Markov Chain Monte Carlo Methods

ECEN 662 Final Project Report

Viswam Nathan

Kai He

Justin Lewis

OUTLINE

I. INTRODUCTION:

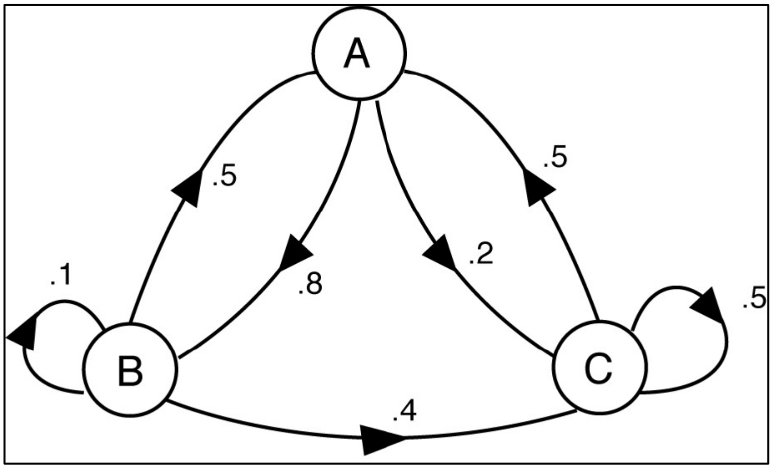
A. Motivation:

In short, Markov Chain Monte Carlo methods facilitate calculations. They circumvent problems which are grisly to calculate directly by producing rapid approximations of the exact. More specifically, MCMC methods hope to approximate things such as multi-dimensional integrals, statistics, and distributions which are commonly found within the Bayesian statistics framework. With hundreds or even thousands of parameters, these intractable structures prove to be a practical hindrance of the Bayesian viewpoint. By utilizing results from the study of Markov chains and ergodic theory, MCMC methods tackle this problem in a pragmatic, yet clever fashion.

B. Markov Chains:

To understand MCMC methods, a brief review of Markov chains is helpful. Simply put, a Markov chain models a random process. This process is characterized by a set of states S and transitions T.

Figure 1): Simple Markov Chain



Within this simple example, there are three possible states $S = \{A, B, C\}$. The arrows between states represent the transitions T. It should be noted these transitions are based only on the current state. Knowledge of previous states does not affect the process's behavior. This

Markov assumption manifests itself through modeling the transition from the current state $s_{current}$ to the next state s_{next} as a random variable X with probability mass or density function $f_X(x)$. As such, a Markov chain can be viewed as a series of random variables.

Definition 1): Discrete-Time Markov Chain

$Random\ Process$ $\qquad$ $s.t.$

$M \triangleq \{X_1, X_2, \dots, X_N\}$ $\qquad$ $X_i \triangleq R.V.under\ f_{X_i}(x_i)$

$Markov\ assumption$

$$\Pr(X_{i+1} | X_1 = x_1, X_2 = x_2, \dots, X_i = x_i) \;=\; \Pr(X_{i+1} | X_i = x_i)$$

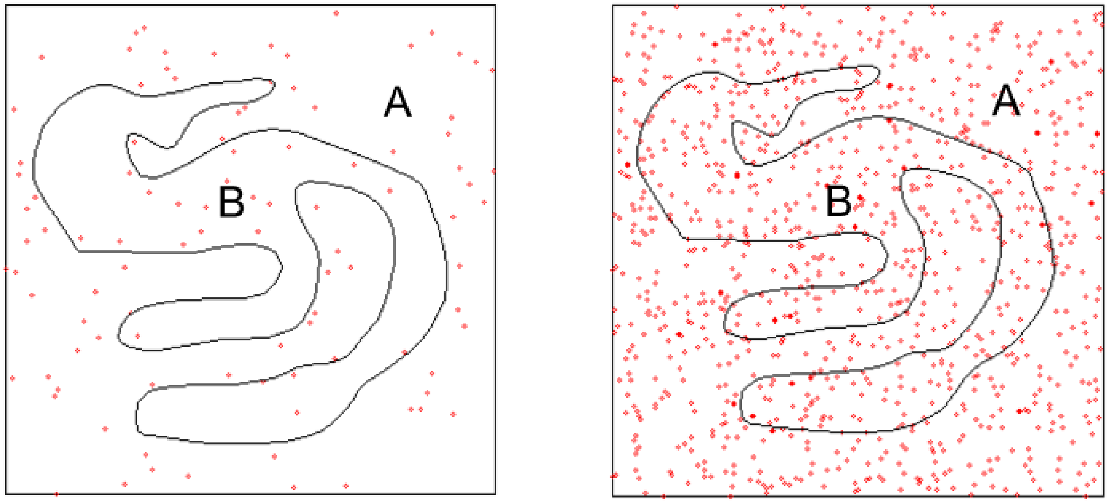
$State\ space$

$$X_i = x_i \quad s.t. \quad x_i \in S \quad \forall\ i = 1,2, \dots N$$

C. Monte Carlo Methods:

The second piece of MCMC is understanding the motivation behind Monte Carlo methods. The goal of common Monte Carlo methods is to use random sampling to approximate a difficult computation. A common example is numerical integration of a non-standard, closed space.

