Markov-chain Monte Carlo for Phase-type Models

Louis JM Aslett and Simon P Wilson

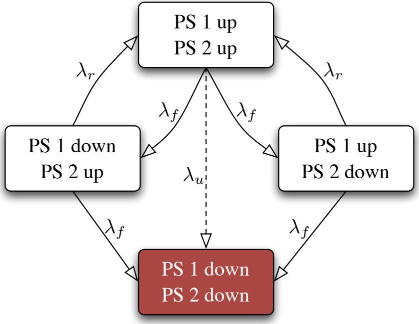
Trinity College, University of Dublin

September 2011

Trinity College Dublin

Toy Example : Redundant Repairable Components

State	Meaning
1	both PS working
2	1 failed, 2 working
3	1 working, 2 failed
4	subsystem failed



$$\Longrightarrow \mathbf{T} = \begin{pmatrix} -2\lambda_\mathrm{f} & \lambda_\mathrm{f} & \lambda_\mathrm{f} & 0 \\ \lambda_\mathrm{r} & -\lambda_\mathrm{r} - \lambda_\mathrm{f} & 0 & \lambda_\mathrm{f} \\ \lambda_\mathrm{r} & 0 & -\lambda_\mathrm{r} - \lambda_\mathrm{f} & \lambda_\mathrm{f} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Trinity College Dublin

Inferential Setting

Cano & Rios (2006) provide conjugate posterior calculations in the context of analysing repairable systems when stochastic process leading to absorption is observed.

Data

For each absorption time, one has:

- Starting state
- Length of time in each state
- Number of transitions between each state
- Ultimate absorption time

Inferential Setting

Cano & Rios (2006) provide conjugate posterior calculations in the context of analysing repairable systems when stochastic process leading to absorption is observed.

Data

For each absorption time, one has:

- Starting state
- ~~Length of time in each state~~
- ~~Number of transitions between each state~~
- Ultimate absorption time

Model by a Phase-type distribution $\Longrightarrow$ Bladt et al. (2003) provides a Bayesian MCMC algorithm.

Definition of Phase-type Distributions

An absorbing continuous time Markov chain is one in which there is a state that, once entered, is never left. That is, the $n+1$ state intensity matrix can be written:

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

where $\mathbf{S}$ is $n \times n$, $\mathbf{s}$ is $n \times 1$ and $\mathbf{0}$ is $1 \times n$, with

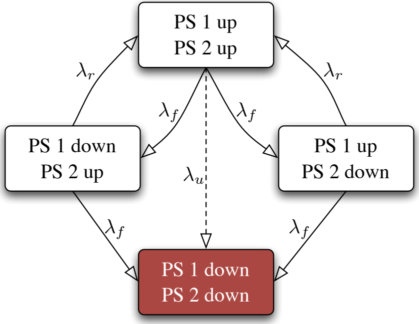
$$\mathbf{s} = -\mathbf{S}\mathbf{e}$$

Then, a *Phase-type distribution* (PHT) is defined to be the distribution of the time to entering the absorbing state.

$$Y \sim \text{PHT}(\boldsymbol{\pi}, \mathbf{S}) \implies \begin{cases} F_Y(y) = 1 - \boldsymbol{\pi}^{\text{T}} \exp\{y\mathbf{S}\}\mathbf{e} \\ f_Y(y) = \boldsymbol{\pi}^{\text{T}} \exp\{y\mathbf{S}\}\mathbf{s} \end{cases}$$

Example

State	Meaning
1	both PS working
2	1 failed, 2 working
3	1 working, 2 failed
4	subsystem failed

