## Fitting SIR Model to Outbreak Data

Student Number: 1493823

**Abstract**

**Background**: Fitting a model is a critical step in modelling to better understand the real situation based on the observed data. Two approaches include Bayesian inference and maximum likelihood. RStan allows its user to perform the inference in R.

**Method**: Three Markov-Chain-Monte-Carlo (MCMC) chains with 3000 interactions and a burn-in period of 1000 were performed to fit three SIR models to the data. Models 1 and 2 are basic SIR models that differ only in the prior of the gamma parameter. Some additional parameters were included in model 3 to consider behaviour change due to increased infection prevalence. Posterior and prior predictive checks were performed in addition to the resulting trace plots and marginal posterior densities to evaluate the reliability of the model. *Rhat* and *neff* were also used to assess the chain convergence. The Bayes factor was assessed to evaluate which model fits the data better. All analyses were performed in R using the RStan package.

**Results**: The resulting posterior distributions for model 1, (mean = 0.14 (95% CrI 0.13 to 0.15)), (mean = 0.09 (95% CrI 0.08 to 0.1)), *R0* (mean = 1.55 (95% CrI 1.48 to 1.64)) and recovery time (mean = 10.87 (95% CrI 9.59 to 12.40)) are normally distributed. Similarly in model 2, the posteriors ( = 0.14 (95% CrI 0.13 to 0.15), = 0.09 (95% CrI 0.08 to 0.10), *R0* = 1.54 (95% CrI 1.48 to 1.62), and recovery time = 10.72 (95% CrI 9.56 to 12.12)) follow normal distributions. Meanwhile, in model 3, the posteriors, (mean = 0.09 (95% CrI 0.06.6 to 0.128)), γ (mean = 0.15 (95% CrI 0.12 to 0.18), *R0*, recovery time (1/γ) are normally distributed. Meanwhile, (mean = 2.85x10-1 (95%CrI 0.0337 to 0.928)), a (mean = 0.002 (95% CrI 2.69 x 10-5 to 0.0128)), and *k* (mean = 0.250 (95% CrI 0.0348 to 0.896)) follow a negative binomial distribution. Prior predictive checks of models 1 and 2 revealed that most trajectories showed that the models are not excessively constrained by the priors. The Bayes factors of models 1 and 3 were compared to determine that model 3 had a better fit than model 1.

**Conclusion**: All of the models provide a good fit for the data but the information model is better than the original model.

**Introduction**

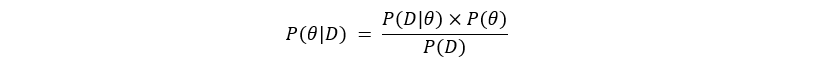
Mathematical modelling in epidemiology provides a means of informing health policy for disease control and prevention strategies. It allows us to evaluate hypothetical scenarios and maximise outcomes under resource constraints [(1, 2)](https://www.zotero.org/google-docs/?bJNsgu). Selecting the most appropriate model for a given scenario can be a challenging task [(2)](https://www.zotero.org/google-docs/?0Yd7HM). Model fitting provides a systematic approach for identifying the model and its parameters that generate the best possible predictions given observed data [(3)](https://www.zotero.org/google-docs/?XJtQli). For example, in infectious disease models, reducing the basic reproduction number (*R0*) can be a key target of an intervention to reduce the potential for an outbreak to occur [(4–6)](https://www.zotero.org/google-docs/?h6cGm3).

While a model generates likely outcomes from a set of parameters, inference is the process of determining the likely parameters based on observations [(7)](https://www.zotero.org/google-docs/?jpsKM6). Two common approaches to model fitting are maximum likelihood estimation (MLE) and Bayesian inference. Both methods use probability principles to estimate the parameters that are most likely to have generated the observed data. However, MLE produces a point estimate for the parameters with the highest likelihood of a given distribution using the observed data [(8)](https://www.zotero.org/google-docs/?uRqCF2). For example, if a population is known to follow a normal distribution, but the mean and variance are unknown, MLE can be used to estimate them, using a limited sample of the population, by finding particular values of the mean and variance so that the observation is the most likely result to have occurred [(9)](https://www.zotero.org/google-docs/?oP65Er).

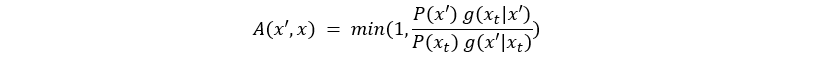
Meanwhile, Bayesian inference considers prior beliefs about the parameters in addition to the likelihood, providing a more comprehensive view of the parameters' uncertainty [(10)](https://www.zotero.org/google-docs/?4xtf6D). In Bayesian inference, Bayes’ theorem is used to obtain a posterior distribution. It is given by the following equation:



Or, in terms of variables:



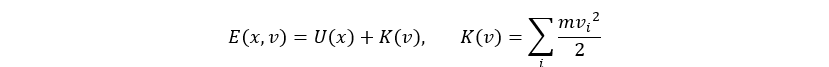
Bayesian inference updates prior beliefs with data to find the most likely parameters. The Metropolis-Hastings algorithm generates samples from probability distributions that converge to the target posterior [(11)](https://www.zotero.org/google-docs/?Cf8yUl). It involves accepting or rejecting proposed values based on the likelihood and prior of the new and previous values through acceptance probability, which is the ratio of the probability density of the proposed move from *xt* to *x’* to that of the reverse move [(11, 12)](https://www.zotero.org/google-docs/?gOZll5). It is given by the following equation:



In the Hamiltonian Monte Carlo (HMC) approach, a parameter space is explored using the gradient of the posterior and its acceleration resulting from the local geometry in the posterior density [(13)](https://www.zotero.org/google-docs/?TWvy5b). However, as it uses gradients for exploring the parameter space, this method does not work for discrete data [(14)](https://www.zotero.org/google-docs/?ecSQws). The HMC is given by the following equation:

,

where the potential and kinetic energy of the system are given by the following equations:



This essay will evaluate how a model fitting can be performed using the HMC approach in RStan. The SIR model will be parameterised for the observed data.

**Methods**

***Model Formulation***

It is assumed that the data follows a closed system SIR model—Susceptible (S), Infected (I), and Recovered (R)—where the total population, *N*, is the sum of the number of individuals in each compartment

A picture containing text, device, screenshot

Description automatically generated

Each compartment's size at time *t* represents a state variable in a mathematical model. The flow chart in Figure 1 illustrates the state variables and processes included in the model, while Tables 1 and 2 provide descriptions of each parameter. The model is expressed as a set of nonlinear ordinary differential equations, where each balance equation controls the rate of change of a state variable. The data are fitted with three models: model 1 and model 2 which are the original SIR models, while model 3 includes a behaviour change as represented by equations *X* and *Y*.

|  |  |  |
| --- | --- | --- |
| **Parameters** | **Description** | **Baseline value** |
| *β* | Transmission rate | To be fitted |
| *𝛾* | Recovery rate | To be fitted |
| *R0* | Basic reproduction number | To be fitted |
| *t0* | Initial simulation time | 0 |
| *tf* | Final simulation time | 200 days |
| *S(t0)* | Initial susceptible population | 19,999 |
| *I(t0)* | Initial infected population | 1 |
| *R(t0)* | Initial recovered population | 0 |
| *N* | Total population | 20,000 |

***Model Fitting***

The prior transmission rate in both models follows a weakly informative prior of . Its value is restricted to be positive. The priors for the recovery rate, 𝛾, in both models are different. In model 1, a weakly informative prior of . In model 2, it is believed that the infectious period lasts between 7 and 14 days. Thus, a normally distributed prior for with mean and standard deviation . A normal distribution instead of a uniform distribution was chosen to allow the inference to explore any relevant values in parameter space.

