Markov chain Monte Carlo for Inference on Phase-type Models

Aslett, Louis J.M.
Trinity College, University of Dublin, Department of Statistics
College Green
Dublin, Ireland
E-mail: louis@maths.tcd.ie

Wilson, Simon P.
Trinity College, University of Dublin, Department of Statistics
College Green
Dublin, Ireland
E-mail: swilson@tcd.ie

Introduction

Phase-type (PHT) distributions provide a natural model for a wide range of stochastic processes where the event of interest is first passage time to a given state or set of states. However, they present interesting inferential challenges. Whilst both general likelihood and Bayesian approaches have been developed in the context of distribution fitting, there has been relatively little work on inference where there is a scientific interpretation of the underlying stochastic process. It is this latter situation which we address in the current work.

Our work builds on the Markov chain Monte Carlo (MCMC) algorithm developed by Bladt et al. (2003) in a number of directions. Firstly, in order to facilitate models in which the stochastic process has some scientifically interesting interpretation the conjugacy properties are shown to hold when constraints are imposed on the structure of the underlying Markov process, both in terms of prohibiting certain state transitions and of fixing certain rate parameters as being equal. The existing algorithm is also shown to naturally incorporate right-censored observations as can be common in real world temporal data.

These modifications can lead to computational issues in some situations and alleviating these in a wide class of situations is the second contribution of this work. The original algorithm is adapted to improve the step involving simulation of the Markov jump process underlying the PHT, which is required for each observation.

This paper provides a high-level summary of our work to this end.

Phase-type distributions

Consider a continuous-time Markov chain on a finite discrete state space of $n+1$ states, one of these states being absorbing. With a possible reordering of states, the generator of the Markov chain can be expressed as:

$$\mathbf{T} = \begin{pmatrix} \mathbf{S} & \mathbf{s} \\ \mathbf{0} & 0 \end{pmatrix}$$

where $\mathbf{S} = (S_{ij})$ is the matrix of transition rates between non-absorbing states i and j for $i \neq j$ and $i, j \in \{1, \ldots, n\}$, whilst $\mathbf{s} = (s_1, \ldots, s_n)^{\mathrm{T}}$ is the vector of transition rates from state i to the absorbing state and $\mathbf{0}$ is the row vector of n zeros.

A (continuous) phase-type distribution (PHT) is defined to be the distribution of the time to entering the absorbing state of a continuous-time Markov process with generator $\mathbf{T}$ and vector of initial state probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^{\mathrm{T}}$.

The distribution and density can be expressed as:

$$Y \sim \mathrm{PHT}(\boldsymbol{\pi}, \mathbf{T}) \implies \left\{ \begin{array}{rcl} F_Y(y) & = & 1 - \boldsymbol{\pi}^{\mathrm{T}} \exp\{y\mathbf{S}\}\mathbf{e} \\ f_Y(y) & = & \boldsymbol{\pi}^{\mathrm{T}} \exp\{y\mathbf{S}\}\mathbf{s} \end{array} \right.$$

where $\mathbf{e}$ is a vector of 1's of the appropriate dimension; $y \in [0, \infty)$ is the time to absorption; and $\exp\{y\mathbf{S}\}$ is the matrix exponential.

Here, we will refer to the stochastic process underlying a PHT random variable Y by $Y(t)$, representing the state of the process at time t. Thus, the process starts in state $Y(0)$, and given a realisation $Y_i = y_i$ of the PHT random variable we know that for the underlying process $Y(y_i) = n+1$.

A good reference for continuous time Markov chains is Grimmett & Stirzaker (2001). Phase-type distributions were introduced in Neuts (1975) and an excellent survey of much of the theory is presented by Asmussen (2000).

Bayesian inference

Bladt et al. (2003) provide a fully Bayesian framework for inference by constructing priors which are conjugate for the unobserved stochastic process. The unobserved stochastic process is simulated as part of a Metropolis-Hastings within Gibbs step. The focus within their work is very much on distribution fitting, with numerous examples of Phase-type approximations to positively supported theoretical densities.

