Transition Probabilities in Markov Chains: Frequentist and Bayesian Perspectives

Oscar Eduardo Flores Hernandez

February 1, 2024

Abstract

This study presents a comprehensive comparison between frequentist and Bayesian frameworks in the estimation of transition probabilities within a Markov chain. The Bayesian approach, leveraging simulation and using a Dirichlet prior distribution, excels in adapting to information gaps, particularly in lower density areas. However, it may display a tendency to overestimate probabilities when faced with true probabilities of 0. In contrast, the frequentist approach, while often conservative, tends to underestimate transition probabilities in scenarios with limited data. This comparative analysis provides valuable insights into the dynamics of each framework, contributing to a subtle understanding of their respective strengths and limitations. To illustrate, a small-scale application with a dataset limited to 1000 samples highlights the practical implications of this comparative study.

1 Introduction

In the realm of statistical inference, two prominent frameworks, the frequentist and Bayesian methodologies, stand as robust alternatives, each with its distinctive principles and strengths. This study delves into a comparative analysis of these methodologies, focusing on their application to the estimation of transition probabilities in a Markov chain.

Markov chains provide a mathematical framework for modeling systems that evolve over time, exhibiting a memoryless property that transitions between states based solely on the current state. Estimating the transition

probabilities within a Markov chain is a critical task with applications in various fields, including finance [6], geology [7], operations research [2], among others.

This research will delve into the foundational principles of both methodologies, showcasing their inherent characteristics and strengths. Additionally, we will apply these methodologies to a specific case, emphasizing the estimation of transition probabilities within a Markov chain.

To ensure a comprehensive comparison, we will conduct a simulation study, where we attempt to replicate a stochastic process using both the frequentist and Bayesian frameworks. Through this simulated process, we seek to uncover the advantages and disadvantages of each methodology, shedding light on their respective capabilities and limitations. By synthesizing insights gained from this comparative analysis, we aim to contribute to a understanding of the frequentist and Bayesian methodologies, providing valuable guidance for choosing the most suitable approach for their specific statistical challenges.

2 Markov Chain

The transition probability matrix for a Markov Chain [8] for a discrete time and finite amount of categories is given by:

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1m} \\ P_{21} & P_{22} & \dots & P_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ P_{m1} & P_{m2} & \dots & P_{mm} \end{bmatrix}$$

Where the transition probability $P(X_{n+1} = j | X_n = i)$ represents the probability of transitioning from state i to state j in one time step. The sum of probabilities in each row must equal 1. Thus $\sum_j P_{ij} = 1$ for all i. If we know the transition probabilities, under conditional independence (conditional on the state i), we can simulate future steps using a multinomial distribution.

3 The Frequentist approach

Assuming the transition matrix is irreducible, under ergodicity, it possesses a unique stationary distribution. In the context of ergodicity, the probabilities of occupying each state will eventually stabilize, forming the stationary distribution $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$. Here, π_i signifies the long-term likelihood of being in state i. With ergodicity in play, as time (t) approaches infinity, the probability distribution of the Markov chain at time t, denoted as $P_t = (P_t(1), P_t(2), \ldots, P_t(n))$, gradually converges to π:

$$\lim_{t\to\infty} P_t = \pi$$

In practice, to approximate the stationary distribution, one estimates the probabilities based on the empirical frequencies of states visited [8]. The longer the observation, the closer the empirical distribution to the stationary distribution.

$$P_t(i) \approx \pi_i$$

This approximation becomes accurate as t becomes large.

Considering a sequence of iid random variables $X_1, X_2, \ldots, X_n$ with a common parameter θ. Let $\bar{X}_n$ be the sample mean:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The Law of Large Numbers (LLN) states that as the sample size n goes to infinity, the sample mean $\bar{X}_n$ converges almost surely to the true parameter θ [1] (for the Markov system, θ represents the true probabilities):

$$\lim_{n\to\infty} \bar{X}_n = \theta$$

or:

$$P\left(\lim_{n\to\infty} \bar{X}_n = \theta\right) = 1$$

With probability 1, the average of the observations approaches the true parameter as the sample size becomes very large. The LLN provides a strong convergence result for the sample mean, ensuring that the law of large numbers holds almost surely.

4 The Bayesian approach

Multinomial Distribution

The Multinomial distribution [12] is a generalization of the binomial distribution to multiple categories. It describes the probability of observing a particular set of counts across multiple categories. For a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ with k categories, the probability mass function (PMF) of the Multinomial distribution is given by:

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{p}, n) = \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_k^{x_i} \tag{1}$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$ is the vector of probabilities for each category, and n is the total number of trials.