Figure 2): Monte Carlo Integration



In figure 2, the seemingly difficult task of calculating the area of the closed space B is made simple by Monte Carlo methods. First, samples are generated uniformly from a space A of known area. Each sample is determined as either inside or outside the closed space B. After a number of iterations, the area of B can be approximated as:

Equation 1): *Monte Carlo Integration Example*

$Given\ closed\ areas\ A, B,$ and $n\ samples \sim U([x_0, y_0], [x_1, y_1])$:	$s.t.$ $m = num\ samples\ \in B$ $n = total\ num\ samples$
$area_B \approx area_A \frac{m}{n}$	

D. MCMC Overview:

So how can these two concepts, Markov chains and Monte Carlo approximation, be used to calculate multi-dimensional integrals? In a Monte Carlo-esque fashion, these integration problems can be solved by generating random samples from an intractable distribution $f_{\vec{X}}(\vec{x})$. These samples are generated by constructing a Markov chain which possesses a continuous state space over the support of $f_{\vec{X}}(\vec{x})$. As such, the Markov chain will traverse the distribution's support through a structured manner. The method by which the chain determines its next state is the key component of MCMC methods.

To determine the structure of the desired Markov chain, ergodic theory is employed. Ergodic theory is interested in the long term behavior of a random process. A random process which is ergodic behaves in a way such that its long term behavior is independent of initial state. In order to prove that a Markov chain is ergodic, it must be shown that the chain is aperiodic, positive recurrent, and irreducible.

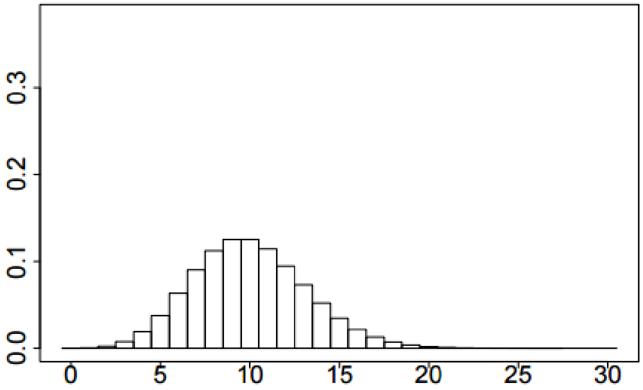
Once constructed, an ergodic Markov chain eventually converges to a stationary distribution. If the Markov chain is created correctly, this stationary distribution can match the distribution of interest $f_{\vec{X}}(\vec{x})$. More precisely, the probability distribution of the chain converges to $f_{\vec{X}}(\vec{x})$ after a sufficient number of iterations. This behavior is shown through the following example:

Problem 1): Estimate Simple Poisson distribution

$Let\ f_X(x) = \frac{10^x}{x!} e^{-10}$

$Estimate\ f_X(x)\ using\ MCMC\ methods$

Figure 3): Plot of $f_X(x)$



Solution:

$Randomly\ initialize\ x_0\ \in [1,3]$

$Markov\ Chain$:

