The Application of Markov Chain Monte Carlo to Infectious Diseases

Alyssa Eisenberg

March 16, 2011

Abstract

When analyzing infectious diseases, there are several current methods of estimating the parameters included in models. However, we propose a newer method involving Bayesian inference and then Markov chain Monte Carlo to estimate the parameters. This allows for more model flexibility and results in more accurate results, especially when there is limited data available.

We will describe the process itself, and then work through two applications of the process to a measles outbreak in Providence, Rhode Island. These applications are very similar, but differ in their level of generality.

1 Introduction

In order to inform our health policies and mechanisms to stop the spread of disease, we must first understand how the disease operates on its own. One of the main techniques used to study infectious diseases is modelling. Beginning from a set of observations, we build a model that specifies the mechanism of the spread of the disease using certain variables. These can include latent periods, variable infectivity rates or natural immunity, immunity upon recovery, time of infection, and more. The variables depend on what disease is being modelled and the delicate balance between including enough to make the model realistic and simplifying it enough to allow us to analyze it.

The model incorporates our knowledge but does not inform us about the disease. The objects of interest are left as variables. The next step becomes attempting to find the values of these variables given a set of data. If the data set were complete, this would be fairly easy. However, the data available from infectious disease outbreaks tends to be minimal. Sometimes we have data from multiple outbreaks of the same disease, and other times we only have data from one larger outbreak.

Generally, the data is one of three types: non-temporal, semi-temporal, or temporal. Non-temporal data is simply the number of cases in a population of a known size. Semi-temporal data is an infection chain, where each entry is a the number of infected individuals in that generation of the infection. Temporal data includes the times at which infections are detected (i.e. when symptoms begin to occur). In this paper, we will focus on semi-temporal data.

There is no standard process of estimating the values of our initial model parameters given the limited data. One process that had been used is maximal likelihood [1], but the estimations are not very accurate. Instead, we consider the use of Bayesian inference and Markov chain Monte Carlo in order to obtain better estimates and allow for more flexibility in the models.

2 Overall Process

We begin with our model and our data set. We then perform Bayesian inference to formulate the problem, and then use Markov chain Monte Carlo techniques to estimate the solution. The following explanation of Bayesian inference comes from [2].

In Bayesian inference, we begin with the prior probabilities for the parameters, also called a prior. These are the probability distribution of the unknown parameters before considering the data. They can be entirely uninformed, such as a general uniform distribution or beta distribution. Or, they could be informed based on previous attempts to estimate the variables or expert opinions. The process will work with either choice, but it could work faster given priors that are closer to the actual distribution of the variable in question. So if our unknown parameters are x and y, we would choose priors $\mathbb{P}(x)$ and $\mathbb{P}(y)$.

We also form an equation to calculate the likelihood function for any given set of parameters. The likelihood is the conditional probability of the data given that set of parameter values. Thus if D is the data, we have the same parameters as above, and L is our likelihood function, $L(x, y) = \mathbb{P}(D|x, y)$.

Finally, we would calculate the posterior distribution of the parameters. This is the conditional distribution of the parameter values given the data. By Bayes' Theorem,

$$\mathbb{P}(x, y|D) = \frac{\mathbb{P}(D|x, y) * \mathbb{P}(x) * \mathbb{P}(y)}{\mathbb{P}(D)}$$

By calculating this, we could estimate the values of the parameters we are looking for. Unfortunately, the normalizing factor, $\mathbb{P}(D)$, is computationally difficult if not impossible to calculate when dealing with infectious diseases. There simply is not enough information and there are too many possibilities of parameters. But, it means that we know that the posterior distribution is proportional to the likelihood times the priors.

At this point, we will use Markov chain Monte Carlo in order to sample from the target distribution (the posterior distribution). The general technique used is the Metropolis-Hastings algorithm. This involves choosing a proposal density Q that may depend only on the current values of the parameters. Once new values are generated, they are accepted with a certain probability and otherwise are left with their current values. The parameters can either be updated as a block, or they can be updated individually in a sequence. Let $\mathbb{P}$ be the target density, x be the original values of the parameters, and x' be the updated values. Then,

$$\mathbb{P}_{accept} = \frac{\mathbb{P}(x') * Q(x|x')}{\mathbb{P}(x) * Q(x'|x)}$$

Notice that the normalizing constant $\mathbb{P}(D)$ does not need to be calculated, and instead we only need to calculate something proportional to the posterior distribution. For more details about the Metropolis-Hastings algorithm, see [3].

This methodology is incredibly useful for modelling infectious diseases. It allows more flexibility than other techniques in the number of unknown parameters. Thus, the model could be made more realistic or missing data could also become a parameter. The method also allows analysis of any and all parameters involved.