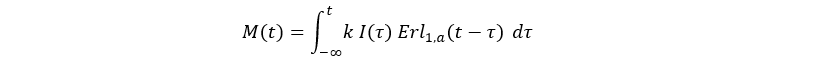
For the dispersion parameter, a prior of is used. This approach is recommended to avoid assigning an excessively high prior probability to models with significant overdispersion. A prior predictive check is conducted before model parameterisations start.

Bayesian inference was performed using three Markov-Chain-Monte-Carlo (MCMC) chains with 3000 iterations and a burn-in period of 1000. The convergence of the chains was assessed through *Rhat* and effective sample size (*neff*) values, in addition to visual inspection of the trace plots. The posterior distribution for each parameter and the associated credible intervals were reported. All analysis was done in R version 4.2.2.

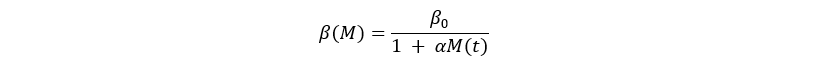
**Incorporating the Impact of Behaviour Change in Model 3**

|  |  |
| --- | --- |
| **Parameters** | **Description** |
| *a* | the inverse of the average information delay *Ta* |
| *k* | Information coverage |
|  | reactivity factor of a voluntary change in contact patterns |

Another assumption of this model is that the force of infection is partially determined on a fully voluntary basis and depends on the available information concerning the spread of the disease in the community. The information is mathematically represented by an information index, *M(t)*, which summarises the information about the current and past values of the disease. It is given by the following equation:

Following [(15)](https://www.zotero.org/google-docs/?GMJiOD), the differential equation for M(t) can be derived as follows:

The transmission rate can be expressed as a function of M, as noted in [(16)](https://www.zotero.org/google-docs/?KkUDYG):



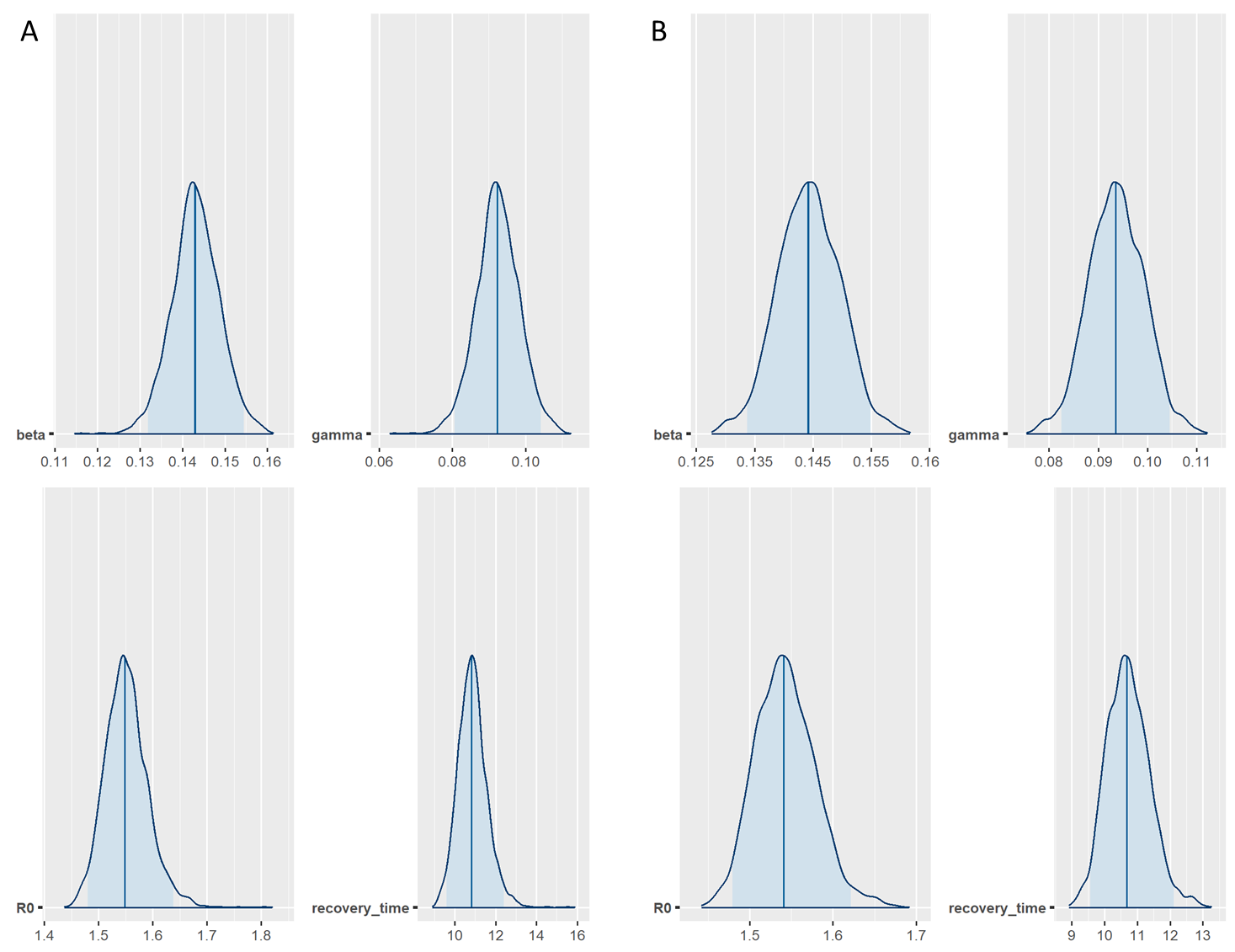
This modification allows the transmission rate to vary with the infection prevalence. Its value decreases when the population in the infected compartment increases.

**Results**

***Model 1 and Model 2***

*Posterior Distributions from the SIR Model*

The resulting posterior distribution of these parameters in both models is similar. As illustrated in Figure 1, these posteriors followed a normal distribution similar to their conjugate priors. The detail of their values and their diagnostic parameters are summarised in Table 1.



**Figure 1.** Posterior distribution from model 1 (panel A) and model 2 (panel B).

Two parameters, (mean = 0.14 (95% CrI 0.13 to 0.15)) and (mean = 0.09 (95% CrI 0.08 to 0.1)), were obtained after fitting model 1 to the data. While *R0* (mean = 1.55 (95% CrI 1.48 to 1.64)) and recovery time (mean = 10.87 (95% CrI 9.59 to 12.40)) were derived from these parameter values.

