Worksheet 3: Model Fitting With MCMC

u1624423

The model and likelihood function

Our population is $N = 55$ million people, each in one of the 3 SIR states. Our initial conditions are $(S(0), I(0), R(0)) = (N - \iota, \iota, 0)$ Each infected person is allocated an infectious period an recovery period drawn independently from a probability distribution. During the infectious period, each individual makes contact with every individual in the population at the points of a Poisson process with rate $\beta/(N - \iota)$. If someone in I contacts someone in S, then that individual begins their infectious period. At the end of an infected person's recovery period, they are moved to R. We assume the incidence of hospitalisation from a week ending on day t is $Y_t \sim \text{Poisson}(\delta I(t))$. Where $\lambda = \delta I(t)$, $P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$ is the probability of observing k counts given λ. With N data points, k_i (where $i = 1, ..., N$), the likelihood of observing a these points in the data for each λ_i, i.e. the likelihood function evaluates to $\mathcal{L} = \prod_{i=1}^{N} \frac{\lambda_i^{k_i} e^{-\lambda_i}}{k_i!}$. Computationally the sum of log likelihoods is used instead as it can be more performant and helps to avoid under/overflow issues. It is common in optimisation methods to minimise the "goodness" of a fit. The log likelihood is $log\mathcal{L} = \sum_{i=1}^{N}(k_i log(\lambda_i) - \lambda_i - log(k_i!))$. We drop the $log(k_i!)$ term as it is independent of the model, it will not affect the model parameters minimising $log\mathcal{L}$.

Parameters and prior information

The model parameters we are trying to estimate are $\boldsymbol{\theta} = (\beta, \gamma, \iota, \delta)$, which are contact rate, recovery rate, initial infections, and proportion of infected people that are hospitalised. The possible values of these parameters are $\beta \in (0, \infty)$, $\gamma \in (0, \infty)$, $\iota \in [0, N]$ and $\delta \in [0, 1]$. We can deduce suitable distributions to use as priors for these parameters, for β and γ we should use gamma distributions, for ι, an $N\times$ beta distribution and for δ, we should use a beta distribution. It is important to note that our priors should not be overly informative. We can use data outside of what we have to estimate reasonable priors. The CDC states RSV is contagious for 3-8 days [1], and recoverable in 7-14 days [2], Reis 2016 [3] supposes an R_0 of 3.0 with $\sigma = 0.6$. Thus for our priors for β, γ, it would be reasonable to consider the midpoints of these ranges. With some plotting of gamma distributions, we find that Gamma$(10.5, 1)$ has most of the mass between 7-14, the infectious period is $1/\gamma$, so we use InverseGamma$(10.5, 1)$ as our prior for γ. Similarly for R_0, Gamma$(30, 1/20)$ has most of the mass between 2.4-3.6, so our prior for β/γ is $\beta/\gamma \sim$ Gamma$(30, \gamma/10)$. For ι, we can't determine a specific amount, so we could choose a Uniform$(0, N) = N\times$ Beta$(1, 1)$. For δ, adults with RSV are typically not hospitalised or even reported as symptoms are much like that of a cold, but infants are. Medical journals such as Hall 2013 [4] suggest a 5.2% chance cases are hospitalised amongst infants. It would be reasonable to choose a distribution where the majority of the mass is about the point 0.052. I chose a distribution of Beta(2,22). Please see `logPrior.m` for my code.

MCMC algorithm

We apply an adaptive MCMC algorithm, where we start with $\boldsymbol{\theta_0}$ and Σ_0, our initial guess of the parameters and covariance of the posterior matrix. Here, after some testing, we find a good starting point of $\boldsymbol{\theta}_0 = (0.5, 0.1, 3, 0.08)$ and $\Sigma_0 = \text{diag}(\boldsymbol{\theta}_0) \times 10^{-}9$. Each iteration, we propose an "optimal" random walk (Gelman, Roberts, Gilks 1996 [5]), $\boldsymbol{\theta'} \sim \mathcal{N}\left(\boldsymbol{\theta}, 2.38^2/d + \epsilon I_d\right)$, where $d = 4$ is the dimension of $\boldsymbol{\theta}$ and ϵ is small (`mvnrnd` is used to draw multivariate random numbers from a normal distribution). We then proceed if the proposed conditions are

