1. `ptmc` R package performance

1.1 Background

Mathematically modelling the propagation of infectious diseases often involves solving complex non-linear SIR-type models. The evolution of these models is dictated by parameters values which are often difficult to accurately estimate due to biases in existing studies or surveillance data, leading to uncertainty in their estimates. The desire to formally incorporate such parameter uncertainty into epidemic models naturally leads to the use of Bayesian inference. However, combining Bayesian inference with complex non-linear systems forms intractable integrals which require Markov Chain Monte Carlo (MCMC) methods to solve. Given the rich, dense literature on MCMC samplers, it is often not clear which MCMC sampler is optimal (i.e. provides convincing convergence over a small time frame) for SIR-type infectious diseases models.

This analysis explores this issue, by comparing several popular MCMC samplers with a promising population-based MCMC sampler (parallel tempering), to sample from the posterior distribution for parameters in an age- and exposure-stratified epidemic model for Respiratory Syncytial Virus. The package finds that the parallel tempering MCMC sampler is quicker and has improved mixing compared to other popular MCMC samplers in existing packages. Therefore, it is encouraged that mathematical modellers who require MCMC to calibrate epidemic models to use parallel tempering or explore similar population based samplers to ensure quick and convincing convergence.

1.2 Methods

1.2.1 Description of epidemic model

I modelled the number of individuals in six different epidemiological states, S_0, I_0, R_0, S_1, I_1, and R_0, where S, I, and R refer to susceptible, infectious and recovered individuals respectively. The subscript 0 relates to primary infection, and subscript 1 refers to secondary and subsequent infections. On average, the infectious period lasts for $1/\gamma$ days and the duration of immunity lasts $1/\omega$ days. The epidemiological states are stratified into two different age groups (0–2 years, and 2+ years), denoted by superscript a. To ensure the population of each age group remains constant, I assume that $\mu = 1861$ live births occur each day into the 0-2 years age group and that the same number of persons die each day from the 2 years and older group. The rate at which individuals acquire infection, λ_i^a depends on contacts between age groups and the number infectious persons in each age group. Thus, the RSV epidemic model is specified in a continuous time system for age group a is given by:

$$
\begin{array}{rcll}
\frac{dS_0^a}{dt} & = & \overbrace{\mu \mathbb{1}_1(a) - \lambda^a S_0^a}^{\text{Transmission terms}} & \overbrace{-\eta^a S_0^a + \eta^{a-1} S_0^{a-1}}^{\text{Ageing terms}} \\
\frac{dI_0^a}{dt} & = & \lambda^a S_0^a - \gamma I_0^a & -\eta^a I_0^a + \eta^{a-1} I_0^{a-1} \\
\frac{dR_0^a}{dt} & = & \gamma I_0^a - \omega R_0^a & -\eta^a R_0^a + \eta^{a-1} R_0^{a-1} \\
\frac{dS_1^a}{dt} & = & \omega R_0^a + \omega R_1^a - \delta_1 \lambda^a S_1^a & -\eta^a S_1^a + \eta^{a-1} S_1^{a-1} \\
\frac{dI_1^a}{dt} & = & \delta_1 \lambda^a S_1^a - \gamma I_1^a & -\eta^a I_1^a + \eta^{a-1} I_1^{a-1} \\
\frac{dR_1^a}{dt} & = & \gamma I_1^a - \omega R_1^a & -\eta^a R_1^a + \eta^{a-1} R_1^{a-1} \\
\frac{dZ^a}{dt} & = & \lambda_0^a S_0^a + \lambda_1^a S_1^a &
\end{array}
\tag{1.1}
$$

where $\mathbb{1}_1(a)$ is the indicator function (non-zero at $a = 1$), η^a is the rate of ageing from age group a to age group $a+1$ (with $\eta^0 = 0$), and $\lambda^a(t)$ is the force of infection for age group a given by

$$
\lambda_i^a = \sum_{b=1}^{2} c^{a,b} \frac{I_0^b + I_1^b}{N^b}
\tag{1.2}
$$

where N^b is the population of age group b, δ_1, is the relative susceptibility of the population to RSV infection after being infected, and $c^{a,b}$ is the number of daily contacts made between age group a and b, estimated from the POLYMOD study through the socialmxr package.[1, 2]