We choose to focus on a Bayesian technique, since it enables the use of dispersed prior densities to inform the model and combat the non-identifiability issue that Phase-type distributions suffer in a frequentist inference context. Since our objective is learning something about the real parameters of the process and not distribution fitting, vague priors can guide the inference away from irrelevant local maxima.

To summarise, the MCMC algorithm developed by Bladt et al. (2003): if the full stochastic process leading to absorption is observed, then $\boldsymbol{\pi} \sim$ Dirichlet and $S_{ij}, s_i \sim$ Gamma are conjugate priors. A Metropolis-Hastings (MH) sampler is used to simulate the *unobserved* process (since here data consist only of absorption time). The MH proposal is a draw from $p(\text{path} \cdot \mid \boldsymbol{\pi}, \mathbf{S}, Y \geq y_i)$ performed by rejection sampling. The acceptance ratio then ensures that after a sufficient number of rejection samples one is sampling from $p(\text{path} \cdot \mid \boldsymbol{\pi}, \mathbf{S}, Y = y_i)$ after truncating to y_i and inserting an absorbing move. This full sample chain from the unobserved process in the MH step gives conjugacy for the Gibbs step. Thus, the algorithm iterates between drawing parameter values and simulating the unobserved processes associated with each absorption time:

$$\begin{tikzcd} p(\boldsymbol{\pi}, \mathbf{S} \mid \text{paths } \cdot, \mathbf{y}) \arrow[d, bend right] \\ p(\text{paths } \cdot \mid \boldsymbol{\pi}, \mathbf{S}, \mathbf{y}) \arrow[u, bend right] \end{tikzcd}$$

We now consider extensions to the methodology.

Censored data

Phase-type distributions measure time to an event of interest and it is not uncommon for such temporal data in scientific experiments to have right censoring. This can, for example, be the case when the event of interest has not occurred by the end of measurement or when some other independent event occurs which masks the event of interest such as so-called 'competing risks' (Prentice et al. 1978).

Such right-censored observations can be dealt with quite simply within the existing algorithm. The data will now consist of some actual and some censored observations, $\mathbf{y} = \{y_1, \dots, y_m, y^{\mathrm{c}}_{m+1}, \dots, y^{\mathrm{c}}_n\}$, where a superscript 'c' denotes right-censoring. Now, when the MH step is performed, it should be used as normal for $\{y_1, \dots, y_m\}$, but for $\{y^{\mathrm{c}}_{m+1}, \dots, y^{\mathrm{c}}_n\}$ it should cease immediately after producing a sample from the rejection sampling. This allows us to effectively incorporate the fact that the unit survived to this time in the likelihood through the Metropolis-Hastings step.

Special structures

As it stands, the algorithm assumes a rate matrix for the phase-type distribution which is completely dense and where every rate can vary independently in the inferential process. However, when the objective is scientific modelling of a process about which there is some known structure or theory this may not be a desirable property. In particular, certain state transitions of the process being modelled may make no physical sense and so their entry must be fixed at zero in the generator of the Markov chain. Additionally, it may be that certain state transitions should (in an idealised modelling sense) have identical parameter values. These desirable modelling assumptions moreover allow a significant reduction in the dimension of the parameter space.

Constraining a parameter to zero is a straight-forward procedure as was the case for the EM-algorithm of Asmussen et al. (1996). The parameter is simply fixed at zero when simulating chains in the MH step and no new values are simulated in the Gibbs step: this solves the problem because we are ceasing to consider this entry of $\mathbf{T}$ as a parameter, excluding it from all inferential procedures.

