Testing for Markovian Property: A Statistical Approach to Transition Matrix Homogeneity

Ryan Rodrigue
Santa Fe Christian Schools, Solana Beach, CA

Abstract

Markov chains are powerful tools for modeling systems that transition between distinct states with the probability of transitioning to each state depending solely on the previous one. This Markov property is central to applications across disciplines, including geology, biology, and economics, as it allows for simplified, tractable analysis. In this paper, we assess the presence of Markovian behavior in several processes by estimating a one-step transition probability matrix and employing a chi-squared test to validate the consistency of transition probabilities over time. Using R software, our study also examines change-point detection to identify shifts in transition matrices, offering insights into cases where Markovian assumptions may or may not hold, such as weather patterns, language structures, and stock price movements.

Keywords: Markov chain, Markovian property, chi-squared test, testing Markovian property

1 Introduction

1.1 Background

Markov chains are mathematical models that describe systems undergoing transitions between a finite or countably infinite set of states $S = \{s_1, s_2, \dots\}$. The system evolves in discrete time steps $n = 0, 1, 2, \dots$, and the probability of moving to the next state depends solely on the current state, not the sequence of previous states. This is formalized by the Markov (or Markovian) property, which states that:

$$P(X_{n+1} = s_{n+1} \mid X_n = s_n, X_{n-1} = s_{n-1}, \dots, X_0 = s_0) = P(X_{n+1} = s_{n+1} \mid X_n = s_n).$$

The transition probabilities between states are summarized in a one-step transition probability matrix $\mathbf{P}$, where ijth entry $p_{ij} = P(X_{n+1} = j \mid X_n = i)$ represents the probability that the chain transitions from state i to state j in one step. Note that these probabilities are constants independent of n, and the probabilities in each row must sum up to one as the chain must transition somewhere with probability one. Markov chains are widely applicable due to their mathematical tractability and "memoryless" property.

In this paper, we examine several processes and employ statistical methods to evaluate whether they exhibit Markovian behavior. Demonstrating that a process adheres to the Markov property rigorously is a formidable challenge, as it requires establishing that the property holds across sequences of states of arbitrary lengths, such as triples, quadruples, and so on. Consequently, rather than attempting to prove the Markov property directly, we take a more practical approach. We estimate the one-step transition probability matrix from a subset of the data and then apply a chi-squared test to assess whether the remaining data are consistent with the estimated transition matrix. In essence, we proceed under the assumption that the process is Markovian, focusing specifically on testing the homogeneity of the transition matrix. In cases where the matrix is not stationary, we also investigate potential change points, i.e., points at which the transition matrix undergoes structural shifts.

1.2 Literature Review

Testing the homogeneity of Markov chains has widespread applications across various fields. In geology, Markov chains are often used to model the stratigraphy of sedimentary rocks and sequence transitions between geological formations [1,2,3]. Markov chain models in biology are frequently applied to sequence analysis, such as DNA or protein sequences, gene expression, and population dynamics [4,5,6]. In engineering, especially in fields like reliability and maintenance, Markov chains model system states over time. Homogeneity testing can reveal if a system's failure process remains consistent or if there are periods of increased or decreased failure rates [7,8]. Markov chains in computer science are commonly found in algorithmic analysis, natural language processing, and network traffic analysis [9,10]. Markov chains are applied in medical studies to model disease progression and patient states over time [11]. In fields like economics and social sciences, Markov chains model consumer behavior, market states, or social processes [12,13,14].

2 Theoretical Framework

2.1 Testing for Homogeneity of Markov Chain

To test for homogeneity in a Markov chain, we assess whether the transition probabilities stay constant over time. The process begins by estimating the one-step transition probability matrix using the first 10–20% of the data. We then examine if the remaining data points align with a Markov chain model that follows these estimated transition probabilities, using a chi-squared test. The details are provided below.

