UNIVERSITY OF MANNHEIM
School of Social Sciences
Department of Political Science

Final Paper for Course
Advanced Quantitative Methods in Political Science

On Using the Metropolis-Hastings Algorithm for Data Imputation

Tobias Stenzel
tobias.stenzel@students.uni-mannheim.de

Prof. Thomas Gschwend, Ph.D.

Submission Date: Mai 02, 2022

Contents

1 Background

1.1 Introduction

The Metropolis-Hastings (MH) algorithm is a method for sampling data points from a probability distribution from which direct sampling is difficult. It places among the top 10 algorithms with the greatest influence on science and engineering in the 20th century (Beichl and Sullivan 2000). The MH algorithm belongs to the class of Markov chain Monte Carlo (MCMC) methods. In my explanation I assume prior knowledge on Monte Carlo sampling. However, I will describe the basics of Markov Chains. This section is structured as follows. First, I motivate the usage of the MH algorithm. Second, I explain the basics of Markov Chains. Third, I derive the algorithm and make clear why it works.

1.2 Motivation

One main application for the MH algorithm is Bayesian inference. Specifically, we want to estimate parameters θ of some probabilistic model f. We have only limited prior knowledge of the distribution of θ, $p(\theta)$, and we have a likelihood sample of f given the unknown θ, namely $p(X|\theta)$. The goal is to estimate the posterior distribution of θ, $p(\theta|X)$, given all information that we have. In practice, we do not have a formal definition of the likelihood but only observations. Therefore, we can only approximate the posterior by numerical integration, i.e., we need to sample many points from the posterior to describe it. We can then use the posterior sample to estimate θ with the maximum a posteriori probability estimate.

Claassen (2019) uses the MH algorithm to impute gaps in a panel data set. His approach consists of four steps: First, assume a data generating process f parameterized by θ. Second, provide the algorithm with the incomplete data X as likelihood and select priors for θ to obtain the posterior distribution $p(\theta|X) = \frac{p(X|\theta)p(\theta)}{P(X)} \propto p(X|\theta)p(\theta)$. Third, use the values for θ with the highest posterior probability as estimates for θ. Finally, insert these estimates into the assumed probabilistic, data generating model f and use it to sample the missing data.

In general and abstracting from Bayesian inference, the MH algorithm generates a sample of random states according to the desired probability distribution $P(X)$. For this purpose, the algorithm employs a Markov process that converges to a unique stationary distribution $\pi(x)$ with $\pi(x) = P(X)$. This distribution can then be used for further steps as previously described. The next section explains the conceptual basics.

1.3 Markov Chains

A Markov chain $(X_t)_{t\in\mathbb{N}}$ is a stochastic process (over time) with the property that the probability of the realization in the next period depends solely on the realization in the current state and not the complete history. This is called the Markov property. Because Markov chains with a countable, or discrete, state space are much more accessible than their continuous variant, in this chapter we will look at the discrete case. Formally, the Markov property writes

$$P(X_{t+1}|X_t, X_{t-1}, ..., X_0) = P(X_{t+1}|X_t). \tag{1}$$

Under some conditions, the stochastic process described by a Markov chain converges to a time-invariant probability distribution, i.e. $P(X_{t+k}|X_{t+k-1}) = P(X_t|X_{t-1}), \forall k > 0$. The crucial step for understanding the MH is to see how it samples a Markov Chain that is certain to converge to a stable posterior distribution. Before exploring how the MH algorithm achieves this result, however, it is necessary to understand its conditions conceptually. To this end, we will use the example depicted by the following graph in Figure 1 that shows the intertemporal transition probabilities between three states representing random events.

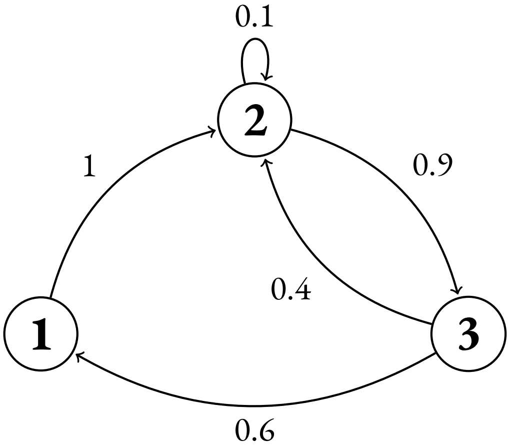

Figure 1: Transition Graph for Markov Chain with 3 states.