The second matter – that of constrained equality for certain parameters – requires examination of the posterior distribution which, when using the conjugate priors above and with the full data (full Markov processes, simulated by Metropolis-Hastings here), can be written as (Bladt et al. 2003):

$$
\begin{aligned}
p(\boldsymbol{\pi}, \mathbf{T} \,|\, \boldsymbol{y}) &\propto \phi(\boldsymbol{\pi}, \mathbf{T}) p(\mathbf{y} \,|\, \boldsymbol{\pi}, \mathbf{T}) \\
&= \left(\prod_{i=1}^{p} \pi_i^{\beta_i - 1} \prod_{i=1}^{p} s_i^{\nu_{i0}-1} e^{-s_i \zeta_i} \prod_{i=1}^{p} \prod_{\substack{j=1 \\ j \neq i}}^{p} S_{ij}^{\nu_{ij}-1} e^{-S_{ij}\zeta_i} \right) \\
&\qquad \times \left(\prod_{i=1}^{p} \pi_i^{B_i} \prod_{i=1}^{p} s_i^{N_{i0}} e^{-s_i Z_i} \prod_{i=1}^{p} \prod_{\substack{j=1 \\ j \neq i}}^{p} S_{ij}^{N_{ij}} e^{-S_{ij} Z_i} \right) \\
&= \prod_{i=1}^{p} \pi_i^{B_i + \beta_i - 1} \prod_{i=1}^{p} s_i^{N_{i0} + \nu_{i0} - 1} e^{-s_i(\zeta_i + Z_i)} \prod_{i=1}^{p} \prod_{\substack{j=1 \\ j \neq i}}^{p} S_{ij}^{N_{ij} + \nu_{ij} - 1} e^{-S_{ij}(Z_i + \zeta_i)}
\end{aligned}
$$

Thus, where we constrain parameters in different parts of the generator $\mathbf{S}$ to be equal, it is possible to ensure the resultant full conditional posteriors still keep the property of conjugacy. However, to do so requires a reduction of the flexibility in specification of the prior distributions. When every entry in the rate matrix is a freely varying parameter, we have the prior specified as $S_{ij} \sim \text{Gamma}(\nu_{ij}, \zeta_i)$ and $s_i \sim \text{Gamma}(\nu_{i0}, \zeta_i)$. Clearly, if we constrain certain S_{ij} to be equal, so that $S_{ij} = \lambda$ for all pairs $(i, j) \in A$ then we must have $\nu_{ij} = \nu_\lambda \; \forall \, (i, j) \in A$ and $\zeta_i = \zeta_\lambda \; \forall \, (i, \cdot) \in A$. The restriction on $\nu_{\cdot\cdot}$ is of no consequence, but the restriction on $\zeta_\cdot$ is important since ζ_i is a common prior parameter for all $S_{i\cdot}$ and s_i

In effect, if a constraint crosses rows in a matrix then *all* parameters in such rows will be constrained to sharing a common parameter ζ. The restriction is not too grave given that there is full freedom in the assignment of ν_{ij} still, but it may prompt even more careful consideration of prior parameter specification.

Computational tractability

The key focus of our work relates to the performance of the algorithm for practical use in modelling. Whilst the original unmodified algorithm had one major source of performance related issues, the expanded modelling options introduced in the previous section can, in certain circumstances, lead to additional severe degradation in the computational performance of the algorithm.

Issue I [Exploring parameter space]: It is common in scientific studies which call upon the Bayesian method of inference to desire use of a relatively diffuse prior density. Additionally, it is desirable for an MCMC sampling scheme to explore the parameter space well (ie the space of vectors and matricies $(\boldsymbol{\pi}, \mathbf{S})$ well here). However, the rejection sampling prescription of the MH step can encounter problems if this is the case.

This can best be seen by considering the acceptance probability for the rejection sampler. Since the rejection sampling is simply forward simulation of an absorbing continuous-time Markov chain, rejecting sample paths that do not reach the observation time y_i before absorbing, the rejection probability is simply $P(Y < y_i \,|\, \boldsymbol{\pi}, \mathbf{S})$. Clearly the density of number of trials required to draw a suitably long simulated chain is Geometric, with parameter $p = \boldsymbol{\pi}^{\mathrm{T}} \exp\{y_i \mathbf{S}\}\mathbf{e}$