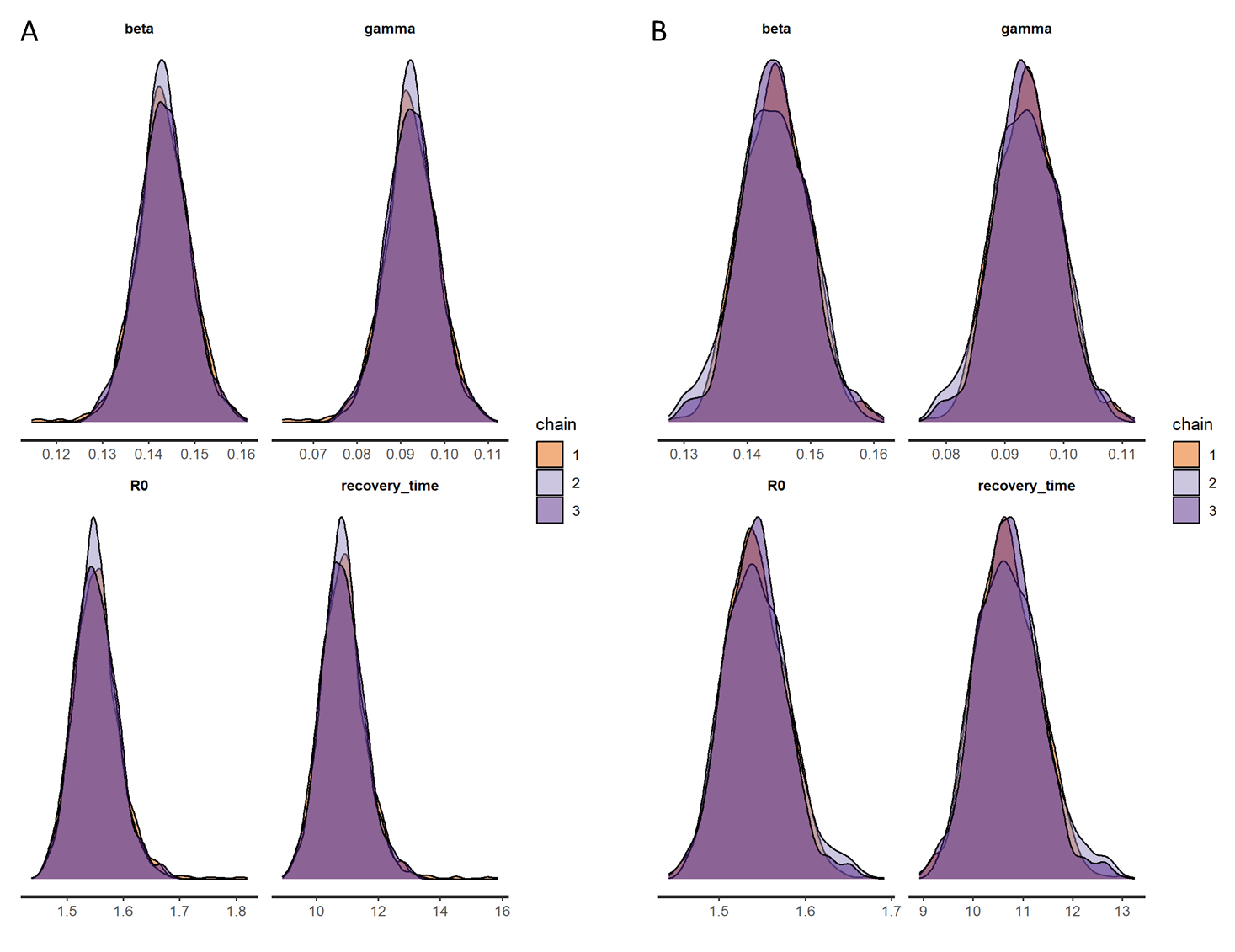
These results were similar to the value of the parameter obtained from fitting model 2 to the data (mean = 0.14 (95% CrI 0.13 to 0.15)); (mean = 0.09 (95% CrI 0.08 to 0.10)); *R0* (mean = 1.54 (95% CrI 1.48 to 1.62)); and recovery time (mean = 10.72 (95% CrI 9.56 to 12.12)).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Parameters | Model 1 | | | | Model 2 | | | |
| Prior | Mean (95% CrI) | *neff* | *Rhat* | Prior | Mean (95% CrI) | *neff* | *Rhat* |
| *β* |  | 0.14 (0.13 to 0.15) | 1020 | 1 |  | 0.14 (0.13 to 0.15) | 1238 | 1 |
| *γ* |  | 0.09 (0.08 to 0.10) | 1021 | 1 |  | 0.09 (0.08 to 0.10) | 1231 | 1 |
| *R0* | - | 1.55 (1.48 to 1.64) | 1009 | 1 | - | 1.54 (1.48 to 1.62) | 1183 | 1 |
| Recovery time | - | 10.87 (9.59 to 12.40) | 992 | 1 | - | 10.72 (9.56 to 12.12) | 1184 | 1 |

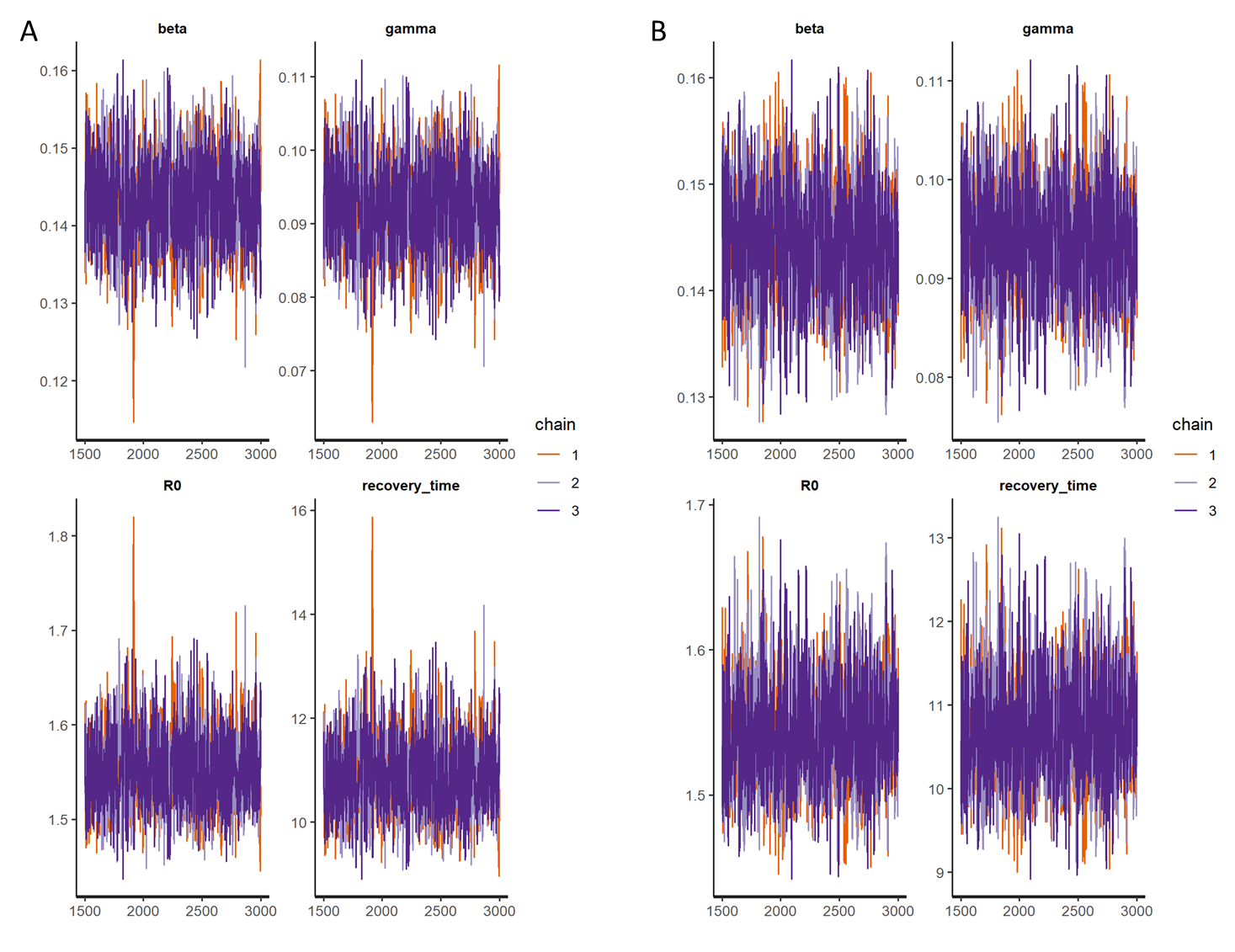
**Table 1.** This table summarises the value for each model parameter and its diagnostic check of indicators, the effective sample size, *neff*, and *Rhat*.