This transition graph can be summarized by the $n \times n$ transition matrix T where each element (i, j) represents the probability of moving from state i in period t to state k in period $t + 1$, and where n represents the number of states, i.e $T_{i,j} = P(X_{t+1} = j | X_t = i)$. For our example, we have

$$T = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{pmatrix}. \tag{2}$$

1.3.1 Limit Distribution

As touched upon in the previous subsection, interesting questions can be what the probabilities of each state $j \in \{1, ..., s\}$ are after a finite number or infinitely many steps. For this purpose let $\pi_t(j) = P(X_t = j)$ denote the probability of being in state j in period t. Of course, the probabilities in $t > 0$ depend on the probabilities for the the initial state π_0. We can use the law of total probability to calculate the probability of each state for the next period $t = 1$ by

$$P(X_1 = j) = \sum_{i=1}^{3} P(X_1 = j | X_0 = i)\pi_0(i). \tag{3}$$

I.e., to compute the probability of being in state j in $t = 1$, for each initial state i, we multiply its probability $\pi_0(i)$ by the probability of moving from i to state j.

This is equivalent to $\pi_1 = \pi_0 T$ in vector notation. Further, we can compute the distributions in an arbitrary future period by repeating the matrix multiplication, e.g, $\pi_2 = \pi_0 TT$, or in general, $\pi_t = \pi_0 T^t$.

Now we are ready to define the limit distribution that describes the probability distribution after infinitely many periods by

$$\pi_\infty = lim_{t \to \infty} \pi_t = lim_{t \to \infty} \pi_0 T^t. \tag{4}$$

We can further ask two additional important questions. First, does a limit distribution exist? And second, is it unique, or in other word, do we have the same limit distribution independent from the realization of the initial state X_0? In our example, there does not only exist a limit distribution with $\pi_\infty = (0.2, 0.4, 0.4)$, it is even unique regardless of start distribution π_0. This means that independent of the start state, the probability of each state converges to the same number. For the context of the MH algorithm, this is an important property because we always want to compute the same estimates for our parameters θ, regardless of the starting values of our simulation. In the next section, we introduce and simplify conditions that guarantee a unique limit distribution.

1.3.2 Irreducibility, Periodicity and Stationarity

Definition 1.1. A Markov chain is called *irreducible* if each state is reachable from any other state in a finite number of steps.

Figure 2 shows a Markov chain represented by a bipartite graph. This graph is composed by two times the graph in Figure 1. Obviously, this chain is not irreducible because the initial state impacts all future distributions. More precisely, starting in one subgraph sets the probability of reaching states in the other subgraph to zero. We see that a Markov Chain is only irreducible if there is at least an indirect link between every pair of states. We also observe that if the Markov Chain is not irreducible there can be no limit distribution.

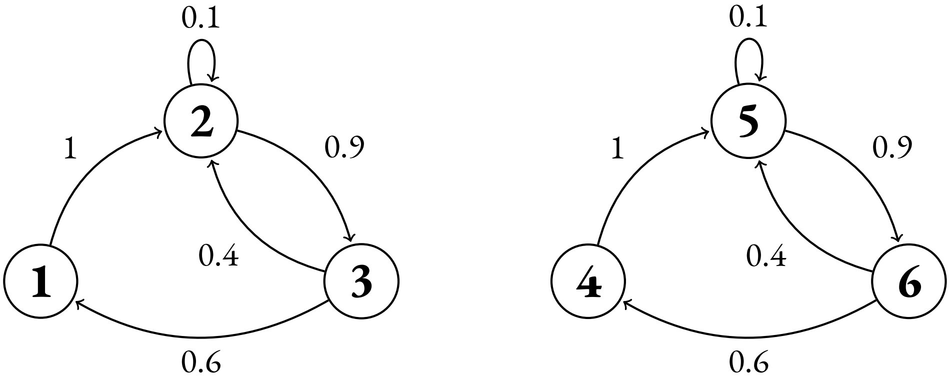

Figure 2: Transition Graph for Irreducible Markov Chain.

Definition 1.2. A state i has a period k if the greatest common denominator of possible revisits is k. A Markov chain is *aperiodic* if the period of all its states is 1.