Consider a finite-state Markov chain with k states, say. Suppose we have a sequence of n observations and would like to test whether the transitions between states are governed by a given one-step transition probability matrix $(p_{ij}),\ i,j = 1,\ldots,k$. For a given initial state i, denote by $n_{i1}, n_{i2}, \ldots n_{ik}$ the number of observed transitions to states $1, 2, \ldots, k$, respectively. Assuming that $n_{i1}+n_{i2}+\cdots+n_{ik} = n_i$, the expected number of transitions is $n_i p_{i1},\ n_i p_{i2}, \ldots,\ n_i p_{ik}$. Under the null hypothesis that the transition probability matrix is the true one, the test statistic $\chi^2 = \sum_{ij} \dfrac{(n_{ij} - n_i p_{ij})^2}{n_i p_{ij}}$ follows a chi-squared distribution with $n(n-1)$ degrees of freedom. To prove this statement rigorously, we note first that the observed transition counts $n_{i1}, n_{i2}, \ldots, n_{ik}$ can be considered as a sample of size n_i from a multinomial distribution with the probability mass function (pmf)

$$p(n_{i1}, n_{i2}, \ldots, n_{ik}) \;=\; \frac{n_i!}{n_{i1}!n_{i2}!\ldots n_{ik}!} p_{i1}^{n_{i1}}\, p_{i2}^{n_{i2}}\, \cdots\, p_{ik}^{n_{ik}}.$$

The natural logarithm of the pmf can be written as

$$\ln p(n_{i1}, n_{i2}, \ldots, n_{ik}) \;=\; \ln n_i! \,-\, \sum_{j=1}^{k} \ln n_{ij}! \,+\, \sum_{j=1}^{k} n_{ij}\, \ln p_{ij}.$$

Now using Stirling's formula that states that for sufficiently large n, $n! \approx \sqrt{2\pi n}\, n^n\, e^{-n}$, or, equivalently, $\ln n! \approx \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln n + n\ln n - n$, we write

$$\begin{aligned}
\ln p(n_{i1}, n_{i2}, \ldots, n_{ik}) \;&=\; \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln n_i + n_i \ln n_i - n_i \\
-\sum_{j=1}^{k}\left(\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln n_{ij} + n_{ij}\ln n_{ij} - n_{ij}\right) &+ \sum_{j=1}^{k} n_{ij}\ln p_{ij} \\
&= \text{constant} - \frac{1}{2}\sum_{j=1}^{k}\ln n_{ij} - \sum_{j=1}^{k} n_{ij}\ln n_{ij} + \sum_{j=1}^{k} n_{ij}\ln p_{ij} \qquad (1)
\end{aligned}$$

where the constant term depends only on n_i and not n_{ij}'s. In the last equality, we used the fact that $\sum_{j=1}^k n_{ij} = n_i$. Next, we apply the transformation $Z_{ij} = \dfrac{n_{ij} - n_i\, p_{ij}}{\sqrt{n_i\, p_{ij}}}$, from where $n_{ij} = n_i\, p_{ij}\left(1 + \dfrac{Z_{ij}}{\sqrt{n_i\, p_{ij}}}\right)$. Using the fact that by Taylor's expansion, for all small x, $\ln(1+x) \approx x + x^2/2$, we obtain

$$\ln n_{ij} = \ln(n_i\, p_{ij}) + \ln\left(1 + \frac{Z_{ij}}{\sqrt{n_i\, p_{ij}}}\right) \approx \ln(n_i\, p_{ij}) + \frac{Z_{ij}}{\sqrt{n_i\, p_{ij}}} + \frac{Z_{ij}^2}{2n_i\, p_{ij}}.$$

Plugging this expression into (1), we get

$$\ln p(n_{i1}, n_{i2}, \ldots, n_{ik}) \approx \text{constant} - \frac{1}{2}\sum_{j=1}^{k}\left(\ln(n_i\, p_{ij}) + \frac{Z_{ij}}{\sqrt{n_i\, p_{ij}}} + \frac{Z_{ij}^2}{2n_i\, p_{ij}}\right)$$

$$- \sum_{j=1}^{k} n_i\, p_{ij}\left(1 + \frac{Z_{ij}}{\sqrt{n_i\, p_{ij}}}\right)\left(\ln(n_i\, p_{ij}) + \frac{Z_{ij}}{\sqrt{n_i\, p_{ij}}} + \frac{Z_{ij}^2}{2n_i\, p_{ij}}\right) + \sum_{j=1}^{k} n_i\, p_{ij}\left(1 + \frac{Z_{ij}}{\sqrt{n_i\, p_{ij}}}\right)\ln p_{ij},$$

which after some simplification and dropping of higher-order terms becomes

$$\ln p(n_{i1}, n_{i2}, \ldots, n_{ik}) \approx \text{constant} - \sum_{j=1}^{k} \frac{Z_{ij}^2}{2}.$$