Both models have the value of *Rhat* = 1, which showed close agreement with one another among 3 Markov chains. Additionally, *neff* is large (> 1007), which means that the Markov chains were able to cohesively explore the parameter space. For a comparison, model 2 has higher *neff* throughout all parameters, indicating that the Markov chains in model 2 were better mixed than those in model 1.

*Model’s Reliability Diagnostics*

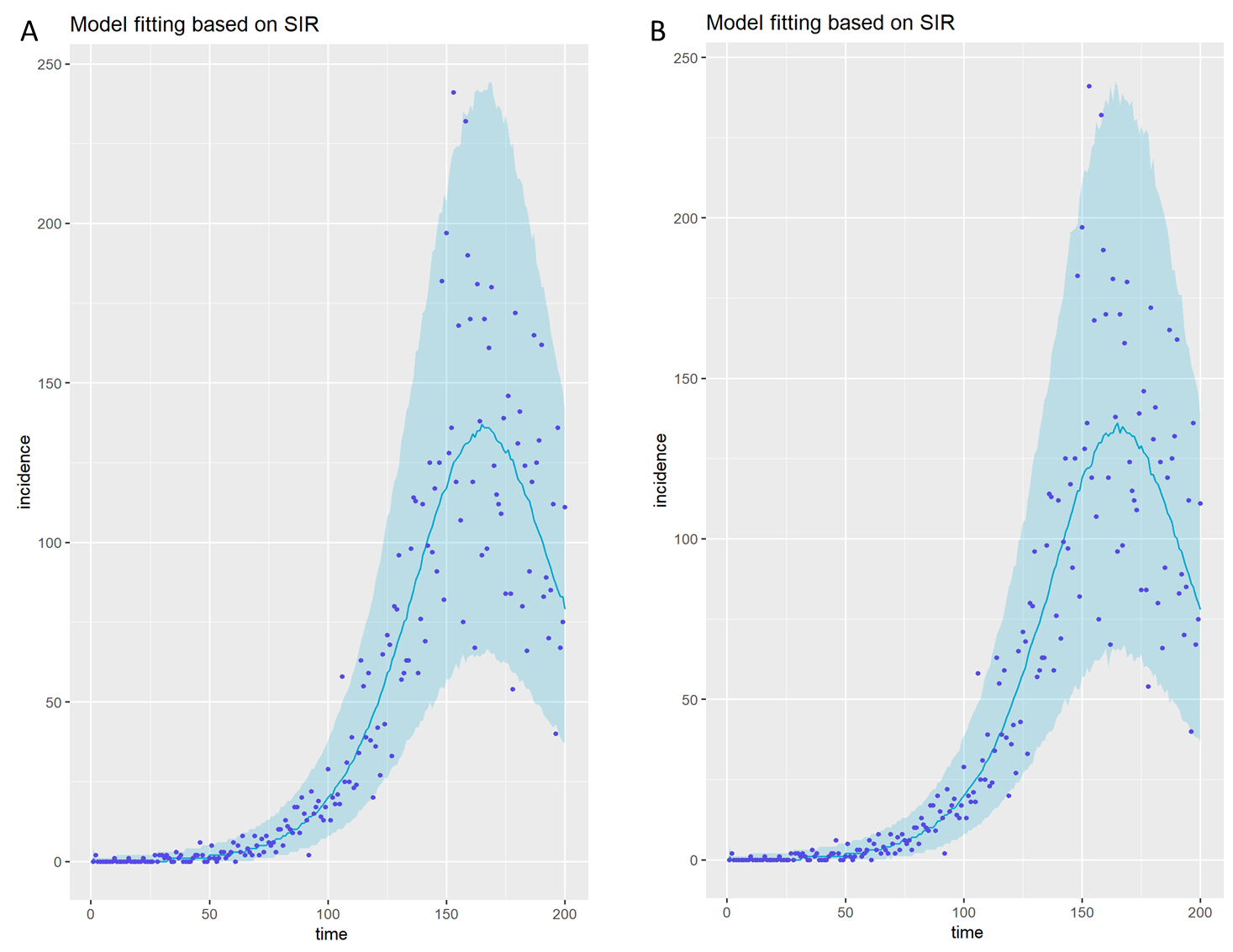


**Figure 2.** The marginal posterior densities of each chain show good agreement between chains in both model 1 (panel A) and model 2 (panel B).



**Figure 3.** Trace plots of each chain confirm that each chain reaches convergence in both model 1 (panel A) and model 2 (panel B).

In a further assessment of the convergence diagnostics, the marginal posterior densities (Figure 2) and trace plots (Figure 3) for both models showed that all chains agree with one another. This indicates that the parameterisation of the Bayesian inference for both models is reliable.