feasible, which requires $\boldsymbol{\theta'}$ to have the conditions $\beta, \gamma > 0$, $\iota \in [0, N]$, $\delta \in (0, 1)$, otherwise we reject the proposal. Note that the proposal distribution does not change. We then calculate log acceptance ratio, $LAR = log\mathcal{L}(\boldsymbol{\theta'}) - log\mathcal{L}(\boldsymbol{\theta}) + log\mathcal{P}(\boldsymbol{\theta'}) - log\mathcal{P}(\boldsymbol{\theta})$ (calling upon `logLikelihoodSIR.m` and `logPrior.m`). If $LAR > log(r)$, where $r \in (0, 1)$ is a uniform random number, we accept the proposal and set $\boldsymbol{\theta} = \boldsymbol{\theta'}$. Otherwise we reject this proposal. We then store our calculated posterior. If $LAR \leq log(r)$, we reject the proposal. The first n_0 iterations are the "burn-in" region. We estimate Σ_n at iteration n of MCMC, $(\boldsymbol{X_0}, ..., \boldsymbol{X_{n_0}})$ like so:

$$\hat{\Sigma} = \Sigma_n = \begin{cases} \Sigma_0, & n < n_0 \\ \text{Cov}(\boldsymbol{X_0}, ..., \boldsymbol{X_n}) + \epsilon I_d, & n \geq n_0 \end{cases}$$

For $n > n_0$, $\bar{\boldsymbol{X}}_n = (n/n+1)\bar{\boldsymbol{X}}_{n-1} + (1/n+1)\boldsymbol{X}_n$, used to calculate our covariance matrix at iteration n, $\Sigma_n = (n - 1/n)\Sigma_{n-1} + \frac{1}{n}(\boldsymbol{X}_n\boldsymbol{X}_n^T + n\bar{\boldsymbol{X}}_{n-1}\bar{\boldsymbol{X}}_{n-1}^T - (n+1)\bar{\boldsymbol{X}}_n\bar{\boldsymbol{X}}_n^T + \epsilon I_d)$.

Results & Convergence

The optimal acceptance rate (Roberts, Gelman, Gilks 1997 [6]) for $d = 1$ is 0.45 and for $d \geq 5$ is 0.234. With 10000 iterations, the same initial parameters discussed previously, and $n_0 = 250$, over 20 runs, the mean acceptance rate was 0.246 (3dp). This agrees with the literature. Here, $d = 4$, so it makes sense to have an acceptance rate slightly higher than that of when $d \geq 5$. Figure 1 is an example of the traceplots of the 4 parameters we are tuning, with an acceptance rate of 0.23043. Looking at these traceplots seem to show convergence, and the acceptance rate suggests that there is indeed convergence. We also look at the mean and variance of the chains to make sure they are (approximately) the same. Due to the stochastic nature, it is expected that convergence does not always occur, but better starting conditions can make it more frequent. Convergence typically occurs after around 6000 iterations. Figure 2 contains histograms of the posteriors. We can see that there is more than a single modal point for β and ι as they explore a large domain. Figure 3 is a plot of the data and our approximated model's median, with the 95% confidence interval. Although the data does not perfectly match up with the median, almost all of the data points are within the 95% confidence interval, meaning we have quite successfuly fit the data. The mean values of $(\beta, \gamma, \iota, \delta) = (0.457, 0.096, 0.063, 0.081)$ (3dp) over 20 runs. This would suggest $R_0 \approx 5$, the mean infectious period of about 10 days, which is not too dissimilar from our suggested priors. We can use these parameters to estimate a final size of the hospitalizations, giving us 4415, which is not too far from the value of 3632 given by the data.

Critical analysis of approach taken

There are various issues with the approach to fit the data to the SIR model. Primarily, the results we see from our approach depend on how correct our SIR is. We used a very basic SIR model as our system, which ignores important factors like demography, immigration and people leaving the country, variation of conditions due to age, people who are already immune to RSV, the potential introduction of a vaccine, seasonal differences, and behavioural changes, such as self isolation. Besides these, the SIR model makes other assumptions that are not necessarily the case. For example, β and γ are unlikely to be constant and every infected individual will experience different infectious and recovery periods. We also assume that the RSV data adheres to a Poisson distribution, which is potentially an unreasonable assumption. We also have not considered analytically what the optimal initial parameters would be. To determine this, we could find the peak of the $d = 4$ dimensional "hill" of $\boldsymbol{\theta}$. To find this point using MATLAB, we can simply use a d dimensional anonymous function of the log likelihood where the variable is $\boldsymbol{\theta}$ and minimise it by passing it to the `fminunc` function which finds the minimum of an unconstrained multivariable function. We have also not considered that the data itself can be inaccurate due to misreporting. Finally, priors are also a subjective choice, so there is no real correct option.

