and the intensity function, and then the intensity function given G and rejected events.

Inference

Assume the modulating function $\lambda(t)$ is known for all t.

In the interval (G_{i-1}, G_i), events from a rate Ω Poisson process were rejected with probability:

$$1 - \frac{\lambda(t)h(t - G_{i-1})}{\Omega}$$

Under the posterior, these rejected events are distributed as an inhomogeneous Poisson process with rate:

$$\Omega - \lambda(t)h(t - G_{i-1})$$

Inference

Assume the modulating function $\lambda(t)$ is known for all t.

In the interval (G_{i-1}, G_i), events from a rate Ω Poisson process were rejected with probability:

$$1 - \frac{\lambda(t)h(t - G_{i-1})}{\Omega}$$

Under the posterior, these rejected events are distributed as an inhomogeneous Poisson process with rate:

$$\Omega - \lambda(t)h(t - G_{i-1})$$

Catch: we know $\lambda(t)$ only at a discrete set of times. Use thinning method of GP Cox processes [Adams et al., 2009].

Inference

Assume the modulating function $\lambda(t)$ is known for all t.

In the interval (G_{i-1}, G_i), events from a rate Ω Poisson process were rejected with probability:

$$1 - \frac{\lambda(t)h(t - G_{i-1})}{\Omega}$$

Under the posterior, these rejected events are distributed as an inhomogeneous Poisson process with rate:

$$\Omega - \lambda(t)h(t - G_{i-1})$$

Catch: we know $\lambda(t)$ only at a discrete set of times. Use thinning method of GP Cox processes [Adams et al., 2009].

We resample the GP on the events and the rejected points using elliptical slice sampling [Murray et al., 2010].

Computational considerations

- Complexity: $O(N^3)$, where $N = |G| + 2|E|$, $|G|$ is the number of observations and $|E|$ is the number of rejected points.
- For large G, we must resort to approximate inference for Gaussian processes [Rasmussen and Williams, 2006].
- Question: how do these approximations compare with time-discretized approximations like [Cunningham et al., 2008]?

Experiments

Three synthetic datasets generated by modulating a Gamma(3) renewal process.

- $\lambda_1(t) = 2\exp(t/5) + \exp(-((t-25)/10)^2, t \in [0, 50]$: 44 events
- $\lambda_2(t) = 5\sin(t^2) + 6, t \in [0, 5]$: 12 events
- $\lambda_3(t)$: a piecewise linear function, $t \in [0, 100]$: 153 events

Three settings of our model and a strawman:

- with the shape parameter fixed to 1 (MRP Exp),
- with the shape parameter fixed to 3 (MRP Gam3),
- with a hyperprior on the shape parameter (MRP Full),
- an approximate discrete time sampler on a regular grid covering the interval, all intractable integrals were approximated numerically.

Experiments

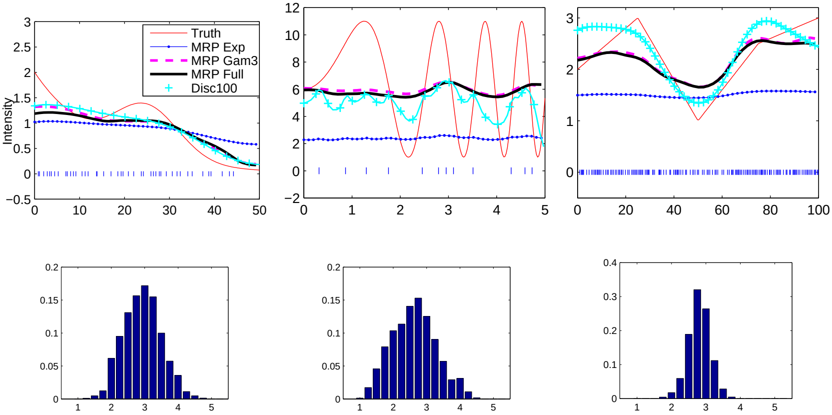


Figure: Synthetic datasets 1-3: Posterior mean intensities (top) and Gamma shape posteriors (bottom). Results from 5000 MCMC samples after a burn-in of 1000 samples.