3 Example: Measles

This example, from [4], comes from semi-temporal data about a measles outbreak in Providence, Rhode Island. Here is the infection chain data for households with three people:

Chain	Frequency
{1}	34
{1,1}	25
{1,2}	239
{1,1,1}	36

In the model, there are N households of three people, where one person always starts infected in the first generation. This makes up the entire population under consideration. Within each household j, there is a probability q_j that an infected individual will not pass on the infection to a susceptible individual in the household. Each susceptible individual in a household with at least one infected individual will undergo one infection attempt for that generation, after which the infected individuals become immune. Each q_j is drawn from the random beta distribution Q on parameters α and β.

Basing our inference on this model, there are two unknown parameters we need to find: α and β, which combine to give us a distribution for the real variable of interest, Q. Since these parameters are a level behind the variable we want to find, they are called hyper-parameters. Their hyper-priors were chosen to be gamma distributed priors with mean of 1 and variance of 1000. Let us call these uninformative hyper-priors p.

We also need to calculate the likelihood function, where $L(\alpha, \beta) = \mathbb{P}(n_1, n_{11}, n_{12}, n_{111}|\alpha, \beta)$. n_1 is the number of households that were recorded with the $\{1\}$ chain of infection in the actual data, and the others follow the other possible chains of infection. By basic probability, we see:

$$L(\alpha, \beta) = (\mathbb{E}[Q^2])^{n_1})(2\mathbb{E}[(1-Q)Q^2])^{n_{11}}(\mathbb{E}[(1-Q^2])^{n_{12}}(2\mathbb{E}[(1-Q)^2Q])^{n_{111}}$$

Replacing the expected values with their function in terms of α and β yields:

$$\begin{aligned} L(\alpha, \beta) = &\left(\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\right)^{n_1}\left(\frac{2\alpha\beta(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)(\alpha+\beta+2)}\right)^{n_{11}} \\ &\times\left(\frac{\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)}\right)^{n_{12}}\left(\frac{2\alpha\beta(\beta+1)}{(\alpha+\beta)(\alpha+\beta+1)(\alpha+\beta+2)}\right)^{n_{111}} \end{aligned}$$

From Bayes' Theorem, we know that the posterior distribution, $\pi(\alpha, \beta)$ is proportional to the likelihood times the hyper-priors. Thus,

$$\pi(\alpha, \beta) \propto L(\alpha, \beta)p(\alpha)p(\beta) = h(\alpha, \beta)$$

Once we have these equations for $L(\alpha, \beta)$, $p(\alpha)$, $p(\beta)$, and $h(\alpha, \beta)$ we are ready to apply the Metropolis-Hastings algorithm. For this, we need to choose a proposal density for updating the values. We chose g, the Gaussian probability densities for α and β centred on their current values with variance $\sigma^2 = .01$. This algorithm chooses to do block updates, so new values α' and β' are chosen at the same time. Then the probability of acceptance becomes:

$$\mathbb{P}_{accept} = min\left(\frac{h(\alpha', \beta')g(\alpha, \beta|\alpha', \beta')}{h(\alpha, \beta)g(\alpha', \beta'|\alpha, \beta)}, 1\right)$$

Thus, the exact steps are to choose any initial values for α and β such that $h(\alpha, \beta) > 0$). Then, update both values by choosing a new value from the proposed density g, and accepting the proposed values with $\mathbb{P}_{accept}$. Otherwise, the original values do not change. These steps form a Markov chain with the target density (the posterior distribution) as its stationary distribution.

Therefore, the results came from a sample of 10000 points from the output of the above Markov chain. The results, $\hat{\alpha} = .276$ and $\hat{\beta} = 1.143$, agree with previous estimates from other methods. This implies that the methodology of using Bayesian inference and Markov chain Monte Carlo on infectious disease modelling is valid. However, we have to be careful because there was a strong correlation between α and β that violates our assumption of independence when we chose different priors for the two variables. This correlation also means that the chain may mix slower, and thus only reach the stationary distribution from which we can sample after a lot more steps.

4 Extension of the Measles Example

The example above demonstrated the use of the application of Markov chain Monte Carlo to infectious disease modelling in a very specific case. In [5], a similar but more general model is developed and then applied to the same Providence measles data as above.

The model of disease spread was very similar to the previous one described. The difference is that instead of setting the households to have three members of which one begins infected, this model generalizes this to each household having M members with i_0 initial infected individuals. The observed data is $\mathbf{y}_\bullet = (y_{\bullet 1}, \ldots, y_{\bullet k})$, where there are k possible infection chains and each $y_{\bullet j}$ is the number of households observed with infection chain j. The probability q_j of avoiding infection in household j remains the same as previously, with Q as the beta distribution $B(\alpha, \beta)$.