The cumulative number of infections in age group a is given by Z^a, so the number of new infections z_w^a per week w in age group a is calculated by the equation:

$$
\left. \frac{dZ^a(t)}{dt} \right|_{7(w-1)}^{7w} = z_w^a
\tag{1.3}
$$

1.2.2 Likelihood

The data, x, used the calibrate the model is the number of weekly (w) confirmed positive RSV samples per age group (a) from the Respiratory DataMart System (RDMS) at PHE between July 2010 and June 2017.[3] This data is patient sensitive so cannot be shared publicly. In the likelihood, I assume that the number of confirmed positive RSV samples per age group and week (x_w^a) is the realisation of the random variable X_w^a,

$$
X_w^a \sim \text{Bin}\left(z_w^a, \epsilon^a\right)
\tag{1.4}
$$

where ϵ^a is the probability that someone with RSV is correctly identified in the RDMS dataset. By representing the set of all parameters to be fitted in the model as

$$
\Psi = \{\gamma, \omega, \delta_1, \alpha, a, b, \phi, \psi, I_1, I_2, I_3, \epsilon^1, \epsilon^2\}
$$

where I_1, I_2, I_3 are seeding parameters at $t = 0$ as defined in **Table 1.1** the likelihood calcualted through:

$$
\mathcal{L}(x|\Psi) = \prod_{t=1}^{52*7} \prod_{a=1}^{2} \mathcal{L}(x_w^a|\Psi) = \prod_{t=1}^{52*7} \prod_{a=1}^{2} \binom{z_w^a}{x_w^a} (\epsilon^a)^{x_w^a} (1 - \epsilon^a)^{z_w^a - x_w^a}
\tag{1.5}
$$

1.2.3 Prior distributions

A summary of all the model parameters with their prior distributions are contained in **Table 1.1**. In implementation the parameters are logit-transformed.

	Parameter	Prior distribution	Source
$1/\gamma$	Duration of infection (days)	$\mathcal{U}(2, 20)$	[4]
$1/\omega$	Duration of immunity (days)	$\mathcal{U}(60, 365)$	[5, 6]
δ_1	Relative proportion of population susceptibility RSV after primary infection	$\mathcal{U}(0, 1)$	—
α	Relative infectiousness of RSV after primary infection	$\mathcal{U}(0, 1)$	—
Parameters associated with transmission			
a	Rate of infection	$\mathcal{U}(0, 1)$	—
b	Amplitude	$\mathcal{U}(0, 1)$	—
ϕ	Peak transmission (offset) (days)	$\mathcal{U}(100, 250)$	[3]
ψ	Width of seasonal peak	$\mathcal{U}(0, 36.5)$	—
Initial parameters (at $t = 0$)			
I_1	Initial proportion infected RSV	$\mathcal{U}(0, 1)$	—
I_2	Initial proportion recovered from RSV	$\mathcal{U}(0, 1)$	—
$1/I_3$	Average duration between infections (years)	$\mathcal{U}(1, 4)$	—
Detection rates			
ϵ^1	0–2 years	$\mathcal{U}(0.0, 0.01)$	—
ϵ^2	2+ years	$\mathcal{U}(0.0, 0.01)$	—

Table 1.1: Prior distributions of the parameters in the mathematical model.