Experiments

	MRP Exp	MRP Gam3	MRP Full	Disc25	Disc100
l_2 error	7.85	3.19	2.55	4.09	**2.43**
log pred.	-47.55	-38.07	**-37.37**	-41.65	-41.02
l_2 error	141.01	**56.22**	58.44	91.32	57.9
log pred.	-3.70	**-2.95**	-3.28	-5.25	-3.85
l_2 error	82.03	**11.42**	13.44	122.34	38.05
log pred.	-89.88	**-48.28**	-48.57	87.17	-55.80

Table: l_2 distance from the truth and mean log predictive probabilities of test sets for synthetic datasets 1 (top) to 3 (bottom).

Experiments

Dataset: the coal mine disaster dataset, recording the dates of a series of 191 coal mining disasters (each of which killed ten or more men [Jarrett, 1979]).

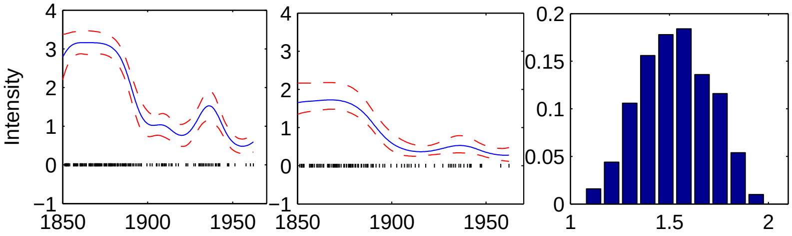


Figure: Left: posterior mean of the intensity function. The posterior for shape parameter was close to 1. Middle and right: results after deleting every alternate event.

Experiments

Dataset: neural spike train recorded from grasshopper auditory receptor cells [Rokem et al., 2006].

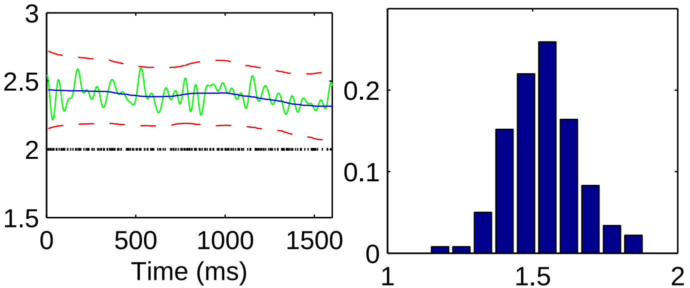


Figure: Left: Posterior mean intensity for neural data with 1 standard deviation error bars. Superimposed is the log stimulus (scaled and shifted). Right: Posterior over the gamma shape parameter.

Experiments

We compare our uniformization based blocked Gibbs sampler with the sampler of [Adams et al., 2009].

	Synthetic dataset 1		
	Mean ESS	Minimum ESS	Time(sec)
Gibbs	93.45 ± 6.91	50.94 ± 5.21	77.85
MH	56.37 ± 10.30	19.34 ± 11.55	345.44
	Coalmine dataset		
	Mean ESS	Minimum ESS	Time(sec)
Gibbs	53.54 ± 8.15	24.87 ± 7.38	282.72
MH	47.83 ± 9.18	18.91 ± 6.45	1703

Table: Sampler comparisons. Numbers are per 1000 samples.

Besides mixing faster our sampler:

- is simpler and more natural to the problem,
- does not require any external tuning.

Conclusions

- The idea of uniformization relates more complicated continuous time discrete state processes to the basic Poisson process.
- We demonstrated how this connection can be used to develop tractable models and efficient MCMC inference schemes.
- We can look into extending the models we discussed here:
  - renewal processes with unbounded hazard rates,
  - semi-Markov jump processes,
  - inhomogeneous MJPs, MJPs with infinite state spaces etc.
- Other applications we wish to study, such as survival analysis, queuing systems etc.

Bibliography I

Adams, R. P., Murray, I., and MacKay, D. J. C. (2009).
Tractable nonparametric Bayesian inference in Poisson processes with gaussian process intensities.
In Bottou, L. and Littman, M., editors, *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 9–16, Montreal. Omnipress.