Thus, the pmf can be approximately written as

$$p(n_{i1}, n_{i2}, \ldots, n_{ik}) \approx C\,\exp\left\{-\sum_{j=1}^{k}\frac{Z_{ij}^2}{2}\right\} = C\,\exp\left\{-\frac{\chi_i^2}{2}\right\}$$

where C denotes a constant, and $\chi_i^2 = \sum_{j=1}^{k} Z_{ij}^2 = \sum_{j=1}^{k} \dfrac{(n_{ij} - n_i\, p_{ij})^2}{n_i\, p_{ij}}$ is a random variable that has a chi-squared distribution with $k-1$ degrees of freedom, the k for the $n_{ij}, j = 1, \ldots, k$, variables minus one for the linear relation $\sum_{j=1}^k n_{ij} = n_i$. Testing for the initial state i is asymptotically independent of i, hence, the overall test statistics is

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{(n_{ij} - n_i\, p_{ij})^2}{n_i\, p_{ij}},$$

which under the null hypothesis, has a chi-squared distribution with $k(k-1)$ degrees of freedom [15].

It is assumed that each $p_{ij} \neq 0$ since division by zero is undefined. Therefore, for each $p_{ij} = 0$, we must reduce the degrees of freedom by one, as each $p_{ij} = 0$ introduces an additional linear constraint.

2.2 Change-point Detection

Change-point detection in Markov chains focuses on identifying the point(s) at which the one-step transition probability matrix changes. Consider a simple situation where a sequence of n observations is available and it is known that the transition matrix changed exactly once somewhere in the middle half of the observations. To estimate the precise point of change, we can use the first 20%, say, to estimate the initial transition probability matrix, denoted as $(\hat{p}_{ij}),\ i, j = 1, \ldots, k$, and the last 20%, say, to estimate the final transition probability matrix, denoted $(\hat{p}^*_{ij}),\ i, j = 1, \ldots, k$. The remaining observations between indices $m_1 = 0.2n+1$ and $m_2 = 0.8n-1$ are assumed to contain the unknown change point m within this interval. Defining n_{ij} as the transition counts of observations between m_1 and m, and n^*_{ij} as those between m and

m_2, we construct the likelihood function:

$$L(n_{11},\ldots,n_{kk},n_{11}^*,\ldots,n_{kk}^*) = \prod_{i=1}^{k}\left[\frac{n_i!}{n_{i1}!\cdots n_{ik}!}\,\hat{p}_{i1}^{n_{i1}}\cdots\hat{p}_{ik}^{n_{ik}}\,\frac{n_i^*!}{n_{i1}^*!\cdots n_{ik}^*!}\,\hat{p}_{i1}^{*\,n_{i1}^*}\cdots\hat{p}_{ik}^{*\,n_{ik}^*}\right]$$

where $n_i = n_{i1}+\cdots+n_{ik}$ and $n_i^* = n_{i1}^*+\cdots+n_{ik}^*$. The value of m for which this likelihood function is the largest is the maximum likelihood estimator of the change point.

3 Applications

Knowing that a process follows the Markov property lets us focus on the present state to make decisions, without any prior context of the system, which allows for efficient modeling of complex systems. In this section, we will explore specific cases where Markov models both are and are not applicable. From predicting stock market behavior to determining the language of a text, we show the applicability of Markovian models.

3.1 Testing Markov Property in Weather Change Patterns

The first example we'll explore in testing for the Markovian property involves weather patterns. A simple Google search reveals that weather is often cited as a common example of a Markov process in the real world. However, in this section, we challenge this notion and conclude that weather does not actually follow a Markov process.