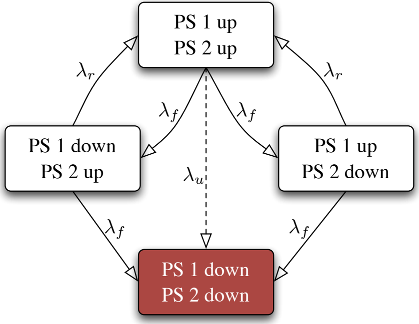


$$\Longrightarrow \mathbf{T} = \begin{pmatrix} -2\lambda_{\mathrm{f}} & \lambda_{\mathrm{f}} & \lambda_{\mathrm{f}} & 0 \\ \lambda_{\mathrm{r}} & -\lambda_{\mathrm{r}} - \lambda_{\mathrm{f}} & 0 & \lambda_{\mathrm{f}} \\ \lambda_{\mathrm{r}} & 0 & -\lambda_{\mathrm{r}} - \lambda_{\mathrm{f}} & \lambda_{\mathrm{f}} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$f_Y(y) = \boldsymbol{\pi}^{\mathrm{T}} \exp\{y\mathbf{S}\}\mathbf{s} \qquad F_Y(y) = 1 - \boldsymbol{\pi}^{\mathrm{T}} \exp\{y\mathbf{S}\}\mathbf{e}$$

Example

State	Meaning
1	both PS working
2	1 failed, 2 working
3	1 working, 2 failed
4	subsystem failed



$$\Longrightarrow \mathbf{T} = \left(\begin{array}{ccc|c} -2\lambda_f & \lambda_f & \lambda_f & 0 \\ \lambda_r & -\lambda_r - \lambda_f & 0 & \lambda_f \\ \lambda_r & 0 & -\lambda_r - \lambda_f & \lambda_f \\ \hline 0 & 0 & 0 & 0 \end{array}\right)$$

$\mathbf{S}$ $\quad$ $\mathbf{s}$

$$f_Y(y) = \boldsymbol{\pi}^{\mathrm{T}} \exp\{y\mathbf{S}\}\mathbf{s} \qquad F_Y(y) = 1 - \boldsymbol{\pi}^{\mathrm{T}} \exp\{y\mathbf{S}\}\mathbf{e}$$

Algorithm 1: Metropolis-Hastings Simulation of Process

In summary:

- Can simulate chain from

$$\mathbb{P}(\text{path } \cdot \,|\, Y_i \geq y_i)$$

  trivially by rejection sampling.
- A Metropolis-Hastings acceptance ratio (ratio of exit rates) exists such that for sufficient draws, the final chain when truncated to time y_i (at which point it absorbs) will be a draw from

$$\mathbb{P}(\text{path } \cdot \,|\, Y_i = y_i)$$

Algorithm 2: Gibbs Sampling from Posterior

The Gibbs step achieves the goal of simulating from

$$p(\boldsymbol{\pi}, \mathbf{S} \,|\, \mathbf{y})$$

by sampling from

$$p(\boldsymbol{\pi}, \mathbf{S}, \text{paths } \cdot \,|\, \mathbf{y})$$

through the iterative process

$$p(\boldsymbol{\pi}, \mathbf{S} \,|\, \text{paths } \cdot, \mathbf{y})$$

$$p(\text{paths } \cdot \,|\, \boldsymbol{\pi}, \mathbf{S}, \mathbf{y})$$

Motivation for Modifications

1. Certain state transitions make no physical sense. (eg $2 \to 3$ in earlier example)

State	Meaning
1	both PS working
2	1 failed, 2 working
3	1 working, 2 failed
4	subsystem failed

$$\implies \mathbf{T} = \begin{pmatrix} -2\lambda_\mathrm{f} & \lambda_\mathrm{f} & \lambda_\mathrm{f} & 0 \\ \lambda_\mathrm{r} & -\lambda_\mathrm{r} - \lambda_\mathrm{f} & 0 & \lambda_\mathrm{f} \\ \lambda_\mathrm{r} & 0 & -\lambda_\mathrm{r} - \lambda_\mathrm{f} & \lambda_\mathrm{f} \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Motivation for Modifications

1. Certain state transitions make no physical sense. (eg $2 \to 3$ in earlier example)

2. When part of a larger system, it is highly likely there will be censored observations.

Motivation for Modifications

1. Certain state transitions make no physical sense. (eg $2 \rightarrow 3$ in earlier example)
2. When part of a larger system, it is highly likely there will be censored observations.
3. Where there is no reason to believe distributional differences between parameters, they should (in ideal sense) be constrained to be equal. This is as much to assist with reducing parameter dimensionality.

Motivation for Modifications

1. Certain state transitions make no physical sense. (eg $2 \longrightarrow 3$ in earlier example)
2. When part of a larger system, it is highly likely there will be censored observations.
3. Where there is no reason to believe distributional differences between parameters, they should (in ideal sense) be constrained to be equal. This is as much to assist with reducing parameter dimensionality.
4. Computation time!