Dirichlet Distribution

The Dirichlet distribution [11] [4] is a multivariate generalization of the Beta distribution. It is often used as a prior distribution for the parameters of the Multinomial distribution (given the data, we are trying to find the real probabilities $\boldsymbol{p}$ which is a vector containing the probabilities of landing in each category). The probability density function (PDF) of the Dirichlet distribution is given by:

$$f(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1} \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ is the vector of concentration parameters, and Γ is the gamma function [10].

Bayesian Conjugate

A conjugate distribution plays an important role in the Bayesian approach to parametric inference. It provides prior distributions which are tractable. Special attention has been paid for exponential families, which include the

Bernoulli, Wischart, Beta, and Dirichlet distributions [5]. All of them representing important pillars of the Bayesian estimation of parameters. A family of distributions whose density functions have the likelihood's kernel structure is called a natural conjugate [9].

The Dirichlet distribution serves as a natural conjugate for the Multinomial distribution. In practical terms, this means that when we use the Dirichlet distribution as a prior for the parameters ($\boldsymbol{p}$) of the Multinomial distribution, the resulting posterior distribution of these parameters is also a Dirichlet distribution.

This relationship becomes crucial when estimating the parameters of the multinomial, often referred to as 'the probabilities.' Our typical approach involves initially establishing a prior assumption about the parameter distribution, representing the expected distribution of probabilities for each category within the multinomial. Subsequently, as we observe data, we update our beliefs, leading to a final output—a distribution representing our understanding of where the probabilities may lie. It's important to note that this approach doesn't yield a single point estimate for the multinomial probabilities. Instead, it provides a range of combinations, offering insights into where the true probabilities are likely to exist.

In mathematical terms, the posterior distribution is proportional to the product of the likelihood and the prior ($posterior \propto likelihood \cdot prior$):

$$f(\boldsymbol{p}|\text{data}, \boldsymbol{\alpha}) \propto P(\text{data}|\boldsymbol{p}) \cdot f(\boldsymbol{p}|\boldsymbol{\alpha}) \tag{3}$$

$$f(\boldsymbol{p}|\text{data}, \boldsymbol{\alpha}) = \big(\frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_k^{x_i}\big) \cdot \big(\frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}\big) \tag{4}$$

Since with respect to $\boldsymbol{p}$, $\frac{n!}{\prod_{i=1}^{k} x_i!}$ and $\frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)}$ are constants:

$$f(\boldsymbol{p}|\text{data}, \boldsymbol{\alpha}) \propto \prod_{i=1}^{k} p_i^{\alpha_i + x_i - 1} \tag{5}$$

The posterior distribution remains a Dirichlet distribution with updated parameters:

$$f(\boldsymbol{p}|\text{data}, \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k} (\alpha_i + x_i))}{\prod_{i=1}^{k} \Gamma(\alpha_i + x_i)} \prod_{i=1}^{k} p_i^{\alpha_i + x_i - 1} \tag{6}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ is the observed counts in each category.

The MCMC model

In instances where the available data for estimating probabilities is limited, an alternative strategy is required. Here, we delve into the Bayesian approach, a versatile framework that excels in accommodating uncertainty and dynamically updating beliefs with the acquisition of additional data. This methodology proves particularly beneficial when confronted with scenarios involving small sample sizes.

To simulate or analyze a Markov chain, it's essential to determine the transition probabilities between different states. In a Bayesian framework, where exact probabilities are unknown, we adopt a simulation approach. Initially, we simulate possible probabilities that, when observed, align with the outcomes we've seen. To facilitate this, we assume a prior distribution for the probabilities, often opting for the Dirichlet distribution due to its natural conjugate properties. Subsequently, we update our beliefs based on observed data, simulate probabilities from the resulting posterior distribution, and employ these simulated probabilities in a multinomial distribution to generate potential scenarios. This entire procedure is commonly referred to as Markov Chain Monte Carlo (MCMC). [3] :

Algorithm 1: DirichletSample

Data: α
Result: p
$\alpha \leftarrow$ **DirichletParameters** ;
$p \leftarrow$ **SampleFromDirichlet**(α) ;
return p ;

Algorithm 2: MultinomialSample

Data: p
Result: x
$x \leftarrow$ **SampleFromMultinomial**(p) ;
return x ;

Algorithm 3: Generate Posterior Distribution, conditional on state i

Data: N, M, α, x

posterior_distribution $\leftarrow$ [] ;

for $i \leftarrow 1$ *to* N **do**

  $p \leftarrow$ **DirichletSample**$(\alpha + x)$;

  for $j \leftarrow 1$ *to* M **do**

    Append *posterior_distribution* $\leftarrow$ **MultinomialSample**(p) ;

return *posterior_distribution* ;