Cohn, I., El-Hay, T., Friedman, N., and Kupferman, R. (2010).
Mean field variational approximation for continuous-time bayesian networks.
J. Mach. Learn. Res., 11:2745–2783.

Cunningham, J. P., Yu, B. M., Shenoy, K. V., and Sahani, M. (2008).
Inferring neural firing rates from spike trains using gaussian processes.
In *Advances in Neural Information Processing Systems,20*.

El-Hay, T., Friedman, N., and Kupferman, R. (2008).
Gibbs sampling in factorized continuous-time Markov processes.
In *UAI*, pages 169–178.

Fan, Y. and Shelton, C. R. (2008).
Sampling for approximate inference in continuous time Bayesian networks.
In *Tenth International Symposium on Artificial Intelligence and Mathematics*.

Fearnhead, P. and Sherlock, C. (2006).
An exact Gibbs sampler for the Markov-modulated Poisson process.
Journal Of The Royal Statistical Society Series B, 68(5):767–784.

Hobolth, A. and Stone, E. A. (2009).
Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution.
Ann Appl Stat, 3(3):1204.

Bibliography II

Jarrett, B. Y. R. G. (1979).
A note on the intervals between coal-mining disasters.
Biometrika, 66(1):191–193.

Jensen, A. (1953).
Markoff chains as an aid in the study of Markoff processes.
Skand. Aktuarietiedskr., 36:87–91.

Murray, I., Adams, R. P., and MacKay, D. J. (2010).
Elliptical slice sampling.
JMLR: W&CP, 9:541–548.

Nodelman, U., Koller, D., and Shelton, C. (2005).
Expectation propagation for continuous time Bayesian networks.
In *Proceedings of the Twenty-first Conference on Uncertainty in AI (UAI)*, pages 431–440, Edinburgh, Scottland, UK.

Nodelman, U., Shelton, C., and Koller, D. (2002).
Continuous time Bayesian networks.
In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 378–387.

Opper, M. and Sanguinetti, G. (2007).
Variational inference for Markov jump processes.
In *NIPS*.

Rao, V. and Teh, Y. W. (2011a).
Fast MCMC sampling for Markov jump processes and continuous time Bayesian networks.
In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*.

Bibliography III

Rao, V. and Teh, Y. W. (2011b).
Gaussian process modulated renewal processes.
In *Advances in Neural Information Processing Systems 23*.

Rasmussen, C. E. and Williams, C. K. I. (2006).
Gaussian Processes for Machine Learning.
MIT Press.

Rokem, A., Watzl, S., Gollisch, T., Stemmler, M., Herz, A. V. M., Watzl, S., Gollisch, T., Stemmler, M., and Herz, A. V. M. (2006).
Spike-Timing Precision Underlies the Coding Efficiency of Auditory Receptor Neurons.
Journal of Neurophysiology, pages 2541–2552.

Algorithm 1 Blocked Gibbs sampler for GP-modulated renewal process on the interval $[0, T]$

Input: Set of event times G, set of thinned times $\tilde{G}_{prev}$ and l instantiated at $G \cup \tilde{G}_{prev}$.

Output: A new set of thinned times $\tilde{G}_{new}$ and a new instantiation $l_{G\cup\tilde{G}_{new}}$ of the $\mathcal{GP}$ on $G \cup \tilde{G}_{new}$.

1: Sample $A \subset [0, T]$ from a Poisson process with rate Ω.

2: Sample $l_A | l_{G\cup\tilde{G}_{prev}}$.

3: Thin A, keeping element $a \in A \cap [G_{i-1}, G_i]$ with probability $\left(1 - \frac{\hat{\lambda}\sigma(l(a))h(a-G_{i-1})}{\Omega}\right)$.

4: Let $\tilde{G}_{new}$ be the resulting set and $l_{\tilde{G}_{new}}$ be the restriction of l_A to this set. Discard $\tilde{G}_{prev}$ and $l_{\tilde{G}_{prev}}$.

5: Resample $l_{G\cup\tilde{G}_{new}}$ using, for example, elliptical slice sampling.