Consider the five-state Markov chain in Figure 3 as an illustration for the above definition and suppose we start in state 1. Observe that, independent of the random draw for next period, we will arrive again in state 1 after two or four steps. Therefore, state 1 has a period of 2. If a state is revisited in random rather than a fixed time period then the state has period 1. This is automatically the case if a state has a positive edge with itself.

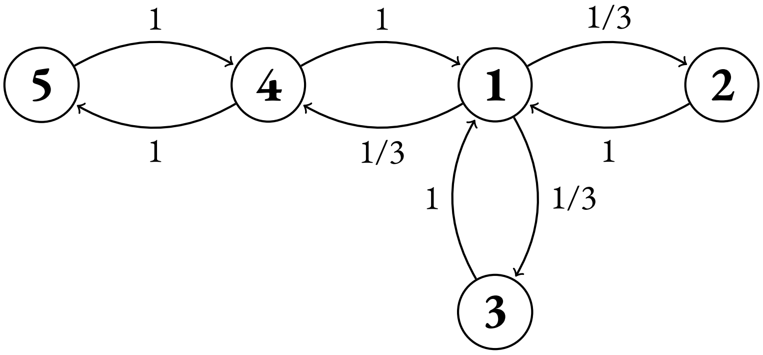

Figure 3: Markov Chain with 2-periodic State 1

Definition 1.3. π^* is the *stationary distribution* of a Markov Chain with Transition matrix T if $\pi^* = \pi^* T$ and π^* is a probability vector.

Verbally, this means that the probability distribution π^* does not change anymore over time. If π^* is also unique, then π^* is our aim, the limit distribution introduces in section 1.3.1, i.e., $\pi^* = \pi_\infty$.

These three definitions are enough to understand the next fundamental theorem.

1.3.3 The Fundamental Theorem of Markov Chains

The next theorem defines formally the condition when a Markov Chain converges to a unique distribution, i.e. the limit distribution.

Theorem 1.1. *(Fundamental Theorem of Markov Chains) If a Markov chain is irreducible and aperiodic (called ergodic) then it has a stationary distribution π^* that is unique (*$\lim_{t \to \infty} P(X_t = i) = \pi_i^*, \forall i$*).*

Therefore, if we want to construct a stable distribution $P(X)$ via Markov chains, we need to ensure that it is irreducible and aperiodic with stationary distribution $\pi^* = P(X)$. In the next subsection, we substitute the stationarity condition by a stronger one before we finally derive the MH algorithm.

1.3.4 Reversibility

Definition 1.4. A Markov chain is *reversible* if there is a probability distribution π over its states such that $\pi(i)T_{ij} = \pi(j)T_{j,i}, \forall i, j$ (reversibility condition).

Theorem 1.2. *A sufficient condition for distribution π^* to be a stationary distribution of a Markov chain with transition matrix T is that it fullfills the reversibility condition.*

Proof. $\sum_i \pi(i)T_{i,j} = \sum_i \pi(j)T_{j,i} = \pi(j)\sum_i T_{j,i} = \pi(j) \implies \pi T = \pi$ □

Reversibility is a stronger condition than stationarity because it requires that the probability flux from i to j is equal to the one from j to i for each possible pair of states. Recall, that stationarity only requires that the probability flux to one state is equal on aggregate and not that it is symmetric between each pair of states over time. Therefore, if we want to achieve a stationary distribution it is enough to ensure that it is reversible.

1.4 The Algorithm

Recall that we want to generate a sample of a desired distribution $P(X)$. For this purpose, we use a Markov process that is uniquely defined by its transition probabilities $P(X_{t+1}|X)$ with limit distribution π so that $\pi = P(X)$. As explained in the previous section, a Markov process has a limit distribution if each transition $X_t \to X_{t+1}$ is reversible and if the stationary distribution π is ergodic. With the MH algorithm, we construct such a Markov process with stationary distribution $\pi = P(X)$. The derivation starts with another way of writing reversibility[1]:

$$P(x'|x)P(x) = P(x|x')P(x') \iff \frac{P(x'|x)}{P(x|x')} = \frac{P(x')}{P(x)} \tag{5}$$