Figures and tables

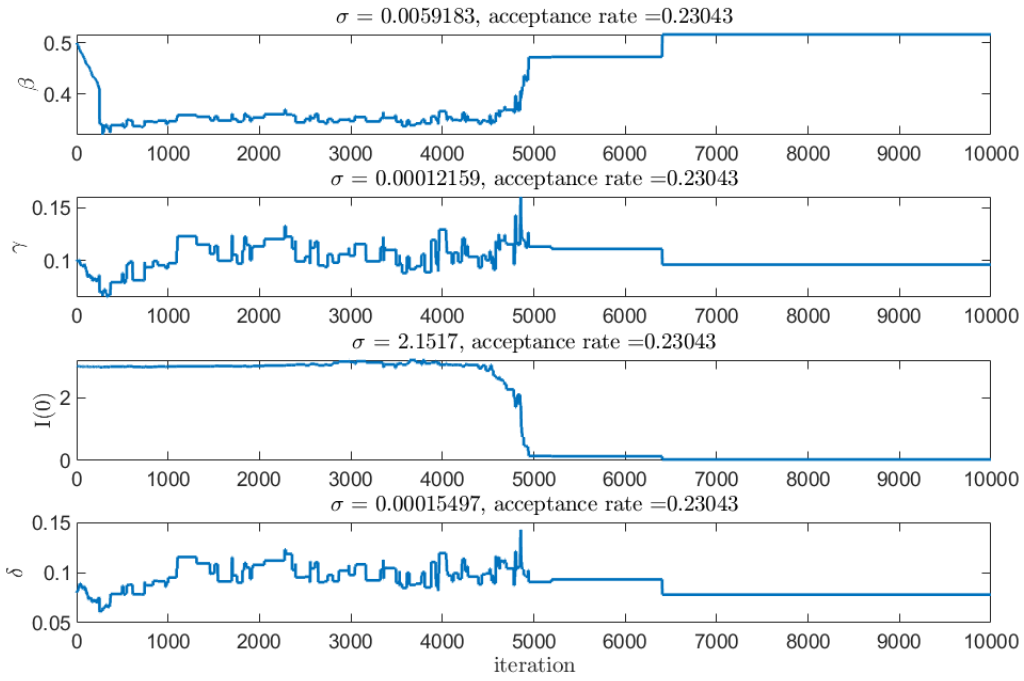

Figure 1: Traceplots of $\beta, \gamma, \iota, \delta$ with acceptance rate 0.23043, converging at iteration 6405.

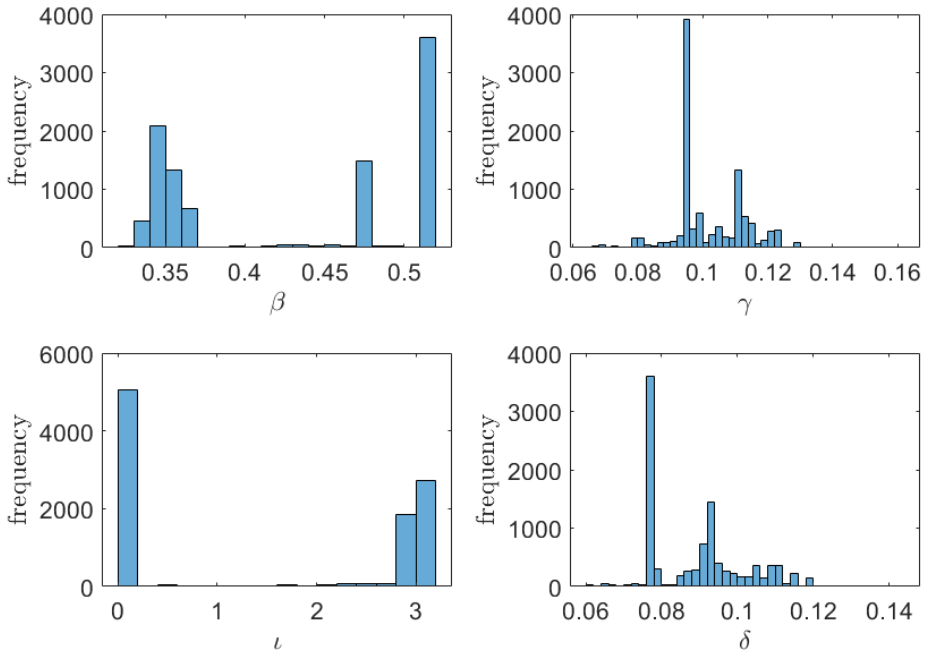

Figure 2: Histograms of posterior distributions of $\beta, \gamma, \iota, \delta$.