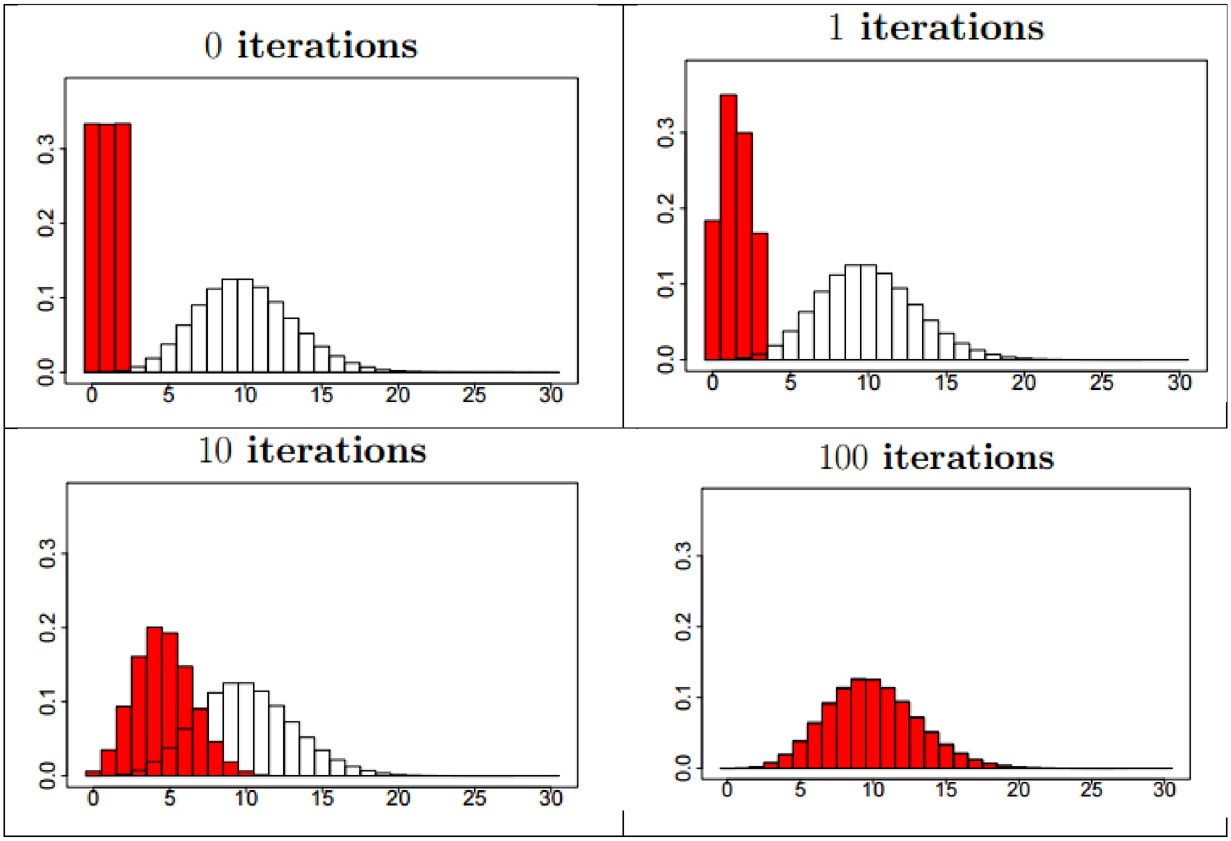
$$\mathbf{P}(x_{i+1} = x_i - 1 | x_i) = \begin{cases} x_i/20 & \textbf{if } x_i \leq 9, \\ 1/2 & \textbf{if } x_i > 9 \end{cases}$$

$$\mathbf{P}(x_{i+1} = x_i | x_i) = \begin{cases} (10 - x_i)/20 & \textbf{if } x_i \leq 9, \\ (x_i - 9)/(2(x_i + 1)) & \textbf{if } x_i > 9 \end{cases}$$

$$\mathbf{P}(x_{i+1} = x_i + 1 | x_i) = \begin{cases} 1/2 & \textbf{if } x_i \leq 9, \\ 5/(x_i + 1) & \textbf{if } x_i > 9 \end{cases}$$

The above Markov chain is a result which comes from a common solution to the MCMC problem known as the Metropolis Hastings algorithm. This solution comes from utilizing a principle known as the detailed balance condition. This sufficient condition ensures that the Markov chain constructed possesses a unique stationary distribution. Under this conditioned viewpoint, a unique formulation becomes fairly straightforward. This solution and the Metropolis Hastings algorithm will be explained in more detail in Section II.

Figure 4): Plots of estimate $f_{\hat{X}}(\hat{x})$ over $f_X(x)$



The plots above demonstrate how the use of a Markov chain can be used to estimate a probability distribution. The usefulness of this particular solution is of little significance considering the distribution is one dimensional; however, it is valuable for instructive reasons. One can see that as the number of iterations increases the estimate improves until convergence is reached. The question of when convergence is reached is very important, especially for distributions of high dimension.

Ultimately, Markov chains prove to be a very practical method for generating samples from a distribution which is analytically intractable. Once these samples are generated, they can be utilized to approximate a distribution, calculate a statistic, or compute an integral. The exact details by which these things are accomplished will be explained in the following sections. This first section stands as a brief exposition of the subject.