Statistical -vs- Stochastic

In other words, we want an MCMC algorithm fit for performing inference when PHTs used for stochastic rather than statistical modelling.

Stochastic Model

"Stochastic models seek to represent an underlying physical phenomenon of interest, albeit often in a highly idealised way, and have parameters that are physically interpretable." — Isham

Statistical Model

"In contrast, statistical models are descriptive, and represent the statistical properties of data and their dependence on covariates, without aiming to encapsulate the physical mechanisms involved." — Isham

Statistical -vs- Stochastic

In other words, we want an MCMC algorithm fit for performing inference when PHTs used for stochastic rather than statistical modelling.

Stochastic Model $\longrightarrow$ Aslett & Wilson

"Stochastic models seek to represent an underlying physical phenomenon of interest, albeit often in a highly idealised way, and have parameters that are physically interpretable." — Isham

Statistical Model $\longrightarrow$ Bladt et al

"In contrast, statistical models are descriptive, and represent the statistical properties of data and their dependence on covariates, without aiming to encapsulate the physical mechanisms involved." — Isham

Solving I-III

1. To prohibit a transition $i \longrightarrow j$,
   - Do not draw S_{ij} from the prior in algorithm 2.
   - Fix S_{ij} at zero when running algorithm 1.
   - Do not draw values for S_{ij} from the full conditional posterior in algorithm 2.
2. To account for censoring:
   - In algorithm 1, if y_i is a censored observation, return immediately once rejection sampling from

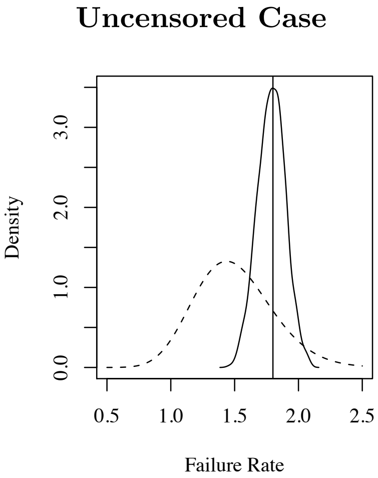
     $$\mathbb{P}(\text{path } \cdot \,|\, Y_i \geq y_i)$$

     is complete, ignoring the MH steps. Else, use algorithm 1 as normal.
3. Some routine algebra shows conjugacy can be maintained with constraints, though naturally the parameter updating changes.

Toy Example Results

100 uncensored observations simulated from PHT with

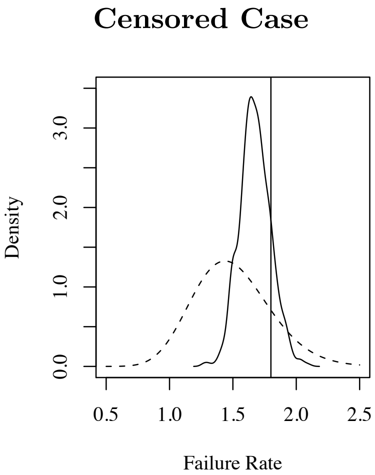
$$\mathbf{S} = \begin{pmatrix} -3.6 & 1.8 & 1.8 \\ 9.5 & -11.3 & 0 \\ 9.5 & 0 & -11.3 \end{pmatrix}$$



Toy Example Results

100 (25%) censored observations simulated from PHT with

$$\mathbf{S} = \begin{pmatrix} -3.6 & 1.8 & 1.8 \\ 9.5 & -11.3 & 0 \\ 9.5 & 0 & -11.3 \end{pmatrix}$$



Trinity College Dublin

The Big Issue

Intractable computation time for many applications!

The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.