1.2.4 MCMC samplers

Overview of sampling methods

To sample values from the posterior distribution of the fitted parameters, we use Bayes rule

$$p(\Psi|D) \propto \mathcal{L}(x|\Psi)p(\Psi) \tag{1.6}$$

which simplifies the problem to sampling from the previously-defined likelihood and prior distributions. To generate the random samples we consider four different MCMC samplers; a random-walk Metropolis-Hastings (MH), an adaptive Metropolis-Hastings (AMH), a Differential Evolution with snooker update (DEzs) and a parallel tempering (PT). For a full explanation of these algorithms please refer to citations.

MH sampler The standard random-walk Metropolis-Hastings MCMC algorithm.[7, 8]

AMH sampler In the adaptive AMH algorithm, previous samples are used to alter the covariance matrix in the proposal distribution of the Metropolis-Hastings algorithm.[9]

DEzs sampler The DE algorithm is a population MCMC sampler, where three chains are run simultaneously and the proposal distribution of one chain is generated using the position of the other two chains and a snooker update.[10, 11]

PT sampler The PT algorithm runs numerous chains simultaneously which have an likelihood which is altered so that some (hotter) chains accept many proposals and other (colder) chains run similar to a Metropolis-Hastings algorithm.[12]

Implementation

The MH, AMH, and DEzs samplers are implemented via the R package `BayesianTools`[13] where explicit details about their implementation can be found. These three samplers run for 1,000,000

steps with thinning occurring every 10th step. The PT sampler is implemented via the R package `ptmc`[14] where explicit details on implementation can be found) and is run for 100,000 steps with a temperature ladder of 10 chains per run with thinning every 10th step (only the coldest chain is used to calculate the posterior distribution). For all four samplers the burn-in is half of the total run time and each are ran ten times so that convergence via Gelman-Rubin statistics and manual inspection of trace plots can be performed.[15] An R Markdown file with all code used for this analysis can be found at the Github repository link `https://github.com/dchodge/ptmc_performace`.

1.3   Results

The trace plots for the log(-log($p(\Psi|D)$))) for 10 runs of each of the four MCMC samplers are given below.