II. MCMC ALGORITHM: METROPOLIS HASTINGS

In this section we provide an overview of one of the popular MCMC algorithms: Metropolis Hastings. The basic idea of the Metropolis Sampling algorithm is to simulate a posterior sample from a probability distribution by making use of

1) Acceptance Function $\alpha\left(\rho^{(i)}\middle|\rho^{(i-1)}\right) = \min\{1, \frac{p(\rho^{(i)}|x)}{p(\rho^{(i-1)}|x)}\}$, in which full joint density function $p(\rho|x) \propto p(x|\rho)p(\rho)$;

2) Proposal distribution: $q(\rho^{(i)}|\rho^{(i-1)})$.

Generally speaking, the Metropolis Hastings Algorithm is to randomly draw a candidate following a certain rule (distribution), and decide whether to accept this candidate or not based on whether it's more likely to happen or not. Here we illustrate the objectives and how the two important functions work in the Metropolis Hastings Algorithm.

A) Proposal Function

The concept of the proposal function is to simulate a candidate 'sample' from a certain region or distribution, which is called a 'proposal distribution'. The randomly generated sample will not always be accepted by the algorithm. Whether to accept or reject the simulated value for the posterior sample depends on the acceptance function α(x). Generally speaking, the simulated candidate achieving higher posterior probability will have more chance to get accepted by the algorithm. Metropolis Hastings Algorithm mainly works with two types of proposal distribution, symmetric or asymmetric one. For symmetric proposal distribution, we have the important property that $q(x_i\,|x_{i-1}) = q(x_{i-1}\,|x_i)$ which means that the distribution of the next step conditioned on the current one is the same as the current one conditioned on the next step. Many distributions like Normal Distribution are very straight forward symmetric proposal distributions. For Normal distribution $N(0,\sigma)$, we have $x_i = x_{i+1} + N(0,\sigma)$, which is a symmetric proposal since $N(x_i - x_{i-1},\sigma) = N(x_{i-1} - x_i,\sigma)$. Symmetric proposal distributions randomly "walk" to a state, and the simulated value will be accepted or rejected based on acceptance function α. This kind of perturbing is generally called 'Random-walk Metropolis Algorithm'.

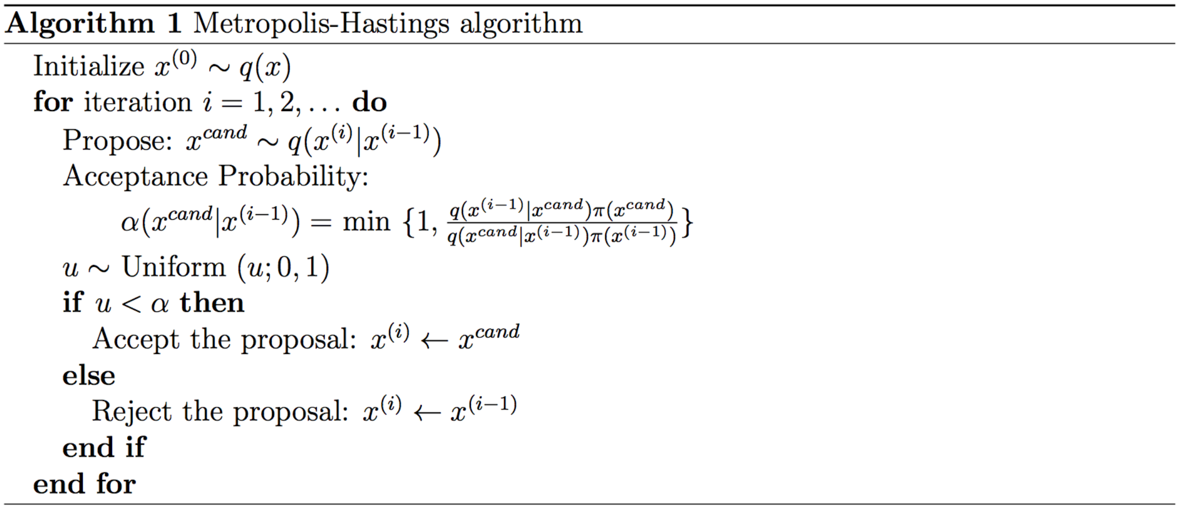
Although Random Walk is the most common proposal distribution for Metropolis Hastings algorithm, sometimes people still need to work with asymmetric proposal distribution when catering to certain data. When the property of the candidate is not symmetric, like if we need to estimate a variance, which is always greater than zero; in such case, a symmetric proposal distribution can no longer be used for us to estimate the posterior distribution.

B) Acceptance Function

Acceptance function has two objectives, one is to lead the sampler to go to the region with higher probability ($\frac{\pi(x^{cand})}{\pi(x^{i-1})}$); another one is to encourage the sampler to explore the space and avoid the situation that the sampler will get stuck in a certain region. ($\frac{q(x^{(i-1)}|x^{cand})}{q(x^{cand}|x^{i-1})}$). With symmetric proposal distribution, the acceptance function becomes proportional to how likely the current state $x^{(i-1)}$ and the next state x^{i} under the full joint distribution. Since the proposal distribution can also be symmetric, under that circumstance, acceptance function takes care of both constraints.