The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.

$$\mathbb{P}(Y_i \geq y_i \,|\, \text{true params}) = 0.01$$

$$p(\boldsymbol{\pi}, \mathbf{S} \,|\, \text{paths } \cdot, \mathbf{y})$$

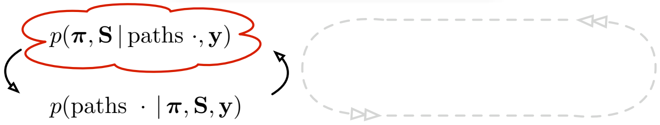
$$p(\text{paths } \cdot \,|\, \boldsymbol{\pi}, \mathbf{S}, \mathbf{y})$$

The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.

$$\mathbb{P}(Y_i \geq y_i \,|\, \boldsymbol{\pi}, \mathbf{S}) = 0.09$$



The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.

$$\mathbb{P}(Y_i \geq y_i \,|\, \boldsymbol{\pi}, \mathbf{S}) = 0.09$$

$$p(\boldsymbol{\pi}, \mathbf{S} \,|\, \text{paths } \cdot, \mathbf{y})$$

$$p(\text{paths } \cdot \,|\, \boldsymbol{\pi}, \mathbf{S}, \mathbf{y})$$

$$\mathbb{E}(\text{RS iter.}) = 11$$

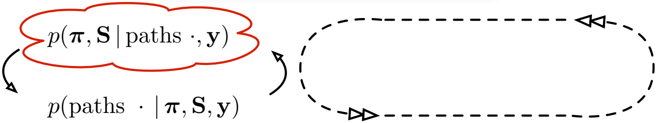
$$95\% \text{ CI} = [1, 39]$$

The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.

$$\mathbb{P}(Y_i \geq y_i \,|\, \boldsymbol{\pi}, \mathbf{S}) = 0.001$$



The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.

$$\mathbb{P}(Y_i \geq y_i \,|\, \boldsymbol{\pi}, \mathbf{S}) = 0.001$$

$p(\boldsymbol{\pi}, \mathbf{S} \,|\, \text{paths } \cdot, \mathbf{y})$

$p(\text{paths } \cdot \,|\, \boldsymbol{\pi}, \mathbf{S}, \mathbf{y})$

$\mathbb{E}(\text{RS iter.}) = 1000$

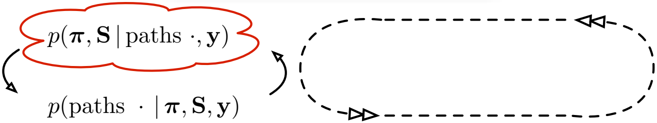
95% CI $= [25, 3687]$

The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.

$$\mathbb{P}(Y_i \geq y_i \,|\, \boldsymbol{\pi}, \mathbf{S}) = 10^{-6}$$

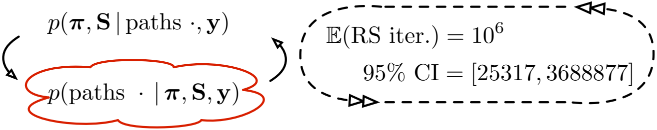


The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.

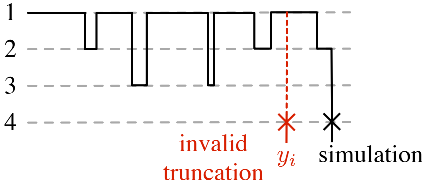
$$\mathbb{P}(Y_i \geq y_i \mid \boldsymbol{\pi}, \mathbf{S}) = 10^{-6}$$



The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.
2. states from which absorption impossible – wasteful to resample whole chain because state at time y_i unsuitable for truncation.



The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.
2. states from which absorption impossible – wasteful to resample whole chain because state at time y_i unsuitable for truncation.

State	Meaning	$\mathbb{P}(\text{state})$
1	both PS working	0.9986
2	1 failed, 2 working	0.0007
3	1 working, 2 failed	0.0007

$\Longrightarrow$ $\mathbb{E}(\text{MH iter}) = 1429$

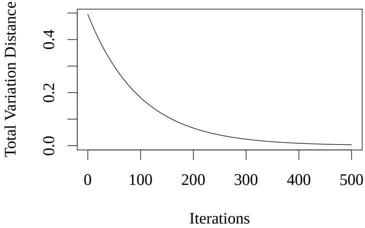
$95\%\ \text{CI} = [36, 5267]$

Trinity College Dublin

The Big Issue

Intractable computation time for many applications!

1. longer chains and MCMC jumps to states for which observations are far in the tails can stall rejection sampling step of MH algorithm.
2. states from which absorption impossible – wasteful to resample whole chain because state at time y_i unsuitable for truncation.
3. time for MH algorithm to reach stationarity can grow rapidly.