The main idea is to separate transition $P(x'|x)$ in two steps: the proposal step and the acceptance-or-rejection step. Let $g(x')$ be the proposal distribution, i.e., the conditional probability of proposing state x' given x. And let $A(x'|x)$ be the probability of accepting proposed state X'. Formally, we have $P(x'|x) = g(x'|x)A(x'|x')$. Inserting this in Equation (5) gives

$$\frac{P(x')}{P(x)} = \frac{g(x'|x)A(x',x)}{g(x|x')A(x',x)} \iff \frac{A(x',x)}{A(x,x')} = \frac{P(x')}{P(x)}\frac{g(x|x')}{g(x'|x)}. \tag{6}$$

The following choice, termed the Metropolis choice, is commonly used as an acceptance ratio for sampling x' from $P(x')$ that fulfills the above reversibility condition:

$$A(x',x) = \min\left(1, \frac{P(x')}{P(x)}\frac{g(x|x')}{g(x'|x)}\right) \tag{7}$$

Note that the minimizer in $A(x',x)$ enforces that the probability is below 1. The MH algorithm writes as follows:

[1] We simplify our notation by using x' and x instead of X_{t+1} and X_t.

Algorithm 1 Metropolis-Hastings algorithm

```
Initialize $X_0$
for $t \leftarrow 0$ to $T-1$ do
    Draw $u \sim \mathcal{U}_{[0,1]}$
    Draw candidate $X^* \sim P(X^*|X_{t-1})$
    if $u < \min\{1, \frac{p(X^*)g(X_t|X^*)}{p(X_t)g(X^*|X_t)}\}$ then
        $X_{t+1} \leftarrow X^*$
    else
        $X_{t+1} \leftarrow X_t$
    end if
end for
```

Obviously, the construction of the acceptance ratio ensures reversibility. Ergodicity is ensured by the random nature with which we accept proposed states: First, the chain is irreducible because each state is reachable from any other state with positive probability at every single step. Second, for each state x, $P(x' = x)$ is always positive and therefore the Markov chain is aperiodic.

In a general setting, the choice for transition distribution $g(x'|x)$ and the number of iterations until the limit distribution is reached are unclear. These two choices are the hyperparameters of the MH algorithm. In the Bayesian inference application in the article series staring from Claassen (2019), additional choices are the prior distrubtion $p(\theta)$ and the model choice f.

References

Beichl, Isabel, and Francis Sullivan. 2000. "The Metropolis Algorithm." *Computing in Science & Engineering* 2(1): 65–69.

Claassen, Christopher. 2019. "Estimating Smooth Country–Year Panels of Public Opinion." *Political Analysis* 27(1): 1–20.

Statutory Declaration

Hiermit versichere ich, dass diese Arbeit von mir persönlich verfasst ist und dass ich keinerlei fremde Hilfe in Anspruch genommen habe. Ebenso versichere ich, dass diese Arbeit oder Teile daraus weder von mir selbst noch von anderen als Leistungsnachweise andernorts eingereicht wurden. Wörtliche oder sinngemäße Übernahmen aus anderen Schriften und Veröffentlichungen in gedruckter oder elektronischer Form sind gekennzeichnet. Sämtliche Sekundärliteratur und sonstige Quellen sind nachgewiesen und in der Bibliographie aufgeführt. Das Gleiche gilt für graphische Darstellungen und Bilder sowie für alle Internet-Quellen. Ich bin ferner damit einverstanden, dass meine Arbeit zum Zwecke eines Plagiatsabgleichs in elektronischer Form anonymisiert versendet und gespeichert werden kann. Mir ist bekannt, dass von der Korrektur der Arbeit abgesehen werden kann, wenn die Erklärung nicht erteilt wird.

Mannheim, den __________ ____

Name und Unterschrift

I hereby declare that the paper presented is my own work and that I have not called upon the help of a third party. In addition, I affirm that neither I nor anybody else has submitted this paper or parts of it to obtain credits elsewhere before. I have clearly marked and acknowledged all quotations or references that have been taken from the works of other. All secondary literature and other sources are marked and listed in the bibliography. The same applies to all charts, diagrams and illustrations as well as to all Internet sources. Moreover, I consent to my paper being electronically stores and sent anonymously in order to be checked for plagiarism. I am aware that the paper cannot be evaluated and may be graded "failed" ("nicht ausreichend") if the declaration is not made.

Mannheim, __________ ____

Name and Signature