C) Implementation of Metropolis Hastings Algorithm

Here is a psuedocode for Metropolis Hastings Algorithm



Algorithm 1 Metropolis-Hastings algorithm

```
Initialize $x^{(0)} \sim q(x)$
for iteration $i = 1, 2, \ldots$ do
    Propose: $x^{cand} \sim q(x^{(i)}|x^{(i-1)})$
    Acceptance Probability:
        $\alpha(x^{cand}|x^{(i-1)}) = \min \{1, \frac{q(x^{(i-1)}|x^{cand})\pi(x^{cand})}{q(x^{cand}|x^{(i-1)})\pi(x^{(i-1)})}\}$
    $u \sim$ Uniform $(u; 0, 1)$
    if $u < \alpha$ then
        Accept the proposal: $x^{(i)} \leftarrow x^{cand}$
    else
        Reject the proposal: $x^{(i)} \leftarrow x^{(i-1)}$
    end if
end for
```

First, we need to initialize the first step of the chain, which usually randomly draws from the prior distribution. In the main loop, we generate the candidate from the proposal distribution $x^{cand} \sim q(x^{i}|x^{i-1})$; the candidate is then plugged into the acceptance function to get the acceptance probability. After we get the acceptance probability α, we randomly generate a value u from a uniform distribution of [0,1], and compare u and acceptance probability α. The proposal is accepted if the acceptance probability α is greater than u; rejected otherwise.

D) A Simple Example

In order to get practical experience for the Metropolis Hastings Algorithm, we consider two streams of observation $x^{1:N}$ and $y^{1:N}$which follow the multivariate Normal Distribution $N(0,\sigma)$ with variance $\sigma_{xx} = 1$, and $\sigma_{yy} = 1$. We are interested in estimating the covariance parameter ρ_{xy}. Following multivariate Normal distribution with the above the parameters, we have the joint distribution $p(x_i, y_i|\rho) = \Pi \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}[x_i^2 - 2\rho x_i y_i + y_i^2]\}$. In order to have a fuller specification of this model, we need to decide the prior distribution of the candidate need to be estimated. The prior for covariance matrices is 'Jefferys' prior, which takes $1/|\Sigma|^{3/2}$. In our case, it takes the following form:

$$p(\rho) = \frac{1}{|\Sigma|^{3/2}} = 1/(1-\rho^2)^{3/2}$$

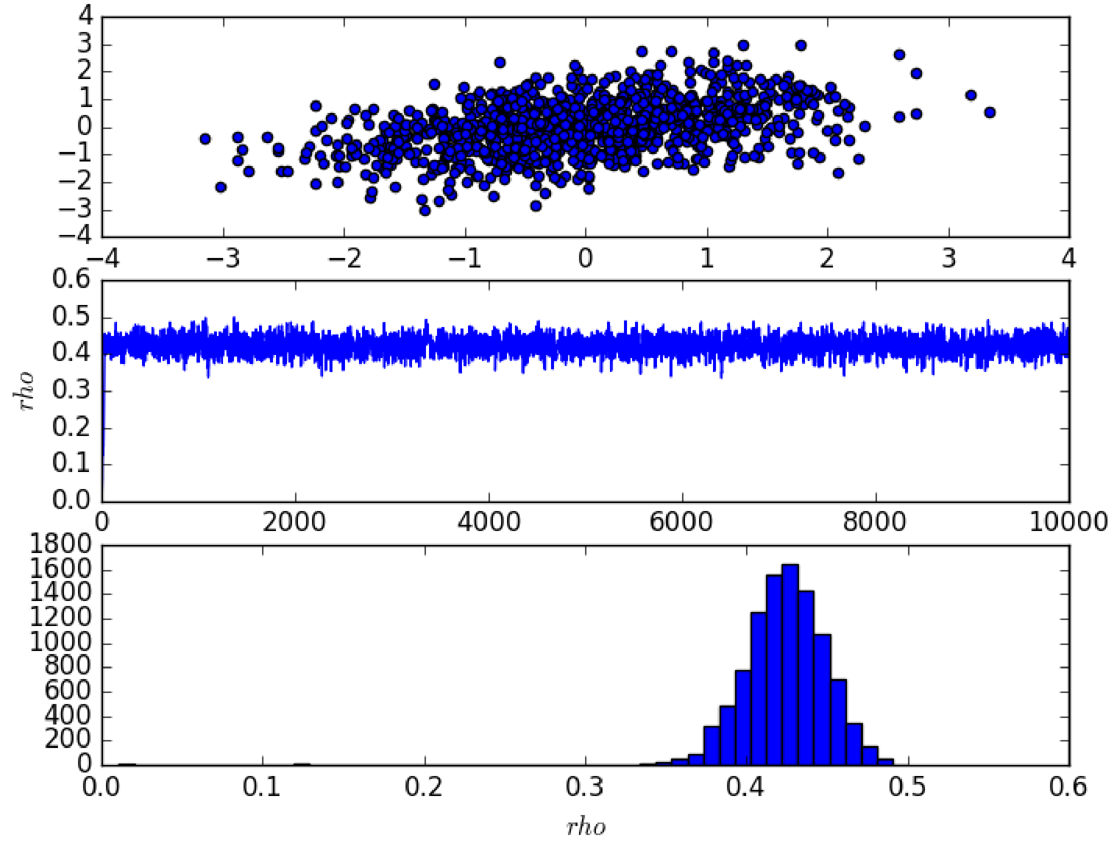
Then we need to compute the posterior distribution. However, since the computation of the integral of $\int p(x_i, y_i|\rho)d\rho$ is intractable, we can only compute $\pi(x)$ which is proportional to the posterior distribution.