1.3.1   MH sampler

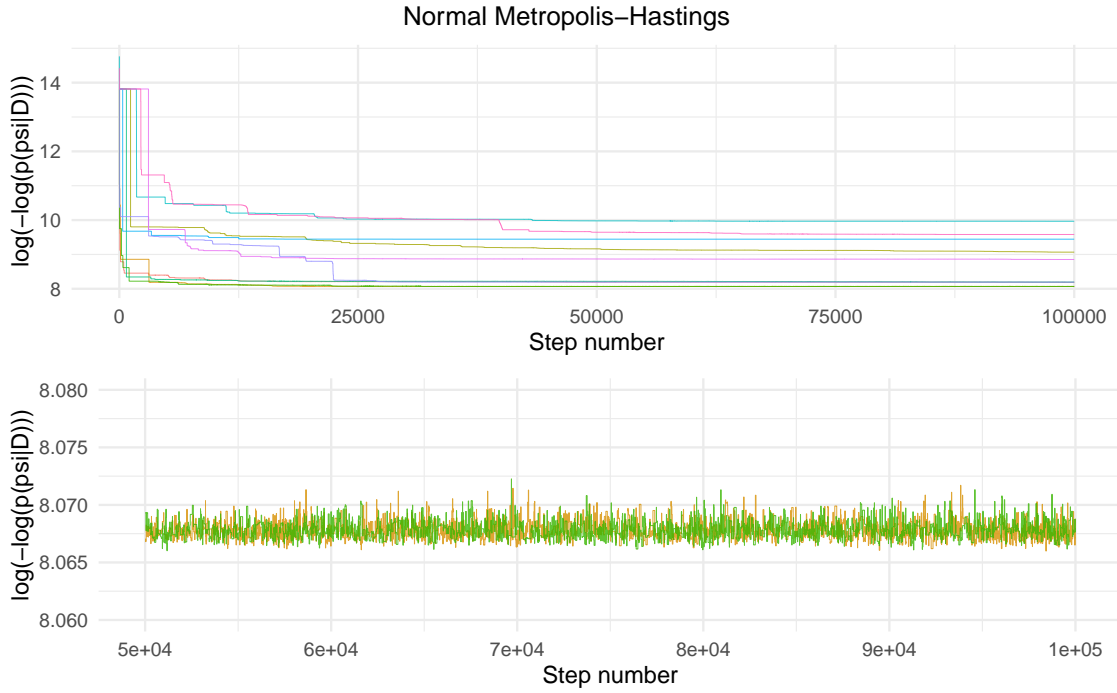For the MH sampler, two of the ten chains appear to reach the convergent state after the burn-in (**Figure 1.1**).
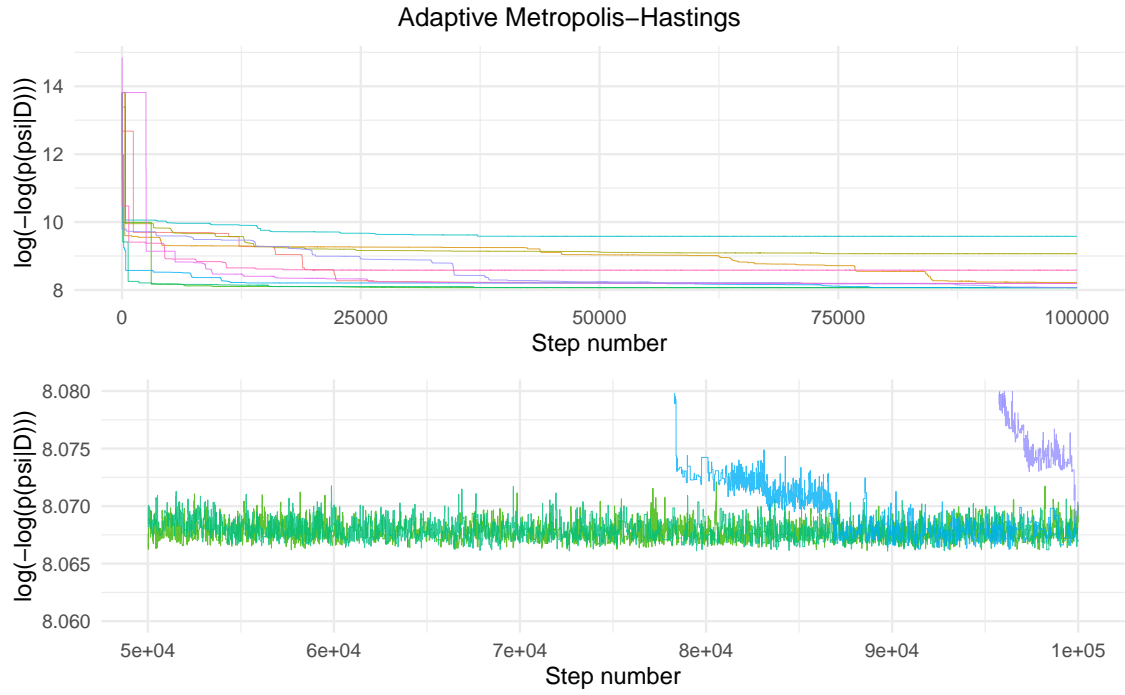

Figure 1.1: Top: Trace plots for the ten chains for the whole run. Bottom: Trace plots for the ten chains for the run after burn-in.