5 Comparing both models

For this chapter, and showcasing the relevance of the markov estimation, we will assume we have an arbitrary Markov Stochastic Process with 11 categories from which we know the 'real' Transition Probabilities (Provided as an Appendix in the last page). From the real probability matrix we know that the process is irreductible, acyclic, and time homogeneous. These properties guarantee that the process is ergodic and the average of the samples from a large enough sample will converge to the true distribution almost surely [8]. A heatmap of the 'real' stochastic transition probabilities is provided:

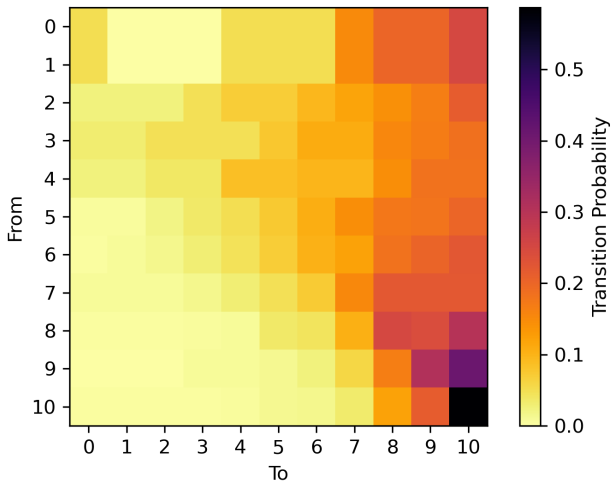
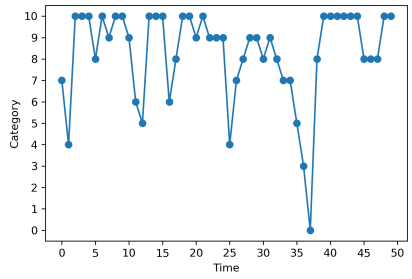
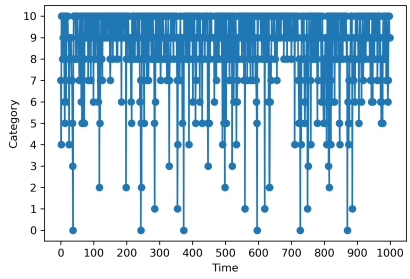


Figure 1: Transition Probability Heatmap

Let's generate a single plausible trajectory from the actual stochastic process, comprising 1000 observations of an individual over time. Subsequently, utilizing this sample trajectory, we aim to reconstruct the genuine probability matrix. This task is particularly complicated due to the need to estimate 121 parameters (one for each transition probability within an 11x11 matrix) with relatively few observations. Given the process's inclination to persist in high-density regions (e.g., 7, 8, 9, 10) and the less frequent transitions to lower categories (e.g., 0, 1, 2, 3), obtaining accurate estimates for the higher density categories is plausible, but challenges may arise in accurately estimating the lower ones. It is also important to notice that the transitions from state 0 to 1, 2, 3 and from state 1 to 1, 2, 3 have a probability of 0, meaning that direct transitions from state 0 to 1, 2, 3 and from state 1 to 1, 2, 3 are not possible according to the real Markov chain model.

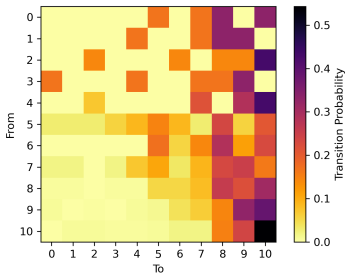


(a) First 50 observed categories for our individual across time.

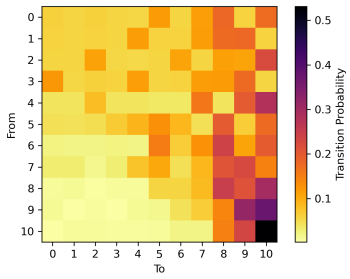


(b) 1000 observed categories for our individual across time.

Approaching this estimation challenge from a frequentist perspective often results in obtaining probability estimates of 0 for certain transition probabilities. This outcome is reasonable since, empirically, those transitions have 'never' occurred in the observed samples. In contrast, the Bayesian approach offers a probability matrix that compensates for the scarcity of information, particularly in areas with lower density.



(a) Frequentist Estimation Heatmap



(b) Bayesian Estimation Heatmap.

Despite the difficulty in achieving a close resemblance to the real probability matrix with only 1000 samples, both frameworks exhibit distinct advantages and drawbacks. The Bayesian approach showcases its strength in compensating for information gaps in lower density areas. However, it has a tendency to overestimate probabilities, particularly in cases where the actual transition probability is 0, as the lack of information may lead the model to simulate non-zero probabilities. Conversely, the frequentist approach leans towards underestimating transition probabilities when data is scarce, often assigning a probability of 0 to processes where the true probability is positive.