$$p(\rho|x_i, y_i) \propto 1/(1-\rho^2)^{3/2}\, \Pi \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}[x_i^2 - 2\rho x_i y_i + y_i^2]\}$$

After we have $\pi(x)$ there is the question about the choice of proposal distribution and acceptance function. For proposal distribution, we simply use a uniform distribution. For the acceptance function, since the proposal distribution is symmetric, it only takes care of the $\pi(x)$ which aims to achieve the higher posterior probability.

The implementation of this model contains the following procedures. First, 1000 pairs of (x,y) are generated by a multivariate-normal function following the distribution $x_i y_i|\rho \sim N(0, \Sigma)$. In which $\Sigma = [\begin{matrix} \sigma_{xx} & \rho \\ \rho & \sigma_{yy} \end{matrix}]$, and $\rho = 0.4$, $\sigma_{xx} = \sigma_{yy} = 1$. Since the parameter of interest has no constraint on the value it can assume, a uniform distribution centered in the current candidate's value with overall width of 0.14 is used for the proposal distribution. It should be noted that the proposal distribution is not unique, we can either take different width for uniform distribution, or we can take other distribution like Normal distribution to perturb. After the proposal distribution is selected, we have the acceptance function to be $\alpha\left(\rho^{(i)}\middle|\rho^{(i-1)}\right) = \min\{1, \frac{p(\rho^{(i)}|x^{1:N},y^{1:N})}{p(\rho^{(i-1)}|x^{1:N},y^{1:N})}\}$.

10,000 is set to be the number of steps of walk. The first graph in the figure below shows the 1,000 generated pairs of (x,y). The trace of ρ can be seen in the second graph, from which we see that the chain converges immediately. Finally, the bottom row is the histogram of ρ, which is the MCMC estimate of its posterior distribution.



III. APPLIED MCMC RESEARCH PAPER REVIEW

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A. J. and Boëlle, P. Y. (2004), A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statist. Med., 23: 3469–3487. doi:10.1002/sim.1912

We provide a summary of this paper, highlighting especially the MCMC components, as an example of how this technique can be applied to model and solve a real-world problem involving limited observations and complex, undefined probability distributions for the parameters of interest.

A) Background

In this paper, studying the spread of influenza within a household, as well as among several households in a community is of interest to the authors. Specifically, the authors would like to know the average duration of infectious period for each individual in the community. The problem is that the start and end dates of the infection are not known directly; this is because the only observations available in an uncontrolled setting usually are the times when symptoms occur and are reported to physicians. However, the virus itself may be infecting the individual (and hence potentially transmit to others) for a few days before and after the symptoms manifest. Since there is no systematic way to bracket the symptomatic period with the infectious period, the research community has relied on MCMC techniques to explore the possible model parameters for inference.

Additionally of interest to the authors are the susceptibility and infectiousness of individuals, which are respectively numbers ranging from 0 to 1 which indicate the probability of catching the flu and infecting someone else with the flu respectively. This can be used to instantaneously evaluate the risk for an individual from the community and from the household on any given day. There is no existing probabilistic model for these parameters either, and hence again the authors rely on MCMC to approximate the posterior distribution. The authors also investigated some supplementary aspects of influenza transmission dynamics, such as the effect of density of infectives in a household and whether the transmission characteristics are different for children and adults.