We analyze one year of weather data (May 1, 2023 to April 30, 2024) from wunderground.com about Long Beach, California, categorizing each day's weather into one of four states: A (sunny), B (cloudy), C (foggy), or D (raining/snowing). We estimate the transition probability matrix using the first 20% of the data, and then we use the remaining 80% to calculate the observed frequencies, along with the expected frequencies based on the assumption that the transition probability matrix remains constant. We obtain the following observed and expected frequencies:

$$\text{Observed} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array}\!\!\begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} 138 & 29 & 10 & 6 \\ 33 & 26 & 2 & 4 \\ 6 & 7 & 10 & 1 \\ 6 & 3 & 2 & 6 \end{pmatrix} \end{array}$$

and

$$\text{Expected} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array}\!\!\begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{pmatrix} 119.654 & 56.308 & 7.038 & 0 \\ 12.717 & 52.283 & 0 & 0 \\ 24 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

We conduct a chi-squared test comparing the corresponding non-zero entries of the observed and expected matrices, with degrees of freedom calculated as $(4)(4-1)-$ the number of zeros in the expected matrix$=12-10=2$. This yields a chi-squared test statistic of $\chi^2 = 76.364$ and a corresponding p-value of $2.616\cdot 10^{-17}$, which is clearly outside the acceptable range. Therefore, the two matrices differ significantly in their corresponding elements, indicating that the process is not Markovian, contradicting the common belief that weather follows a Markov process.

We suspected that Long Beach, California—a city that doesn't experience snow—might yield different results compared to a city with snowy days. To explore this, we conducted the same test with one year (July 19, 2023 to July 18, 2024) of data from wunderground.com about Detroit, MI, and found quite similar results ($\chi^2 = 19.792$, df=5, and p-value=0.001). Thus, we can reasonably conclude that weather is not a Markov process. These findings suggest that weather forecasters might benefit from taking into account multiple previous days of weather data when making predictions for the next day.

3.2 Testing Markov Property in Consonant-Vowel Transitions in Literary Works

The next example of testing for the Markov property involves transitions between consonants and vowels in literature. We consider two states: C for consonants and V for vowels. To determine if the sequence of consonants and vowels in English literature follows a Markov process, we choose Ray Bradbury's book *Fahrenheit 451*. After converting the text to a plain format and substituting every consonant with a C and every vowel with a V, we analyze the first 20% of the text to estimate the one-step transition probability matrix. The remaining 80% of the text gives us the observed frequencies of transitions along with the expected ones, assuming that the transition matrix persists. The observed and the expected matrices are presented below:

$$\text{Observed} = \begin{array}{c} \\ C \\ V \end{array} \begin{array}{c} \begin{array}{cc} C & \quad V \end{array} \\ \begin{pmatrix} 8330 & 49221 \\ 49219 & 47091 \end{pmatrix} \end{array}$$

and

$$\text{Expected} = \begin{array}{c} \\ C \\ V \end{array} \begin{array}{c} \begin{array}{cc} C & \qquad V \end{array} \\ \begin{pmatrix} 8443.422 & 49103.590 \\ 49259.329 & 47048.67 \end{pmatrix} \end{array}.$$

We conduct a chi-squared test comparing the corresponding entries of the observed and expected matrices, with degrees of freedom calculated as $(2)(2-1) -$ the number of zeros in the expected matrix $= 2 - 0 = 2$. This yields a chi-squared test statistic of $\chi^2 = 1.875$ and a p-value of 0.392, which is above the 0.05 threshold for rejecting the null hypothesis. Therefore, we conclude that the sequence of consonants and vowels in English literature follows a Markovian process.

Having established that this process is Markovian, we can now compare the transition probability matrices of various languages and utilize these matrices to identify the language of an unknown text. For instance, we analyzed texts from four different languages to derive approximate transition probability matrices for English, Spanish, French, and Russian. The works we selected include *A Tale of Two Cities* by Charles Dickens for English, *Cien Años de Soledad* by Gabriel García Márquez for Spanish, *Le Petit Prince* by Antoine de Saint-Exupéry for French, and *Eugene Onegin* by Alexander Pushkin for Russian. The estimated one-step transition probability matrices are given below.

$$\text{English} = \begin{array}{c} \\ C \\ V \end{array} \begin{array}{c} \begin{array}{cc} C & \quad V \end{array} \\ \begin{pmatrix} 0.1464 & 0.8536 \\ 0.5206 & 0.4794 \end{pmatrix} \end{array}$$

$$\text{Spanish} = \begin{array}{c} \\ C \\ V \end{array} \begin{array}{c} \begin{array}{cc} C & \quad V \end{array} \\ \begin{pmatrix} 0.1544 & 0.8456 \\ 0.7027 & 0.2973 \end{pmatrix} \end{array}$$

$$\text{French} = \begin{array}{c} \\ C \\ V \end{array} \begin{array}{c} \begin{array}{cc} C & V \end{array} \\ \begin{pmatrix} 0.2284 & 0.7716 \\ 0.6069 & 0.3931 \end{pmatrix} \end{array}$$

$$\text{Russian} = \begin{array}{c} \\ C \\ V \end{array} \begin{array}{c} \begin{array}{cc} C & V \end{array} \\ \begin{pmatrix} 0.1370 & 0.8630 \\ 0.6681 & 0.3319 \end{pmatrix} \end{array}$$