6 Conclusion

The empirical application comparing a frequentist and a Bayesian framework has provided valuable insights into their respective approaches to estimating transition probabilities in a Markov chain. The findings suggest that the frequentist approach tends to be more conservative, often underestimating probabilities. This conservatism can be advantageous, particularly when dealing with situations where true probabilities might be very close to zero, as the frequentist approach guards against overestimation.

On the other hand, the Bayesian framework, by its nature, allows for a more flexible modeling of uncertainty. Through the incorporation of prior beliefs and the ability to update these beliefs as more data becomes available, Bayesian methods exhibit a capacity to estimate transition probabilities that may be closer to the true underlying values. This is particularly advantageous when dealing with scenarios where empirical data is scarce, and the Bayesian approach can provide more informative estimates by leveraging prior knowledge.

Furthermore, the Bayesian approach's ability to incorporate noise as a form of compensation for the lack of information is noteworthy. By allowing for the integration of uncertainty through prior distributions and updating based on observed data, Bayesian methods offer a mechanism to quantify and express the inherent variability in transition probabilities. This feature allows Bayesian estimates to be more robust, providing a meaningful estimate even in cases where empirical information is limited or absent.

In summary, the empirical application underscores the nuanced differences between the frequentist and Bayesian frameworks in estimating transition probabilities. While the frequentist approach tends to be more conservative, which can be advantageous in certain contexts, the Bayesian approach excels in handling uncertainty and incorporating prior beliefs, thereby offering a more comprehensive and flexible approach to probability estimation, especially in situations with limited empirical information.

If necessary, you can access the original code utilized for generating the images, estimations, and results through the GitHub-hosted service at the following link: `https://github.com/OscarFlores-IFi/Frequentist_vs_Bayesian`

References

[1] Patrick Billingsley. *Probability and Measure*. Wiley, New York, 1995.

[2] Jose Blanchet, Guillermo Gallego, and Vineet Goyal. A markov chain approximation to choice modeling. *Operations Research*, 64(4):927–944, 2016.

[3] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 2013.

[4] R. D. Gupta and Donald St. P. Richards. The history of the dirichlet and liouville distributions. *International Statistical Review / Revue Internationale de Statistique*, 69(3):433–446, 2001.

[5] E. Gutiérrez-Peña, A.F.M. Smith, J.M. Bernardo, et al. Exponential and bayesian conjugate families: Review and extensions. *Test*, 6:1–90, 1997.

[6] Mujtaba Malik and Lyn C. Thomas. Modelling credit risk of portfolio of consumer loans. *The Journal of the Operational Research Society*, 61(3):411–420, 2010.

[7] Takashi Nishioka and Haresh C. Shah. Application of the markov chain on probability of earthquake occurrence. *Journal of the Japan Society of Civil Engineers (Doboku Gakkai Ronbunshu)*, 298:137–145, 1980.

[8] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[9] Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Harvard University Press, Boston, 1961.

[10] Reinhold Remmert. *The Gamma Function*, pages 33–72. Springer New York, New York, NY, 1998.

[11] George A. F. Seber. Continuous distributions. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 922–924. Springer, Berlin, Heidelberg, 2011.

[12] George A. F. Seber. Multinomial distribution. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 882–884. Springer, Berlin, Heidelberg, 2011.

7 Appendix

From\To	0	1	2	3	4	5	6	7	8	9	10
0	0.050	0.000	0.000	0.000	0.050	0.050	0.050	0.150	0.200	0.200	0.250
1	0.050	0.000	0.000	0.000	0.050	0.050	0.050	0.150	0.200	0.200	0.250
2	0.024	0.024	0.024	0.048	0.071	0.071	0.095	0.119	0.143	0.167	0.214
3	0.031	0.031	0.046	0.046	0.046	0.077	0.108	0.108	0.154	0.169	0.184
4	0.024	0.024	0.037	0.037	0.085	0.085	0.098	0.098	0.146	0.183	0.183
5	0.005	0.005	0.02	0.035	0.05	0.075	0.106	0.146	0.176	0.181	0.201
6	0.004	0.008	0.016	0.028	0.045	0.069	0.101	0.121	0.182	0.202	0.224
7	0.007	0.007	0.007	0.015	0.029	0.046	0.073	0.153	0.221	0.221	0.221
8	0.003	0.003	0.004	0.006	0.009	0.035	0.043	0.104	0.25	0.242	0.301
9	0.002	0.002	0.002	0.007	0.008	0.013	0.023	0.058	0.166	0.309	0.41
10	0.003	0.003	0.003	0.004	0.005	0.012	0.014	0.034	0.121	0.214	0.587

Table 1: Transition Probabilities