The data used to build the posteriors is taken from a real flu outbreak in France, which involved 334 households monitored over a 15-day period; the aggregates of dates and identities of the patients reporting symptoms in that area was taken to be the observation set.

B) Model Specification

The fundamental equation that encapsulates the model for this problem is as follows:

$$P(Y, v, \psi, \theta) = P(Y|v, \psi)P(v, \psi|\theta)P(\theta)$$

Where,

Y is the set of observations in all households of whether or not an individual reported symptoms for a given day, *i.e.*, it is either 1 or 0 for each individual for each day.

v and ψ are respectively the set of start and end dates of the infectious period for each individual. These dates were defined to be in continuous time so as to provide explicit ordering always.

θ encapsulates a set of model parameters that have to do with risk of infection, susceptibility and infectiousness of an individual. These will be defined in the 'Transmission Level' subsection below, but they basically govern the dynamics of the transmission.

The three terms on the right represent respectively the three levels of their hierarchical model: observation level, transmission level, and prior level. These three levels will be explained in detail in the following subsections. The objective of the MCMC algorithm is to approximate the posterior distribution denoted by: $P(\theta, v, \psi|Y)$.

Observation Level

This level has to do with the first term on the right of the model equation $P(Y|v, \psi)$ which basically ensures that the 'augmented' data v and ψ are consistent with the observations Y. We know from established medical knowledge that the start of the infectious period is 1 to 3 days before the appearance of symptoms. There is no established relationship between the symptoms and the end of the infectious period. So the authors basically defined this term as an indicator function that checked whether v was within the expected limit from the observations Y for each individual on each day.

Transmission Level

This is the second term in the model equation, and it describes the influenza transmission within a household. In a household of population *n*, for a subject *s* susceptible just before time *t*, the instantaneous risk of infection was given by:

$$\lambda_s(t) = \alpha_s + \varepsilon_s \sum_{i\epsilon I(t)} \frac{\beta_i}{n}$$

Where,

α_s is the instantaneous risk of infection from the community

ε_s is the susceptibility to infection for individual *s*

β_i is the ability of infective *i* to infect others

The other parameter of interest when looking at transmission dynamics is the duration of infection for a given individual *i* which is $\psi_i - v_i$. This was assumed to be a gamma distribution with parameters mean μ_i and standard deviation σ_i for the density *d*. Putting it together, the transmission model can be defined as:

$$P(v, \psi|\theta) = \prod_{i\epsilon I} d_{\mu_i,\sigma_i}(\psi_i - v_i) \prod_{i\epsilon I-\{1\}} \lambda_i(v_i)\, e^{-\int_{v1}^{v_i} \lambda_i(t)dt} \prod_{s\epsilon S} e^{-\int_{v1}^{15} \lambda_s(t)dt}$$

Prior Level

This level represents the final term in the model equation and it involves choosing the prior distribution for the unknown parameters $\mu, \sigma, \alpha, \beta$ and ε. The mean duration for an influenza infection was known to be 2-5 days, and so a gamma distribution with mean 3 and standard deviation 2 was chosen as the prior distribution for both μ and σ. Since there is no previously established distribution for the infection risk parameters α and β, the authors chose an exponential distribution with parameter 0.001, which is basically almost like a flat line that reflects the lack of information on the structure of the probability distribution for these. Susceptibility ε_a for adults is defined as 1, and susceptibility for children ε_c is a logistic function such that probability $\frac{\varepsilon_a}{\varepsilon_c} > 1$ is equal to the probability $\frac{\varepsilon_a}{\varepsilon_c} < 1$. Moreover, all these parameters were defined separately for adults and children, but the prior distributions are the same.

C) MCMC Sampling

The stationary distribution of the Markov chain is defined to be the posterior distribution $P(\theta, v, \psi|Y)$. For each individual, the initial values of the parameters were drawn from uniform distributions, with ranges respectively defined to be conservative. For example, the duration was drawn from a uniform distribution with range (0,20) days.

After the initialization, the Metropolis-Hastings sampling algorithm was used. For all of the parameters except v, a new proposal candidate $\rho^*$was generated such that $\log(\rho^*) = \log(\rho) + \delta u$ with u drawn from a normal distribution N(0,1) and δ heuristically defined differently for the different parameters. Since the parameter v was known to have a fixed range, a random walk was not suitable and so it was drawn from the uniform distribution with the range of 1 to 3 days before the date of the first symptom occurrence. Once a new v value was drawn for a given infective, the corresponding ψ was changed such that the duration $\psi_i - v_i$ did not change.