This process can be expanded by including a larger sample of languages and multiple texts for each language, resulting in a more accurate and comprehensive "dictionary" that pairs each language with its corresponding transition probability matrix. When we encounter a text in an unknown language, we can calculate its transition probabilities and compare them to the transition probability matrices of the known languages to identify the language of the text. For instance, we analyze the book *Moby-Dick* by Herman Melville, which has the following determined transition probabilities:

$$\text{“Moby-Dick”} = \begin{array}{c} \\ C \\ V \end{array} \begin{array}{c} \begin{array}{cc} C & V \end{array} \\ \begin{pmatrix} 0.1422 & 0.8578 \\ 0.5132 & 0.4868 \end{pmatrix} \end{array}$$

The transition probabilities from Moby-Dick closely align with those of the English matrix, allowing us to confidently conclude that it is an English text.

As a side note, this methodology could hold significant value for the military, which places a high priority on accurately identifying the language of intercepted communications or captured foreign documents.

3.3 Testing Markov Property in Daily Stock Price Changes

Our next example of testing for the Markov property focuses on Tesla's daily closing prices (from August 3, 2023 to August 2, 2024) found on yahoofinance.com. While this method can apply to any stock, we use Tesla for illustration. We categorize daily price changes into three states: U (up), D (down), and C (constant). An "up" day is one where the closing price increases by more than 1.5% from the previous day, while a "down" day sees a decrease of more than 1.5%. "Constant" days are those that do not meet the criteria for up or down.

Following the approach used in previous examples, we utilize the initial 20% of the data to estimate the one-step transition probability matrix. The estimated probabilities are then applied to the remaining 80% of the dataset to calculate the expected frequencies of state transitions. Below, we present both the observed and expected transition probability matrices:

$$\text{Observed} = \begin{array}{c} \\ U \\ D \\ C \end{array} \begin{array}{c} \begin{array}{ccc} U & D & C \end{array} \\ \begin{pmatrix} 13 & 18 & 17 \\ 15 & 19 & 24 \\ 19 & 22 & 31 \end{pmatrix} \end{array}$$

and

$$\text{Expected} = \begin{array}{c} \\ U \\ D \\ C \end{array} \begin{array}{c} \begin{array}{ccc} U & D & C \end{array} \\ \begin{pmatrix} 12.158 & 12.632 & 20.211 \\ 15.130 & 20.174 & 22.696 \\ 16.258 & 23.226 & 32.516 \end{pmatrix} \end{array}.$$

To assess the distance between the observed and expected matrices, we perform a chi-squared test, comparing their corresponding entries. The degrees of freedom are calculated as $(3)(3-1) -$ the number of zeros in the expected matrix$= 6 - 0 = 6$. This test yields a chi-squared test statistic of $\chi^2 = 3.841$ and a p-value of 0.698. Since this p-value is greater than the 0.05 threshold, it lies within the acceptable range, indicating that the sequence of up, constant, and down days in Tesla's stock exhibits the Markovian property. This method can be similarly applied to other stocks to determine if the daily price changes follow a Markov process, offering traders insights into next-day performance based on the previous day's trend.

3.4 Testing Markov Property in DNA Nucleotide Sequences

The final application of testing for the Markovian property that we examine involves nucleotide sequences within human DNA. Specifically, we use a sample of GRCh38, a detailed human reference genome developed by the Genome Reference Consortium, which we found on kaggle.com. GRCh38 serves as an internationally accepted DNA sequence, widely utilized in genetic research. In this analysis, we define four distinct states corresponding to the four nucleotides that constitute DNA: A (adenine), T (thymine), G (guanine), and C (cytosine).