They chose the independent hyper-priors $\gamma(\alpha) = \text{Gamma}(2,10)$ and $\gamma(\beta) = \text{Gamma}(1.5,10)$. The likelihood is

$$\psi(\alpha, \beta) = \prod_{j=1}^{k}\left(\frac{B(\alpha_j, \beta_j)}{B(\alpha, \beta)}\right)^{y_{\bullet j}}$$

where $\alpha_j = \alpha + \sum_{l=1}^{g_j+1} s_j(l)$ and $\beta_j = \beta + M - i_0 - s_j(g_j)$. The term $s_j(l)$ is the number of susceptible people left after generation l in a household with chain j, and g_j is the number of infection generations in chain j. For more details about the derivation of this, see [5].

Finally, for the posterior distribution we know:

$$f(\alpha, \beta|\mathbf{y}_\bullet) \propto \gamma(\alpha)\gamma(\beta)\psi(\alpha, \beta)$$

At this point we apply the Metropolis-Hastings algorithm to estimate the values for α and β. The proposal densities for α and β are the same as their hyper-priors. In this paper, they are updated separately instead of in a block. We start with arbitrary initial values for α and β. Then, we generate α' from $\gamma(\alpha)$. We accept this value with a probability:

$$\mathbb{P}_{accept} = min\left(\frac{\psi(\alpha', \beta)}{\psi(\alpha, \beta)}, 1\right)$$

If accepted, α' becomes the new value of α, otherwise it remains the same. Then, we update β in a similar fashion. Choose a β' from $\gamma(\beta)$ and accept it with probability:

$$\mathbb{P}_{accept} = min\left(\frac{\psi(\alpha, \beta')}{\psi(\alpha, \beta)}, 1\right)$$

Otherwise, do not change β. This forms the Markov chain with the posterior distribution as its stationary distribution.

Results from this methodology using sample data showed a slow mixing time where 7640 iterations were required for convergence. Additionally, as before α and β were highly correlated. Thus, a new parameterization was chosen with hyper-parameters $\mu = \frac{\alpha}{\alpha+\beta}$ and $\rho = \frac{1}{\sqrt{\alpha+\beta+1}}$. For the details of adapting the previous calculations and algorithm for the new parameters, see [5]. The results from this new parameterization mixed much more rapidly and the two parameters were not correlated, thus fixing our two main problems. When applied to the actual Providence measles data, the expected values for the different chains were:

Chain	Observed	Calculated
{1}	34	35.1
{1,1}	25	23.0
{1,2}	239	237.7
{1,1,1}	36	38.2

This demonstrates that with the correct model and parameterization, the Bayesian inference and Markov chain Monte Carlo process can deliver accurate estimates with reasonable convergence times.

5 Conclusion

In conclusion, the Bayesian inference and Markov chain Monte Carlo method has been demonstrated to be useful in analyzing infectious disease data. As we mentioned from the beginning, one of the advantages of the Markov chain Monte Carlo method is that it allows for a lot of flexibility in the model and in allowing for missing data. However, as a word of caution, we also have to be careful to construct parameters and an algorithm that will converge in a reasonable amount of time, which is not necessarily always obvious.

The examples we discussed focused on modelling the transmission of a disease with variable infectivity depending on household. But this method is applicable in a much more general sense than that. An extension of the current examples might be to switch from the Greenwood model where the probability of infection only depends on there being an infective agent to the Reed-Frost model where the probability of the spread of infection depends on the number of infected individuals in the household. Other applications may include other variables, such as the possibility of infection from the community, or variable immunity among the population to begin with. Or, when dealing with temporal data, one could use this method to determine the unknown infection times instead of the times at which symptoms appeared. Finally, this methodology can be applied to diseases other than measles as long as there is a model that describes the interaction between people and the spread of the disease.

References

[1] I.M. Longing and J.S. Koopman (1982), Household and Community Transmission Parameters from Final Distributions of Infections in Households, *Biometrics* **38**, 115–126.

[2] C.M. Grinstead and J.L. Snell (1997), *Introduction to Probability*, 2nd edition, American Math Society, USA.

[3] S. Chib and E. Greenberg (1995), Understanding the Metropolis-Hastings Algorithm, *Th American Statistician* **49**, 327–335.

[4] P.D. O'Neill (2002), A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods, *Mathematical Biosciences* **180**, 103–114.

[5] L. Ning, G. Qian and R. Huggins (2003), A random effects model for diseases with heterogeneous rates of infection, *Journal of Statistical Planning and Inference* **116**, 317–332.