Approach: Exact Conditional Sampling

Replace rejection sampling + MH with exact conditional sampling.

- Sample a starting state, i, from the probability mass function:

$$\mathbb{P}(Y\{0\} = i \,|\, \boldsymbol{\pi}, \mathbf{S}, Y = y) = \frac{\mathbf{e}_i^{\mathrm{T}} \exp\{\mathbf{S}y\}\mathbf{s}\ \pi_i}{\boldsymbol{\pi}^{\mathrm{T}} \exp\{\mathbf{S}y\}\mathbf{s}}$$

  and set $t = 0$

Approach: Exact Conditional Sampling

Replace rejection sampling + MH with exact conditional sampling.

- Sample a starting state, i, from the probability mass function:

$$\mathbb{P}(Y\{0\} = i \,|\, \boldsymbol{\pi}, \mathbf{S}, Y = y) = \frac{\mathbf{e}_i^{\mathrm{T}} \exp\{\mathbf{S}y\}\mathbf{s}\ \pi_i}{\boldsymbol{\pi}^{\mathrm{T}} \exp\{\mathbf{S}y\}\mathbf{s}}$$

  and set $t = 0$
- With probability

$$\mathbb{P}(Y[t,y) = i \cap Y\{y\} = n+1 \,|\, \mathbf{S}, Y = y, Y\{t\} = i) = \frac{\exp\{S_{ii}(y-t)\}\ s_i}{\mathbf{e}_i^{\mathrm{T}} \exp\{\mathbf{S}(y-t)\}\mathbf{s}}$$

  set $Y[t,y) = i$ and $Y\{y\} = n+1$ and end the algorithm; else continue

Approach: Exact Conditional Sampling

- Sample the sojourn time in the current state, δ, before a non-absorbing move from

$$p(\delta = d \,|\, \mathbf{S}, Y = y, Y[t, t+\delta) = i, Y\{t+\delta\} \in \{1, \dots, n\} \backslash i) = \frac{\mathbf{p}_{i\cdot}^{\mathrm{T}} \exp\{\mathbf{S}(y-t-d)\}\mathbf{s}\ (-S_{ii}) \exp(S_{ii} d)}{\int_0^{y-t} \mathbf{p}_{i\cdot}^{\mathrm{T}} \exp\{\mathbf{S}(y-t-\delta)\}\mathbf{s}\ (-S_{ii}) \exp(S_{ii}\delta)\, d\delta}$$

and set $Y[t, t+d) = i$

Approach: Exact Conditional Sampling

- Sample the sojourn time in the current state, δ, before a non-absorbing move from

$$p(\delta = d \,|\, \mathbf{S}, Y = y, Y[t, t+\delta) = i, Y\{t+\delta\} \in \{1, \dots, n\} \backslash i) = \frac{\mathbf{p}_{i\cdot}^{\mathrm{T}} \exp\{\mathbf{S}(y-t-d)\}\mathbf{s}\;(-S_{ii}) \exp(S_{ii} d)}{\int_0^{y-t} \mathbf{p}_{i\cdot}^{\mathrm{T}} \exp\{\mathbf{S}(y-t-\delta)\}\mathbf{s}\;(-S_{ii}) \exp(S_{ii}\delta)\, d\delta}$$

  and set $Y[t, t+d) = i$

- Sample a state move, $i \to j$, from

$$\mathbb{P}(Y\{t+d\} = k \,|\, \mathbf{S}, Y = y, Y[t, t+d) = i, Y\{t+d\} \in \{1, \dots, n\} \backslash i) \propto S_{ik} \exp\{\mathbf{S}(y-t-d)\}\mathbf{s}$$

  and set $Y\{t+d\} = j$

Trinity College Dublin

Approach: Exact Conditional Sampling

- Sample the sojourn time in the current state, δ, before a non-absorbing move from

$$p(\delta = d \,|\, \mathbf{S}, Y = y, Y[t, t+\delta) = i, Y\{t+\delta\} \in \{1, \dots, n\} \backslash i) = \frac{\mathbf{p}_{i\cdot}^{\mathrm{T}} \exp\{\mathbf{S}(y-t-d)\}\mathbf{s}\ (-S_{ii}) \exp(S_{ii} d)}{\int_0^{y-t} \mathbf{p}_{i\cdot}^{\mathrm{T}} \exp\{\mathbf{S}(y-t-\delta)\}\mathbf{s}\ (-S_{ii}) \exp(S_{ii}\delta)\, d\delta}$$