The authors ran the MCMC algorithm for 200,000 iterations with the first 5000 discarded to allow the output to converge. The output was then recorded once every 10 iterations to be a sample for the corresponding posterior distribution. With a Pentium III processor, one such procedure took 30 minutes to execute. The authors used the Gelman-Rubin criterion to determine convergence.

D) Results

Among the many conclusions drawn from the MCMC output, the authors determined that the mean of the infectious period duration was 3.8 days, with a standard deviation of 2 days. They found that, depending on the size of the household, the risk of infection from a fellow household member was 11 to 29 times higher than the risk of infection from the community in general. As expected, they confirmed that the risk of infection was inversely proportional to the duration of infectious period. Finally, they found that children were significantly more susceptible to infection than adults, as well as significantly more infectious than adults despite similar infectious periods. There were no other significant differences between the parameters of adults and children. Some examples of the MCMC outputs and the corresponding sampled posterior distributions are shown in the figure below:

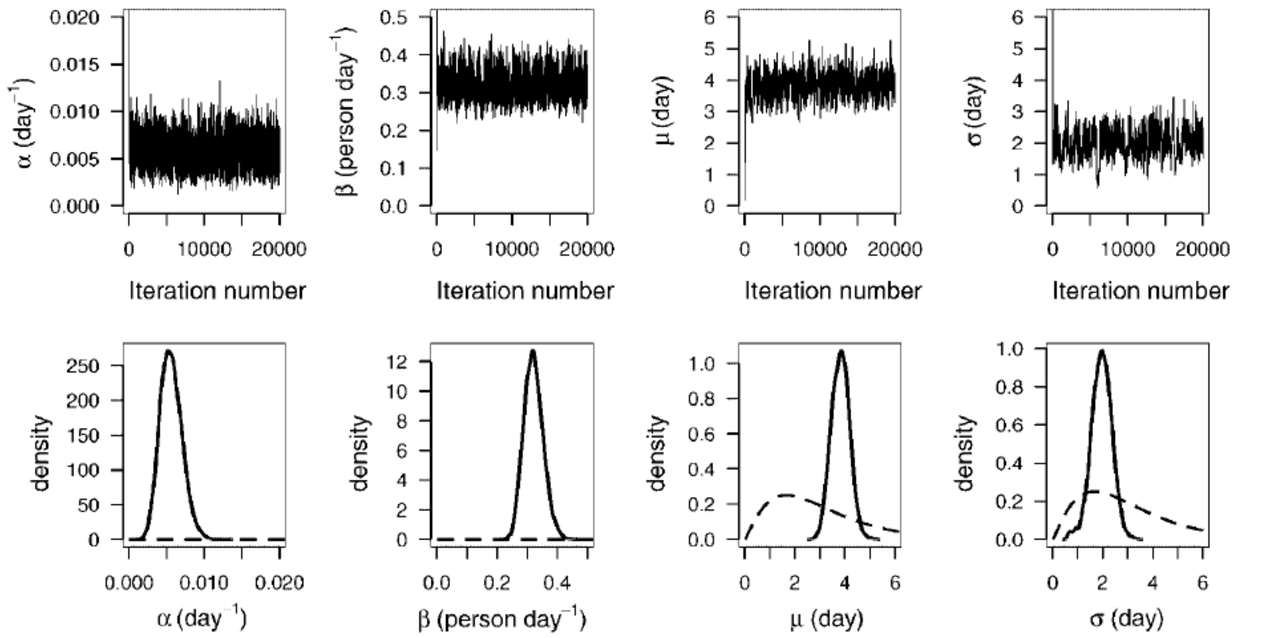


Figure 2. MCMC output, *prior* and *posterior* distributions for the community risk of infection (α), the household risk of infection (β), the mean duration of infectious period (μ) and the standard deviation of infectious period (σ). Dotted lines represent *priors*.

The authors also ran simulations of epidemics in communities of the same size using the newly derived parameters, and confirmed that there was good agreement between the observed number of infections in each household and the expected number.

E) Discussion

In terms of researching MCMC for this report, the main take-away from this paper was how to set up the model equation that would be used in the MCMC algorithm to maximize the posterior by substituting different proposals. The paper showed how when some model parameters can depend on other model parameters (like how v, ψ and θ were interdependent) you could just nest the conditional probabilities within the same framework. The most impressive aspect of this demonstration of MCMC was how even though the priors were somewhat naïve or even completely different from the truth, the algorithm was still able to converge to an accurate representation of the posterior distribution. For example, in the figure above, none of the first three posterior distributions in the bottom row even resemble the prior distribution shown with the dotted line. This shows the power of MCMC and its potential for use in applications wherein the expected shape of the probability distribution is not known or is too complex to define in a closed form.