So, algorithm 1 will from time-to-time be expected to simulate chains which are highly improbable under the given parameter values. Thus, the rejection step of algorithm 1 can completely stall because it is so unlikely that any chain will be produced which absorbs so far into the tail. For example,

$$\begin{aligned} P(Y > y_i \,|\, \boldsymbol{\pi}, \mathbf{S}) = 0.001 \quad &\Longrightarrow \mathbb{E}(\text{Rej Samp iter}) = 1000, \qquad 95\% \text{ CI} = [25, 3687] \\ P(Y > y_i \,|\, \boldsymbol{\pi}, \mathbf{S}) = 10^{-6} \quad &\Longrightarrow \mathbb{E}(\text{Rej Samp iter}) = 1000000, \;\; 95\% \text{ CI} = [25317, 3688877] \end{aligned}$$

Thus, chains are being wastefully sampled, potentially millions of times, to find one chain absorbing beyond y_i. It is not hard to see that for diffuse priors the above problem will effectively stall the algorithm, particularly since this could conceivably arise for all observations for a given data set.

Issue II [Zero constraints for absorbing moves]: Once we admit the possibility of constraining some of the rates of moves to absorption, s_i, to zero, there is the possibility that truncating a rejection sampled chain will produce an impossible move. For example, if $s_j = 0$ is a constraint and our rejection sample is such that $Y(y_i) = j$, then attempting to truncate and insert a move $j \to n+1$ will be invalid. Of course, this is dealt with automatically insofar the MH acceptance ratio has probability zero of accepting such a chain.

However, it can cause serious computational issues when the states most commonly occupied by the process are those from which absorbing moves are disallowed. A simple example we have encountered when modelling a repairable redundant electronic system comprising of reliable components exhibited such behaviour. Here, a move 1 → absorption is disallowed as it represents the probability zero event of precisely identical failure times for two independent units:

State, j	**Meaning**	$P(Y(y_i) = j)$
1	both units working	0.9986
2	1 failed, 2 working	0.0007
3	1 working, 2 failed	0.0007

This leads us to conclude that there will be a large number of unusable chain paths produced by rejection sampling under the extended methodology which enables disallowed state moves. In this simple example:

$$\mathbb{E}(\text{no. unusable MH results}) = 1429 \quad \text{and} \quad 95\% \text{ CI} = [36, 5267]$$

Issues I and II can compound and we have often observed the algorithm effectively stalling for days of computation time on a single iteration.

Explicit conditional sampling

We propose replacing the Metropolis-Hastings step entirely by constructing an algorithm which samples a chain explicitly conditional on the time an absorbing move occurs.

The new algorithm produces a sample chain from $p(\text{path} \cdot \,|\, \boldsymbol{\pi}, \mathbf{S}, Y = y)$ as follows (we suppress $\boldsymbol{\pi}, \mathbf{S}$ for readability):

1. Choose the starting state j from the discrete distribution defined by $\boldsymbol{\pi}$ and set the current 'clock time' $t = 0$.

2. Decide if an additional state move occurs before absorption or whether the next move is the absorbing move. To do this, draw $U \sim \text{Unif}(0, 1)$ and remain in j to absorption if

   $$U < P(\delta > y - t \cap j \to n + 1 \,|\, Y(t) = j, \boldsymbol{\pi}, \mathbf{S}, Y = y) = \frac{e^{S_{jj}(y-t)} s_j}{\mathbf{1}_j^{\mathrm{T}} e^{\mathbf{S}(y-t)} \mathbf{s}}$$

   where $\mathbf{1}_j$ hereon denotes a vector with 1 in the j^{th} slot and 0 elsewhere. Otherwise, continue to step 3:

3. Here, a state move must occur between t and absorption at y. Select the time of the next jump according to the density (here $\mathbf{p}_{j\cdot}$ are the jump probabilities from j excluding absorption, $\frac{S_{ji}}{-S_{jj} - s_j}$):

   $$p(\delta = d \,|\, \delta < (y - t), Y(t) = j, \boldsymbol{\pi}, \mathbf{S}, Y = y) \propto -S_{jj} e^{S_{jj} d} \; \mathbf{p}_{j\cdot}^{\mathrm{T}} e^{\mathbf{S}(y-t-d)} \mathbf{s}$$