1.3.2   AMH sampler

For the AMH sampler, two of the ten chains appear to reach the convergent state after the burn-in, with one additional chain reaching the convergent state after the burn-in (**Figure 1.2**).
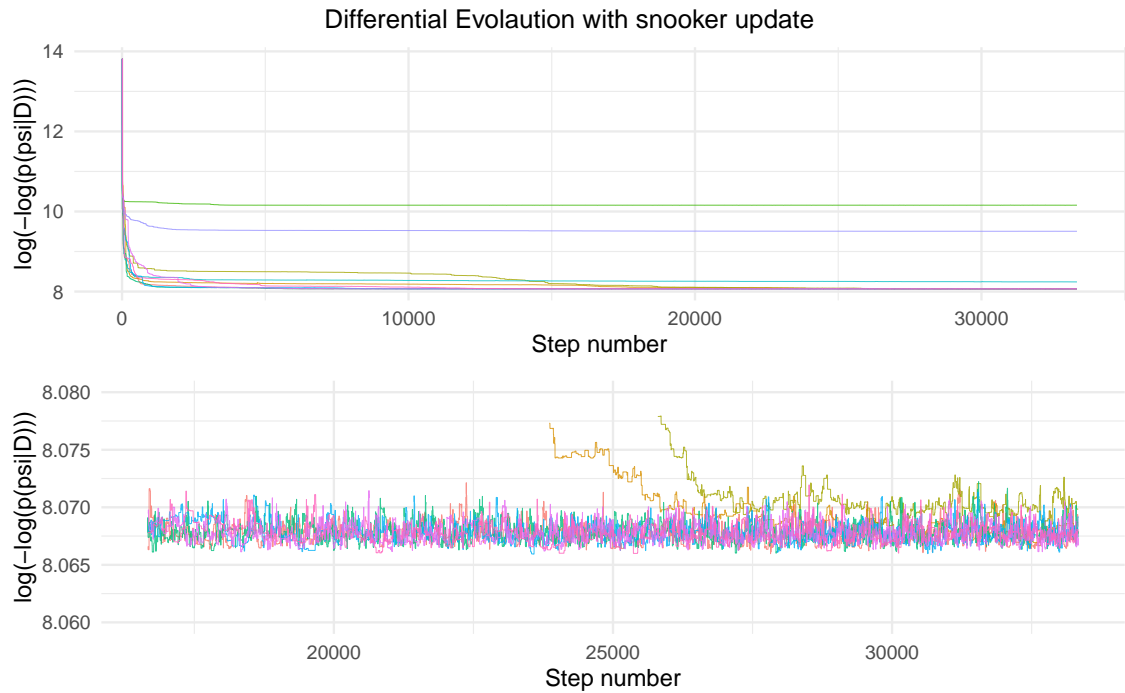
Figure 1.2: Top: Trace plots for the ten chains for the whole run. Bottom: Trace plots for the ten chains for the run after burn-in.

1.3.3 DEzs sampler

For the DEzs sampler, five of the ten chains appear to reach the convergent state after the burn-in, with two additional chain reaching the convergent state after the burn-in **Figure 1.3**).

Figure 1.3: Top: Trace plots for the ten chains for the whole run. Bottom: Trace plots for the ten chains for the run after burn-in.