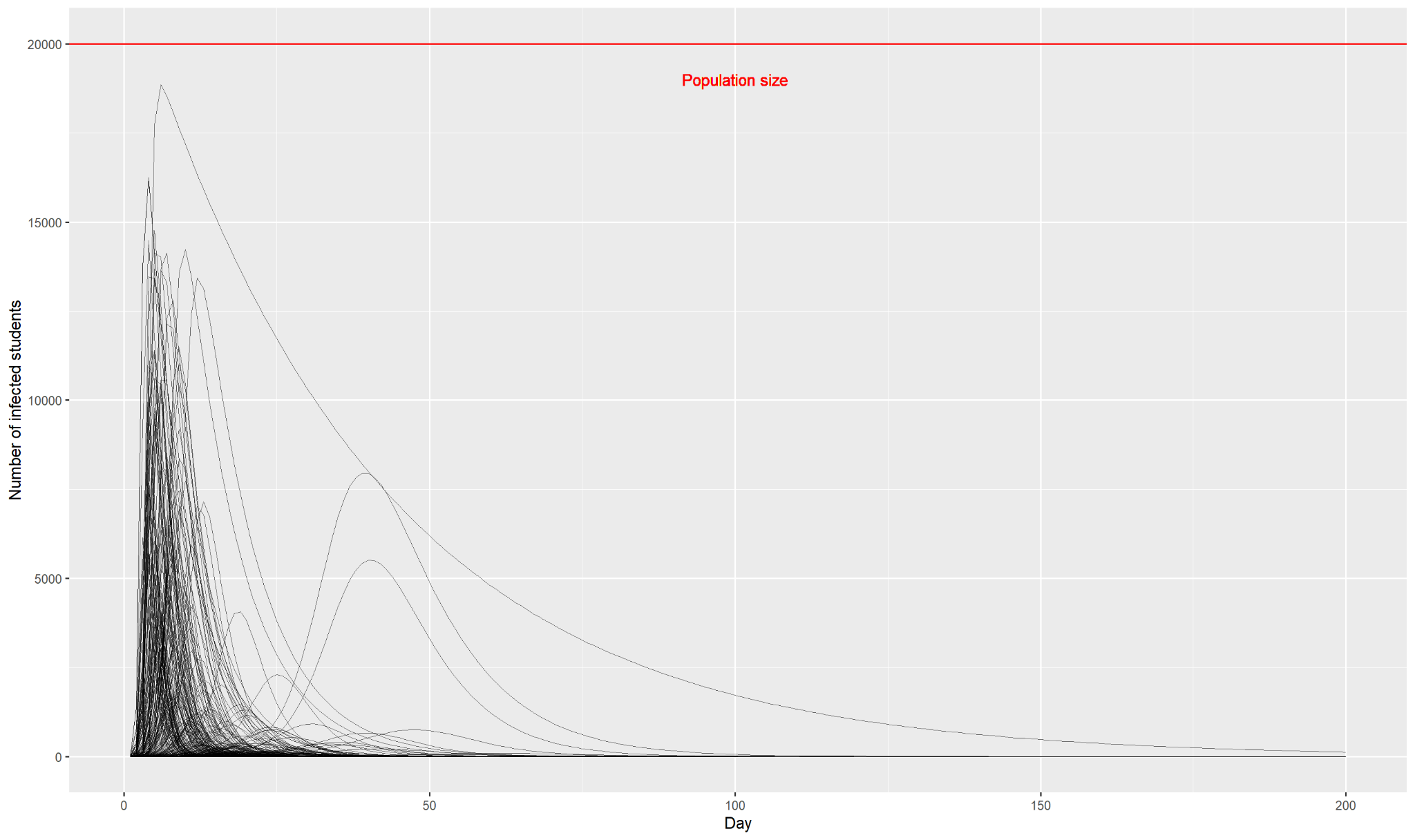


**Figure 4.** Predictive checks from model 1 (panel A) and model 2 (panel B). These plots show that both models capture the data since most of the purple dots lie within the 95% CrI area for both models.

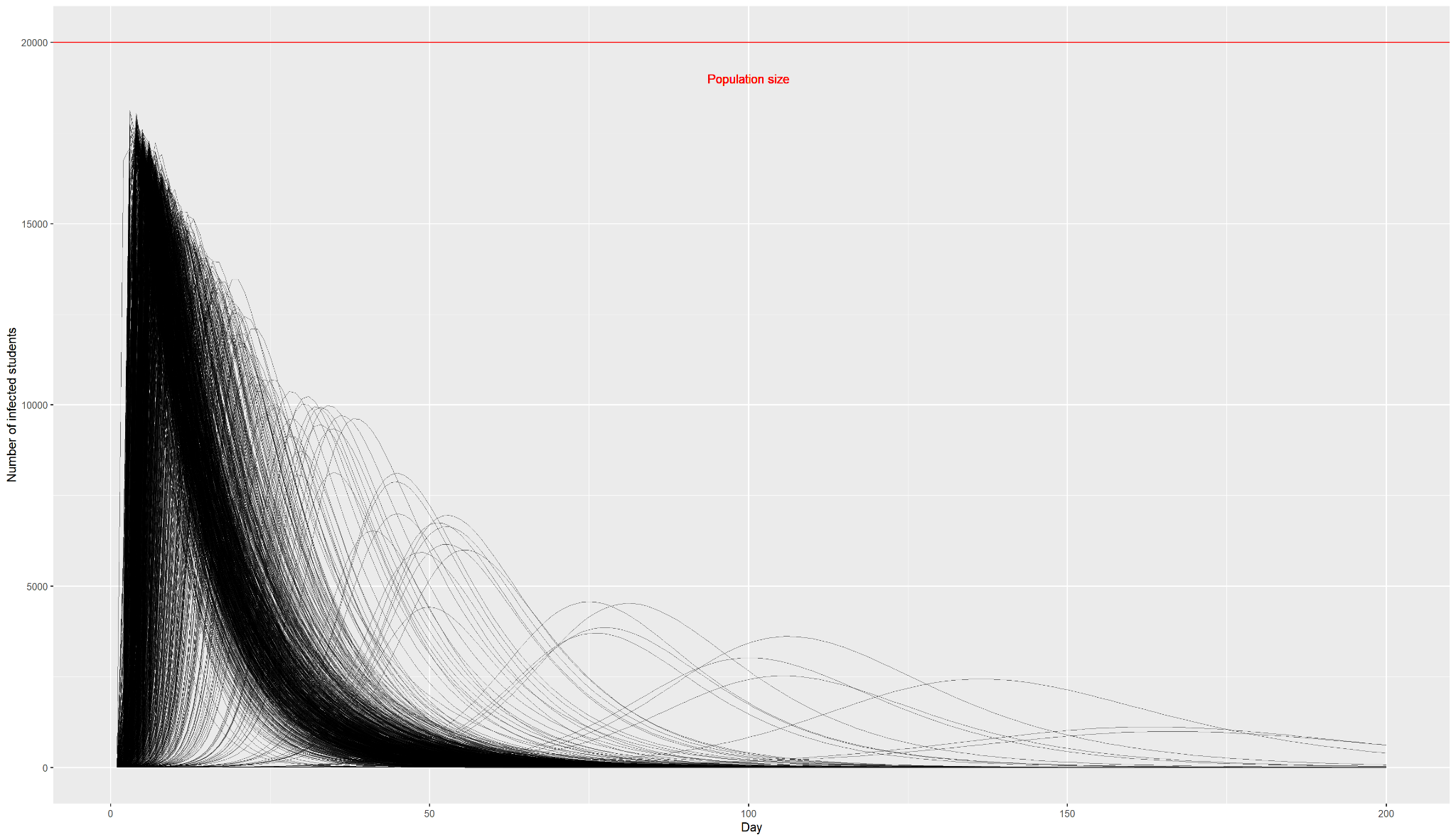
The predictive checks for both models were illustrated in Figure 4. The vast majority of the observed data lies within the 95% CrI of the predicted model, showing that the simulations are consistent with the observed data. These results verify that both models capture the structure of the data and provide a good fit to the data.

*Prior Predictive Checks*

Prior predictive checks were performed to evaluate prior choices in both models. The results are illustrated inFigure 5 for model 1 and in Figure 6 for model 2.



**Figure 5.** This graph illustrates a set of 1000 epidemic trajectories. Each line is a unique simulated trajectory sampled from prior distributions of a weakly informative prior of and a moderately informative prior of . The red line represents the total population size of 20,000.