  and set $Y[t, t+d) = i$

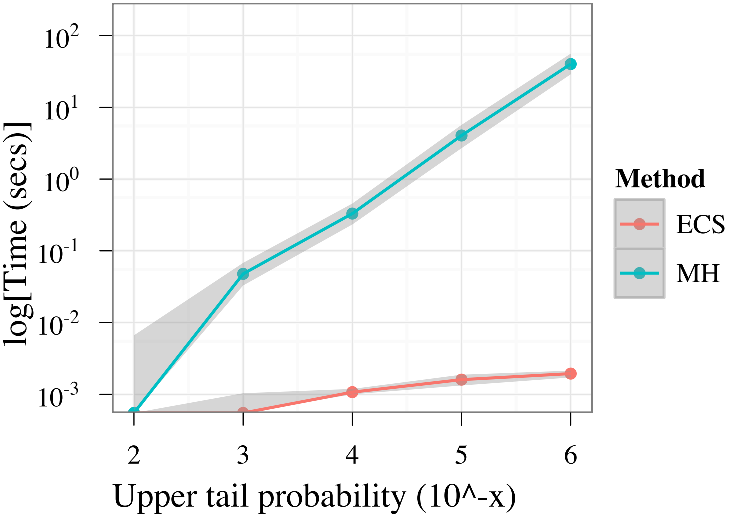
- Sample a state move, $i \to j$, from

$$\mathbb{P}(Y\{t+d\} = k \,|\, \mathbf{S}, Y = y, Y[t, t+d) = i, Y\{t+d\} \in \{1, \dots, n\} \backslash i) \propto S_{ik} \exp\{\mathbf{S}(y-t-d)\}\mathbf{s}$$

  and set $Y\{t+d\} = j$

- Update $t = t + d$ and $i = j$, then loop to second step

Tail Depth Performance Improvement



Trinity College Dublin

Overall Performance Improvement

This shows the new method keeping pace in 'nice' problems and significantly outperforming otherwise.

$$\mathbf{T} = \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

No problems i-iii

	MH	ECS
$\bar{t}$	1.6 µs	7.2 µs
s_t	104 µs	19 µs

$$\mathbf{T} = \begin{pmatrix} -2 & 0.01 & 1.99 & 0 \\ 1 & -300 & 0 & 299 \\ 299 & 0 & -300 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

All problems i-iii

	MH	ECS
$\bar{t}$	10.2 hours	0.016 secs
s_t	9.4 hours	0.015 secs

Overall Performance Improvement

This shows the new method keeping pace in 'nice' problems and significantly outperforming otherwise.

$\mathbf{T} = \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$			$\mathbf{T} = \begin{pmatrix} -2 & 0.01 & 1.99 & 0 \\ 1 & -300 & 0 & 299 \\ 299 & 0 & -300 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	
No problems i-iii			All problems i-iii	
	MH	ECS	MH	ECS
$\bar{t}$	1.6 µs	7.2 µs	10.2 hours	0.016 secs
s_t	104 µs	19 µs	9.4 hours	0.015 secs

2,300,000 × faster on average in hard problem

Trinity College Dublin

Future Work

- At the moment, further speedup simulating the latent process might only be possible though extra efficiencies in computing the matrix exponential. Tough problem, see Moler & Van Loan (2003).
- Beyond this, of interest is whether it is possible to ascertain exact/approximate distributions for the sufficient statistics:

$\mathbf{N} = N_{ij}$	matrix of no. transitions $i \to j$
$\mathbf{z} = z_i$	vector of total time spent in state i
$\mathbf{B} = B_i$	vector of no. times started in state i

  of a CTMC, given the generator matrix. This may allow these to be sampled directly rather than indirectly as here?
- Can we combat the creep of increasing autocorrelation for these constrained models?

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.

Cano, J. & Rios, D. (2006), 'Reliability forecasting in complex hardware/software systems', *Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06)* .

Moler, C. & Van Loan, C. (2003), 'Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later', *SIAM Review* **45**(1), 3–49.