1.3.4 PT sampler

For the PT sampler, eight of the ten chains appear to reach the convergent state after the burn-in, with the final two chains reaching the convergent state after the burn-in shortly after **Figure 1.4**).
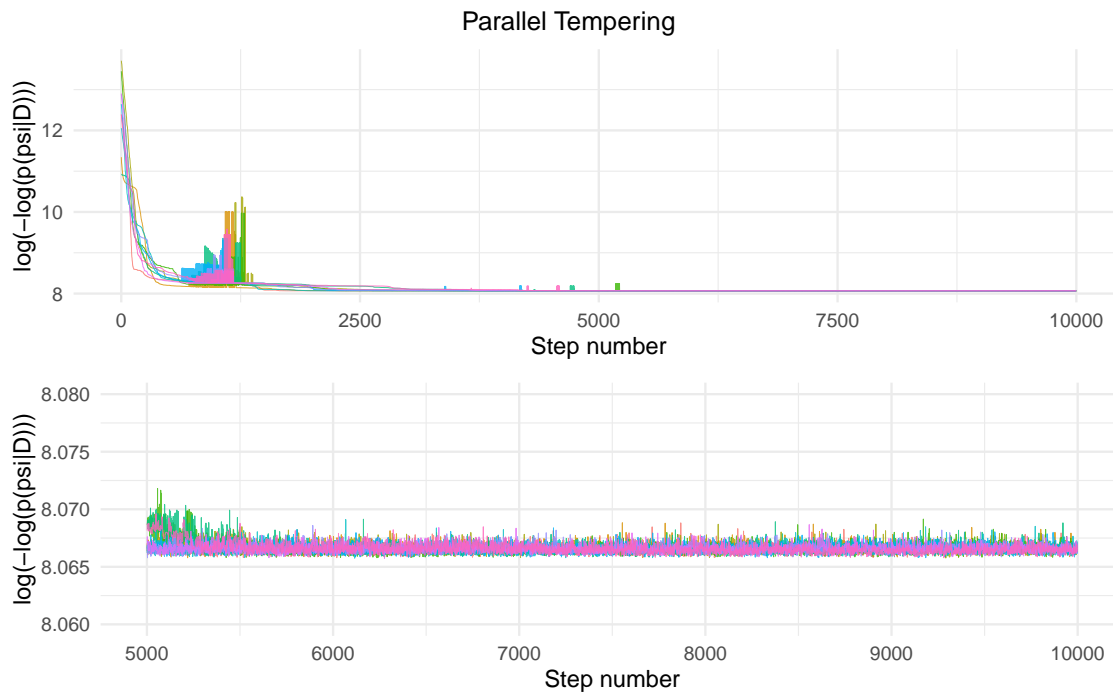

Figure 1.4: Top: Trace plots for the ten chains for the whole run. Bottom: Trace plots for the ten chains for the run after burn-in.

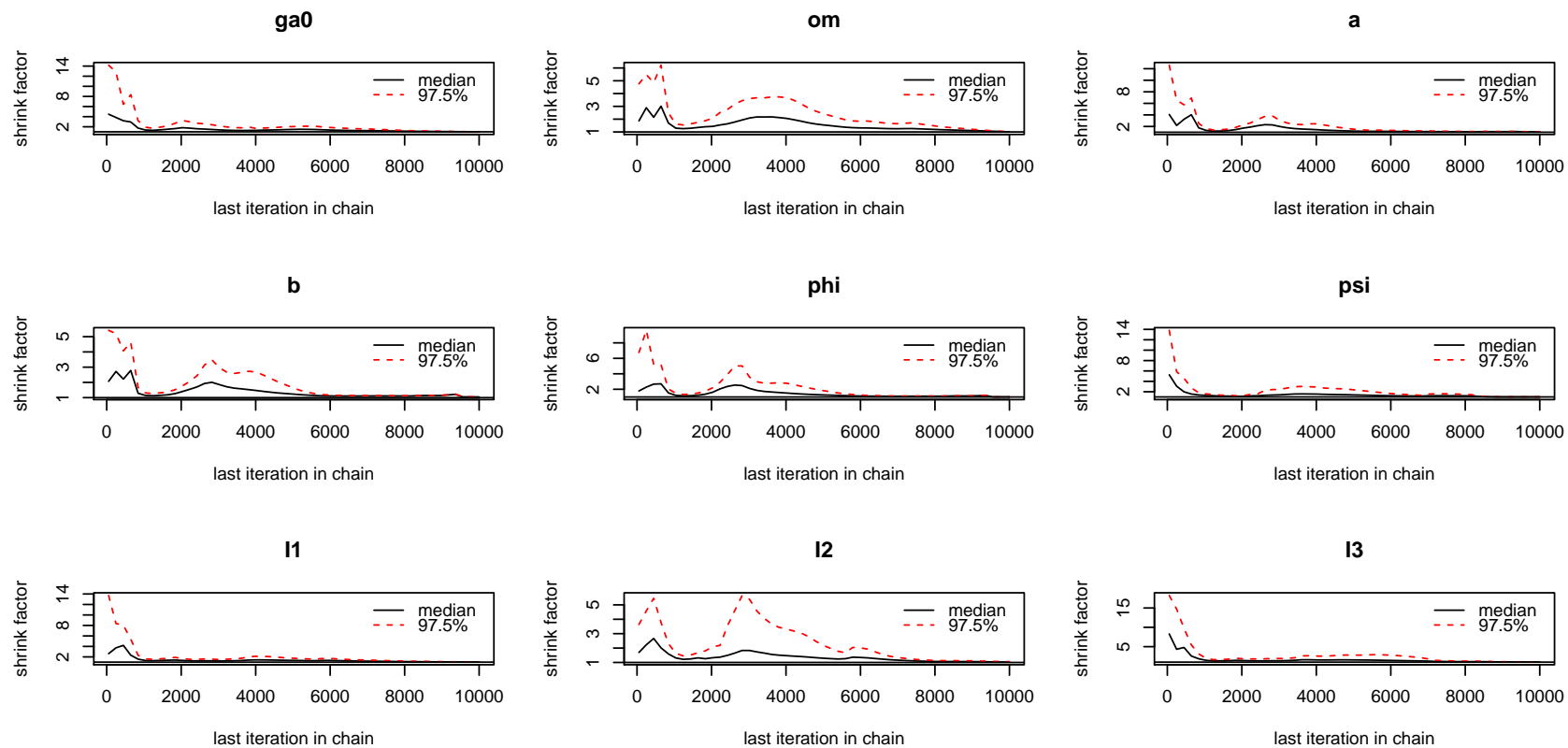The Gelman-Rubin statistics over time for the PT sampler is given in **Figure 1.5-1.6**.