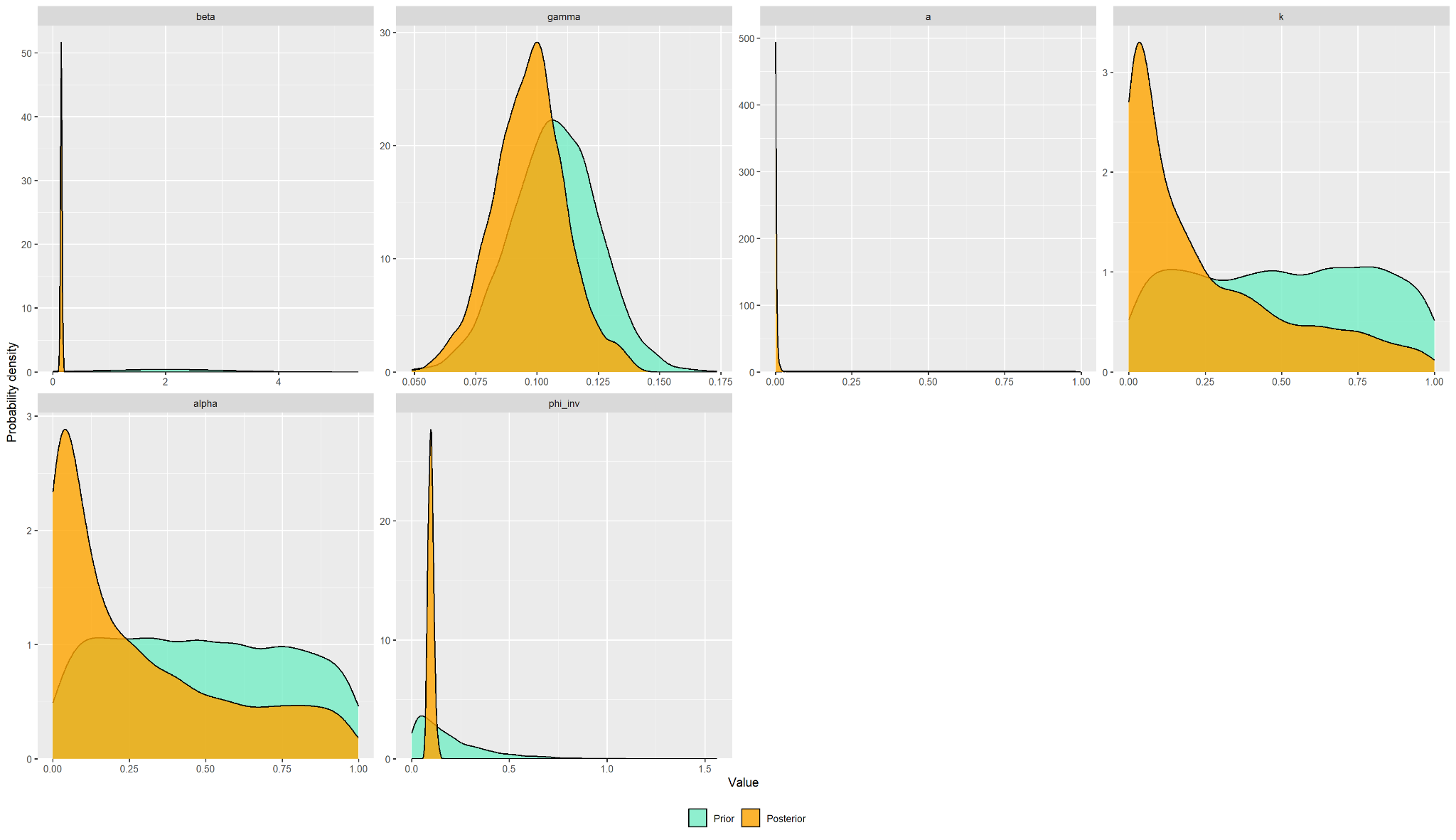
**Figure 6.** This graph illustrates a set of 1000 epidemic trajectories. Each line is a unique simulated trajectory sampled from prior distributions of a weakly informative prior of ) and a moderately informative prior of . The red line represents the total population size of 20,000.

In model 1, the distribution of and are weakly informative and truncated at 0 as the transmission rate and recovery rate cannot be negative. In model 2, these parameters are also believed to be non-negative, but for the recovery time, (), a moderately informative prior was chosen such that 95% of its value is between 7 days and 14 days. A prior of was chosen to represent this prior belief.

The resulting trajectories show that the outbreak seems to spread more slowly using prior distribution in model 1 compared to model 2. In model 2, it appears that the magnitude of the outbreak is a lot higher and occurs more quickly than the observed data.

***Model 3***

*Posterior Distribution*



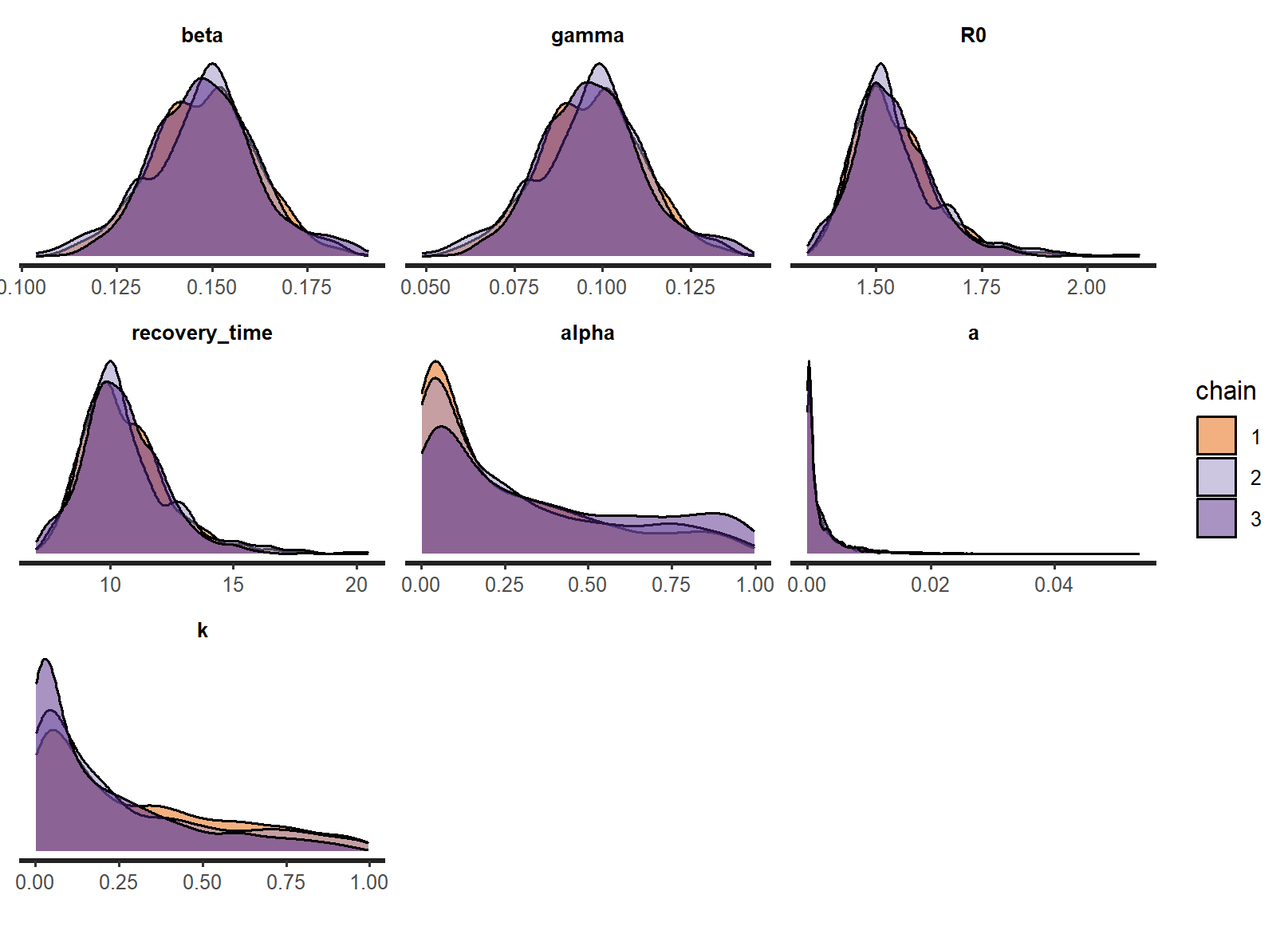
**Figure 7.** Posterior distributions for the information model and their priors.

Parameterisation of the information model on the SIR dynamics is summarised in Table 3. The distribution of (mean = 0.09 (95% CrI 0.06.6 to 0.128)), (mean = 0.15 (95% CrI 0.12 to 0.1.8), R0, and recovery time () follow a normal distribution. Meanwhile, (mean = 0.285 (95% CrI 0.0337 to 0.928)), *a* (mean = 0.00220 (95% CrI2.69 x 10-5 to 0.01.28)), and *k* (mean = 0.250 (95% CrI 0.0348 to 0.896)) follow a negative binomial distribution (Figure 7).