4. The non-absorbing move is then chosen from the probability mass function ($j \neq i$):

   $$P(j \to i \,|\, \delta = d < (y - t), Y(t) = j, \boldsymbol{\pi}, \mathbf{S}, Y = y) \propto S_{ji} \; \mathbf{1}_i^{\mathrm{T}} e^{\mathbf{S}(y-t-d)} \mathbf{s}$$

5. Update $j = i$ and $t = t + d$, then loop to 2.

The calculations in steps 2 and 4 are routine save for the usual complication associated with computing matrix exponentials. Step 3 is somewhat more involved, though it is possible to attack the distribution function for the jump time d analytically (numerical stability is improved if the rate matrix is eigendecomposed as $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$)

$$F(d) = \frac{\displaystyle\int_0^d -S_{jj} e^{S_{jj}x} \; \mathbf{p}_{j\cdot}^{\mathrm{T}} e^{\mathbf{S}(y-t-x)} \mathbf{s} \, dx}{\displaystyle\int_0^{y-t} -S_{jj} e^{S_{jj}x} \; \mathbf{p}_{j\cdot}^{\mathrm{T}} e^{\mathbf{S}(y-t-x)} \mathbf{s} \, dx} = \frac{-S_{jj} \mathbf{p}_{j\cdot}^{\mathrm{T}} \mathbf{Q} e^{\mathbf{\Lambda}(y-t)} (-S_{jj}\mathbf{I} - \mathbf{\Lambda})^{-1} \left\{ \mathbf{I} - e^{-d(-S_{jj}\mathbf{I} - \mathbf{\Lambda})} \right\} \mathbf{Q}^{-1} \mathbf{s}}{-S_{jj} \mathbf{p}_{j\cdot}^{\mathrm{T}} \mathbf{Q} e^{\mathbf{\Lambda}(y-t)} (-S_{jj}\mathbf{I} - \mathbf{\Lambda})^{-1} \left\{ \mathbf{I} - e^{-d(y-t)(-S_{jj}\mathbf{I} - \mathbf{\Lambda})} \right\} \mathbf{Q}^{-1} \mathbf{s}}$$

Note that much of the above (including eigendecomposition) is unchanged on repeated evaluation for different d, so the overhead is much less than it may at first seem. Random variates can be generated from this distribution using a numerical root solver to invert a Uniform(0,1) random number.

We also developed this conditioning instead explicitly on $Y_i \geq y_i$ for censored data.

Conclusions and future work

The methodological changes enable application of Bayesian inference for phase-type models to a wider class of problems than was previously possible, including those where there is a physical meaning to the underlying stochastic process. The replacement of the Metropolis-Hastings step also reduces the number of situations in which the MCMC scheme will stall during sampling. Work is currently focused on the most efficient techniques for sampling step 3 in the new algorithm.

Funding

Louis Aslett is supported by a PhD scholarship from the Irish Research Council for Science, Engineering and Technology (IRCSET).

REFERENCES

Asmussen, S. (2000), ‘Matrix-analytic models and their analysis’, *Scandinavian Journal of Statistics* **27**(2), 193–226.

Asmussen, S., Nerman, O. & Olsson, M. (1996), ‘Fitting phase-type distributions via the EM algorithm’, *Scandinavian Journal of Statistics* **23**(4), 419–441.

Bladt, M., Gonzalez, A. & Lauritzen, S. L. (2003), ‘The estimation of phase-type related functionals using Markov chain Monte Carlo methods’, *Scandinavian Journal of Statistics* **2003**(4), 280–300.

Grimmett, G. R. & Stirzaker, D. R. (2001), *Probability and Random Processes*, 3rd edn, Oxford University Press.

Neuts, M. F. (1975), ‘Probability distributions of phase type’, *Liber Amicorum Prof. Emeritus H. Florin, Dept. Math, Univ. Louvain, Belgium* pp. 173–206.

Prentice, R. L., Kalbfleisch, J. D., A. V. Peterson, J., Flournoy, N., Farewell, V. T. & Breslow, N. E. (1978), ‘The analysis of exponentially distributed life-times with two types’, *Biometrics* **34**(4), 541–554.