Figure 1.5: The shrink-factor (Gelman-Rubin statistic) over time for the PT sampler. Low values suggest convergence.
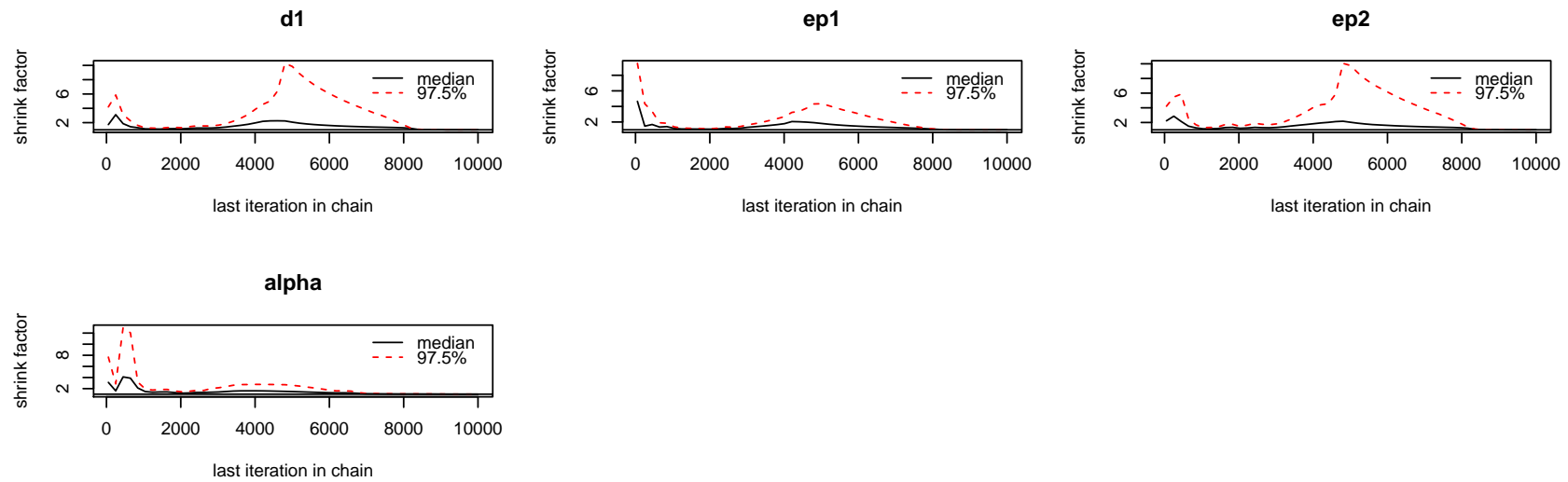
Figure 1.6: The shrink-factor (Gelman-Rubin statistic) over time for the PT sampler. Low values suggest convergence.