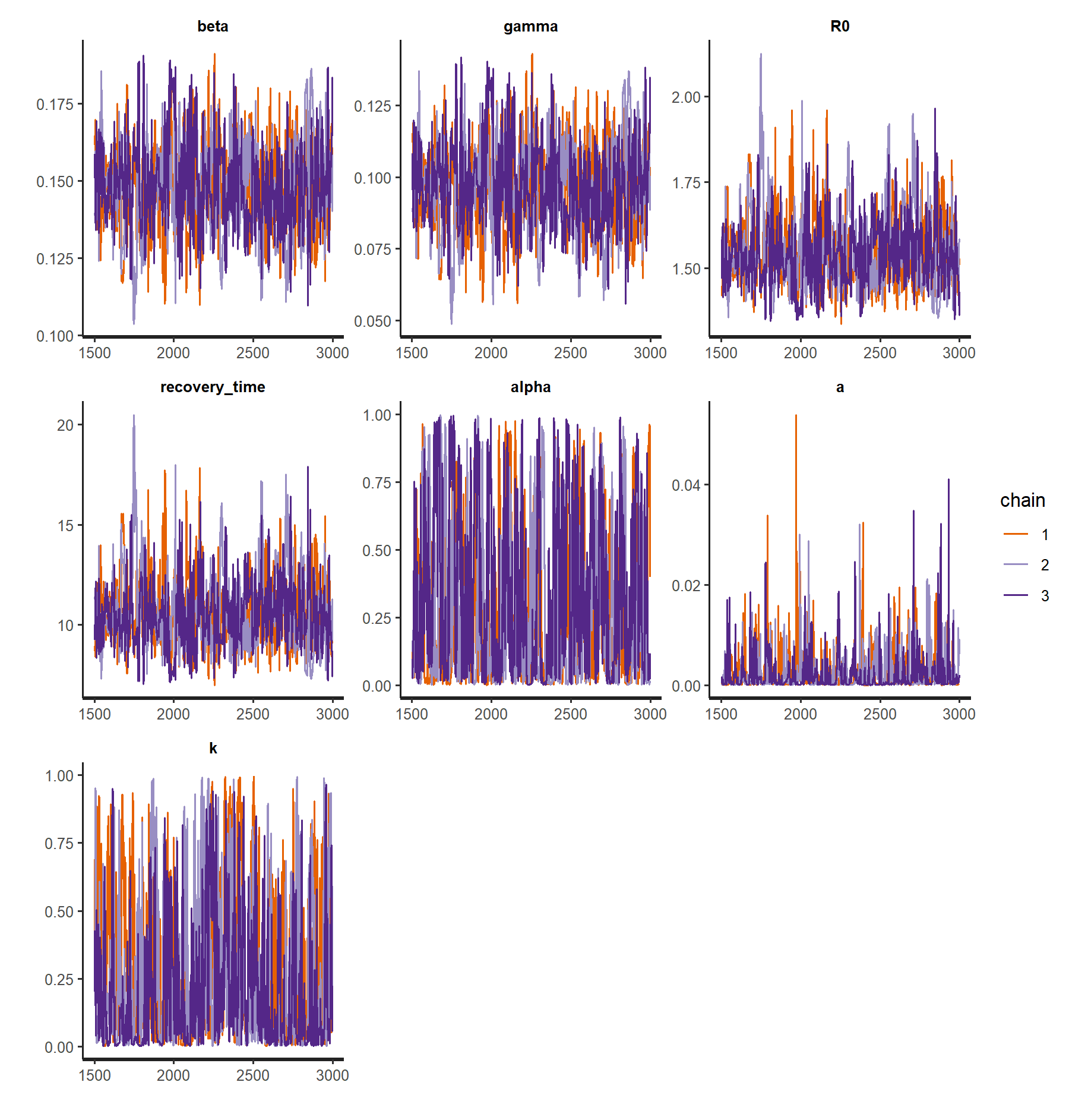
**Table 3.** Summary of the results of model fitting using the information model.

|  |  |  |  |
| --- | --- | --- | --- |
| Parameters | Information model | | |
| Value (95% CrI) | *neff* | *Rhat* |
|  | 9 x 10-2 (6.6 x 10-2 to 1.28 x 10-1) | 359 | 1.01 |
|  | 1.5 x 10-1 (1.2 x 10-1 to 1.8 x 10-1) | 360 | 1.01 |
|  | 2.85 x 10-1 (3.37 x 10-2 to 9.28 x 10-1) | 384 | 1.02 |
| *a* | 2.20 x 10-3 (2.69 x 10-5 to 1.28 x 10-2) | 639 | 1.00 |
| *k* | 2.50 x 10-1 (3.48 x 10-2 to 8.96 x 10-1) | 220 | 1.01 |

The diagnostic checks show that this model converges well, but with lower *neff* and higher *Rhat* than model 1 and 2.

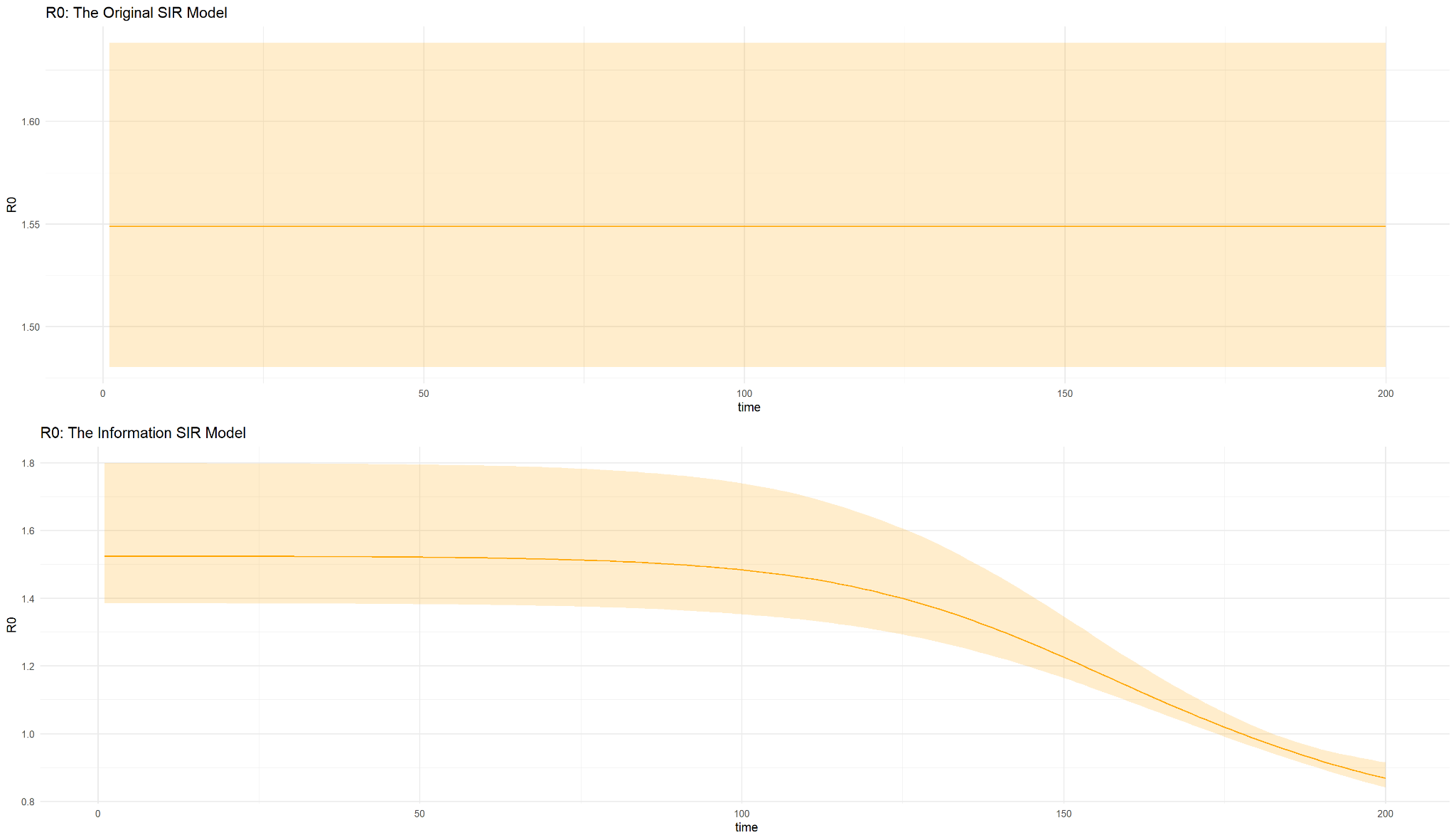


**Figure 8.** The marginal density plot for each parameter in the information model. This shows that all three MCMC chains agree with one another.