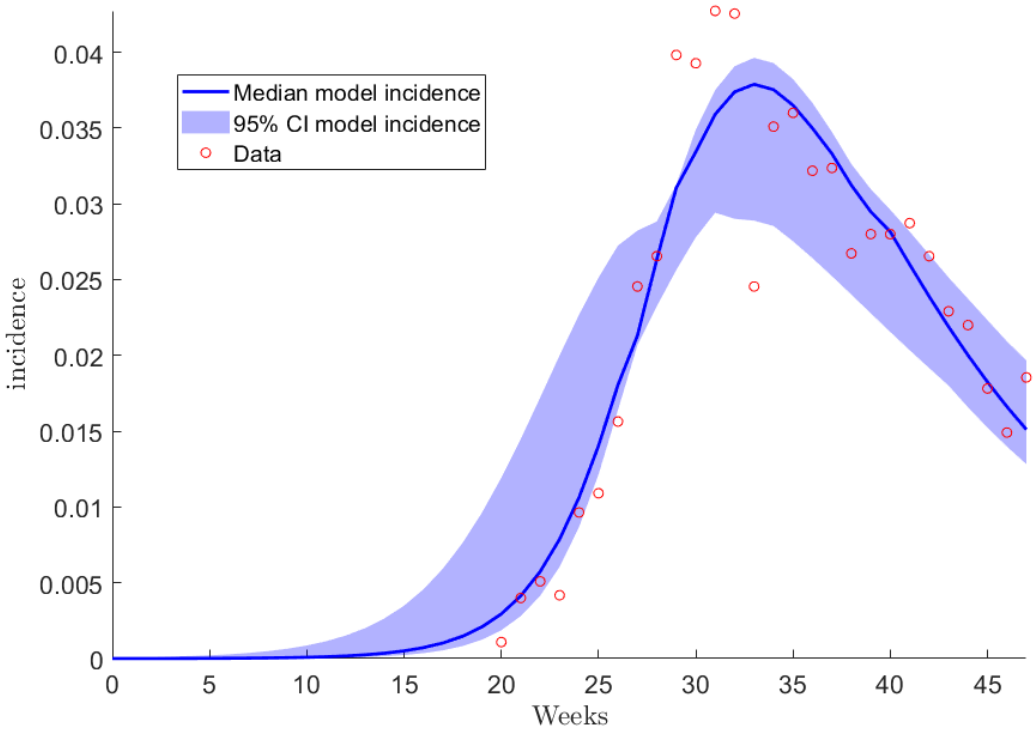

Figure 3: Model fitted to data with median and 95% confidence interval

Iteration	β	γ	ι	δ
0	0.5	0.1	3	0.08
1000	0.342	0.093	3	0.089
2000	0.354	0.113	3	0.108
3000	0.347	0.104	3	0.098
4000	0.367	0.129	3	0.119
5000	0.472	0.113	0	0.091
6000	0.473	0.111	0	0.093
6405	0.517	0.096	0	0.078

Table 1: Parameters change every 1000 iterations (3dp)

References

[1] CDC page on RSV Transmission, https://www.cdc.gov/rsv/about/transmission.html

[2] CDC page on RSV symptomps, https://www.cdc.gov/rsv/about/symptoms.html

[3] Retrospective Parameter Estimation and Forecast of Respiratory Syncytial Virus in the United States, 2016, Reis, Julia; Shaman, Jeffrey L., https://doi.org/10.7916/D8862GZP

[4] Respiratory syncytial virus-associated hospitalizations among children less than 24 months of age, 2013, Caroline Breese Hall 1, Geoffrey A Weinberg, Aaron K Blumkin, Kathryn M Edwards, Mary A Staat, Andrew F Schultz, Katherine A Poehling, Peter G Szilagyi, Marie R Griffin, John V Williams, Yuwei Zhu, Carlos G Grijalva, Mila M Prill, Marika K Iwane, https://doi.org/10.1542/peds.2013-0303

[5] Efficient Metropolis Jumping rules, 1996, G. O. Roberts, A. Gelman and W. R. Gilks, http://people.ee.duke.edu/ lcarin/baystat5.pdf

[6] Weak convergence and optimal scaling of random walk Metropolis algorithms, 1997, A. Gelman, W. R. Gilks, G. O. Roberts, https://doi.org/10.1214/aoap/1034625254