1.4 Discussion

The MCMC samplers are run for a fixed duration to see which provides the most reassuring convergence in a given space of time. The parallel tempering has the most convincing convergence compared to the other 3 samplers, with all 10 of the chains converging before the end of the run. The multivariate potential scale reduction factor for the PT sampler is 1.03 and the scale reduction factor is less than 1.08 for all the fitted parameters suggesting convergence.[15]

It is encouraged that future epidemic models will use the parallel tempering MCMC sampler (through the `ptmc` R package) to sample from posterior distributions. Using such pre-built algorithms will assuage mathematical modellers concerns surrounding convergences of epidemic models, allowing more time for analysis more relevant to policy makers (i.e. impact of intervention programmes or cost-effectiveness analysis) to be performed.

Bibliography

[1] CRAN - Package socialmixr;. Available from: https://cran.r-project.org/web/packages/socialmixr/.

[2] Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 2008;5(3):e74.

[3] Zhao H, Green H, Lackenby A, Donati M, Ellis J, Thompson C, et al. A new laboratory-based surveillance system (Respiratory Datamart System) for influenza and other respiratory viruses in England: Results and experience from 2009 to 2012. Eurosurveillance. 2014;19(3):1–10.

[4] Okiro EA, White LJ, Ngama M, Cane PA, Medley GF, Nokes DJ. Duration of shedding of respiratory syncytial virus in a community study of Kenyan children. BMC Infectious Diseases. 2010;10:15.

[5] Scott PD, Ochola R, Ngama M, Okiro EA, James Nokes D, Medley GF, et al. Molecular Analysis of Respiratory Syncytial Virus Reinfections in Infants from Coastal Kenya. The Journal of Infectious Diseases. 2006;193(1):59–67. Available from: https://academic.oup.com/jid/article-lookup/doi/10.1086/498246.

[6] Hall CB, Walsh EE, Long CE, Schnabel KC. Immunity to and frequency of reinfection with respiratory syncytial virus. J Infect Dis. 1991;163(4):693–698.

[7] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications; 1970. 1. Available from: https://academic.oup.com/biomet/article-abstract/57/1/97/284580.

[8] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of State Calculations by Fast Computing Machines. Citation: J Chem Phys. 1953;21(6):1087.

[9] Haario H, Saksman E, Tamminen J. An adaptive Metropolis algorithm. Bernoulli. 2001;7(2):223–242.

[10] Ter Braak CJF. A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: Easy Bayesian computing for real parameter spaces. Statistics and Computing. 2006 sep;16(3):239–249.

[11] Ter Braak CJF, Vrugt JA. Differential Evolution Markov Chain with snooker updater and fewer chains. Statistics and Computing. 2008 dec;18(4):435–446.

[12] Miasojedow B, Moulines E, Vihola M. Adaptive parallel tempering algorithm. Journal of Computational and Graphical Statistics. 2012 may;22(3):649–664. Available from: http://arxiv.org/abs/1205.1076.

[13] General-Purpose MCMC and SMC Samplers and Tools for Bayesian Statistics [R package BayesianTools version 0.1.7];.

[14] dchodge/ptmc: A general-purpose Parallel Tempering MCMC sampler. Github;. Available from: https://github.com/dchodge/ptmc.

[15] Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics. 1998;7(4):434–455.