**Figure 9.** Trace plots for each parameter in the information model, showing that all chains have reached a convergence and agree with one another.

Model comparison: *R0* and the Bayes factor



**Figure 10.** This graph illustrates how R0 is changing over time for the information SIR model while for the original SIR model the R0 remains constant.

A visual inspection of marginal posterior densities (Figure 8) and the trace plots (Figure 9) shows that all three chains agree with one another. For the information model, the value of *R0* is changing over time as opposed to the value of *R0* in the original model because of the behaviour change (Figure 9). Additionally, the Bayes factor ratio of model 1 to model 3 is less than 1, favouring model 3 for the fitness of the model to the data.

**Discussion**

It was demonstrated that both model 1 and model 2 exhibit a good fit for the observed data. Nevertheless, based on convergence diagnostic checks, model 2 appears to be more reliable than model 1. The posterior predictive checks for these models are similar, and they capture most of the data within their 95% CrI, although the priors for the *γ* between the two models are different. Additionally, the prior predictive checks results showed that the models are not excessively constrained by the priors. It is still possible to include a wide variety of situations and data, although it is likely that prior choices are too broad and include scenarios that are highly unlikely.

Prior distributions usually do not affect the sampling process as priors are updated throughout the inference, but it can shift the resulting posterior distributions away from likelihood [(10)](https://www.zotero.org/google-docs/?rK2VhK). Some distributions, e.g. normal distribution, allow RStan to explore parameter space from infinity. Thus, a flat or weakly informative prior should be recommended to be used as a placeholder before analyses or when there is a lack of information about parameter values [(17)](https://www.zotero.org/google-docs/?cKO7YU). However, in a situation where observation data is limited or the model is too complex, prior choice becomes important as it can highly influence the inference process. Thus, highly informative priors may significantly determine the statistical characteristics of the model and can lead to bias and model falsification [(10, 18)](https://www.zotero.org/google-docs/?fykQgb).

The diagnostic checks for model 3 showed that it exhibits a good fit for the observed data. Additionally, the Bayes factor ratio suggests that this model is preferred over the original model. The mean time delay of the behaviour change (*1/a*) is 454 days, indicating that the dynamics of the disease are minimally impacted by the number of infected individuals. However, the value of indicates that the information has a considerable influence on disease transmission. In the original model, *Rt* , which is an indicator for the disease transmission dynamics, has a constant *R0*. In the modified model, *R0* becomes a function of time. As a result, it is more appropriate to describe it as *Reff*, where .

**Conclusion**

All of the models provide a good fit for the data but the information model is better than the original model. Several diagnostic checks are available in RStan to guide the model development.

References

[1. Garnett GP, Cousens S, Hallett TB, Steketee R, Walker N. Mathematical models in the evaluation of health programmes. The Lancet. 2011 Aug;378(9790):515–25.](https://www.zotero.org/google-docs/?8J8pub)

[2. Grüne-Yanoff T. Choosing the right model for policy decision-making: the case of smallpox epidemiology. Synthese. 2021 May;198(S10):2463–84.](https://www.zotero.org/google-docs/?8J8pub)

[3. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Fam Med Com Health. 2020 Feb;8(1):e000262.](https://www.zotero.org/google-docs/?8J8pub)

[4. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the Basic Reproduction Number (R 0 ). Emerg Infect Dis. 2019 Jan;25(1):1–4.](https://www.zotero.org/google-docs/?8J8pub)

[5. Dietz K. The estimation of the basic reproduction number for infectious diseases. Stat Methods Med Res. 1993 Mar;2(1):23–41.](https://www.zotero.org/google-docs/?8J8pub)

[6. Heesterbeek JAP. A Brief History of R0 and a Recipe for its Calculation. Acta Biotheoretica. 2002 Sep 1;50(3):189–204.](https://www.zotero.org/google-docs/?8J8pub)

[7. Upton GJ, Cook IT. A Dictionary of Statistics. 3rd ed. Oxford: Oxford University Press; 2014.](https://www.zotero.org/google-docs/?8J8pub)

[8. Cole SR, Chu H, Greenland S. Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. American Journal of Epidemiology. 2014 Jan 15;179(2):252–60.](https://www.zotero.org/google-docs/?8J8pub)

[9. Rossi RJ. Mathematical Statistics : An Introduction to Likelihood Based Inference. New York: John Wiley & Sons; 2018.](https://www.zotero.org/google-docs/?8J8pub)

[10. van de Schoot R, Depaoli S, King R, Kramer B, Märtens K, Tadesse MG, et al. Bayesian statistics and modelling. Nat Rev Methods Primers. 2021 Jan 14;1(1):1.](https://www.zotero.org/google-docs/?8J8pub)

[11. Metropolis N, Rosenbluth AW, Rosenbluth MN. Equation of State Calculations by Fast Computing Machines.](https://www.zotero.org/google-docs/?8J8pub)

[12. Schmon SM, Gagnon P. Optimal scaling of random walk Metropolis algorithms using Bayesian large-sample asymptotics. Stat Comput. 2022 Apr 15;32(2):28.](https://www.zotero.org/google-docs/?8J8pub)

[13. Betancourt M. A Conceptual Introduction to Hamiltonian Monte Carlo [Internet]. arXiv; 2018 [cited 2023 Mar 8]. Available from: http://arxiv.org/abs/1701.02434](https://www.zotero.org/google-docs/?8J8pub)

[14. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. *Stan* : A Probabilistic Programming Language. J Stat Soft [Internet]. 2017 [cited 2023 Mar 8];76(1). Available from: http://www.jstatsoft.org/v76/i01/](https://www.zotero.org/google-docs/?8J8pub)

[15. Buonomo B, Della Marca R. Effects of information-induced behavioural changes during the COVID-19 lockdowns: the case of Italy. R Soc open sci. 2020 Oct;7(10):201635.](https://www.zotero.org/google-docs/?8J8pub)

[16. d’Onofrio A, Manfredi P. Information-related changes in contact patterns may trigger oscillations in the endemic prevalence of infectious diseases. Journal of Theoretical Biology. 2009 Feb;256(3):473–8.](https://www.zotero.org/google-docs/?8J8pub)

[17. Rufibach K, Burger HU, Abt M. Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development: K. Rufibach, H. U. Burger and M. Abt. Pharmaceut Statist. 2016 Sep;15(5):438–46.](https://www.zotero.org/google-docs/?8J8pub)

[18. García-Pérez MÁ. Bayesian Estimation with Informative Priors is Indistinguishable from Data Falsification. Span J Psychol. 2019;22:E45.](https://www.zotero.org/google-docs/?8J8pub)