Once again, we use the first 20% of the DNA data points to estimate the one-step transition probabilities. We then calculate the observed and expected matrices of transition frequencies using the remaining 80% of the data. The resulting matrices are shown below.

$$\text{Observed} = \begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} 122 & 116 & 115 & 89 \\ 55 & 122 & 152 & 118 \\ 112 & 97 & 114 & 94 \\ 147 & 122 & 30 & 97 \end{pmatrix} \end{array}$$

and

$$\text{Expected} = \begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} 118.589 & 96.627 & 101.020 & 118.588 \\ 42.881 & 132.542 & 152.034 & 124.746 \\ 105.000 & 109.375 & 96.250 & 109.375 \\ 145.167 & 126.556 & 33.500 & 85.611 \end{pmatrix} \end{array}.$$

We conduct a chi-squared test with degrees of freedom calculated as $(4)(4-1) -$ the number of zeros in the expected matrix$= 12 - 0 = 12$. This test yields a chi-squared statistic of $\chi^2 = 27.297$ and a p-value of 0.007. Since this p-value is less than 0.05, it indicates that the nucleotide sequence does not satisfy the Markov property.

4 Applications of Change-point Detection

To evaluate the effectiveness of our change-point detection method, we can simulate the trajectories of two sequences composed of states 1, 2, and 3, each with distinct transition probabilities. After generating these sequences, we will combine them and assess whether we can accurately identify the point at which the change in transition probabilities occurs.

We choose the following two transition probability matrices:

$$
\text{tm1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.7 & 0.1 & 0.2 \\ 0.1 & 0.5 & 0.4 \\ 0.5 & 0.2 & 0.3 \end{pmatrix} \end{array}
$$

and

$$
\text{tm2} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.2 & 0.3 & 0.5 \\ 0.3 & 0.6 & 0.1 \\ 0.2 & 0.1 & 0.7 \end{pmatrix} \end{array}.
$$

We simulate a sequence of 200 states using the transition probability matrix `tm1` and create another sequence of the same length based on the transition probability matrix `tm2`. After appending the two sequences, we apply the theory outlined in Subsection 2.2 to estimate the point at which the change occurs. We begin by using the first 20% and the last 20% of the data to approximate the one-step transition probabilities before and after the change. Next, we analyze the middle 60% of the data to generate a set of values for the log-likelihood function by sliding the change-point value and identifying the point at which the log-likelihood function reaches its global maximum. In this instance, we found the value of 192, which is quite close to the actual change-point of 201. The graph of the log-likelihood function is displayed for the middle 240 observations, representing 60% of the total 400 observations.

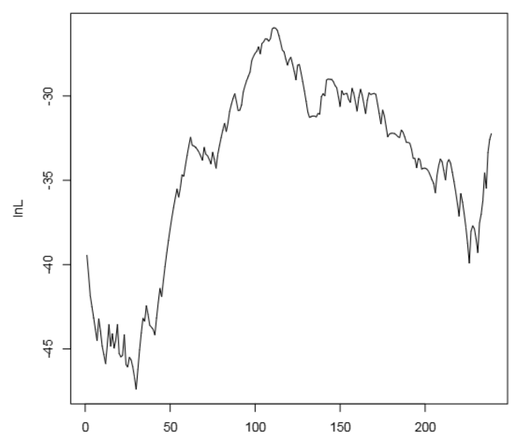

Figure 1: Plot of Log-likelihood Function for the Middle Observations

Change-point detection has numerous practical applications. For instance, when examining Detroit's weather patterns, we can anticipate a change-point in transition probabilities occurring at the start of the snowy season, as this marks a significant shift in weather conditions. In the 2023-2024 winter season, the first snowfall in Detroit was recorded on November 26, 2023 according to wunderground.com. Therefore, when analyzing the weather data from wunderground.com between July 19, 2023, to July 18, 2024, we expect the change point to be around November 27, 2023. Upon conducting the analysis using the log-likelihood function, we identified the change point on December 15, 2023, which, while slightly later than our predicted date, is still reasonably close. This observed transition between distinct seasonal patterns illustrates why our initial tests indicated that weather cannot be modeled as a Markov process: the transition probabilities vary significantly with the changing seasons.

5 Future Work

To enhance the chi-squared test approach for testing homogeneity in Markov chains, future research could incorporate the likelihood ratio test for sparse data scenarios, Bayesian methods for robustness with prior knowledge, a non-parametric Kolmogorov-Smirnov test to avoid strict assumptions, cumulative sum and sequential tests for real-time monitoring, and neural networks to capture complex transitions. These tests offer opportunities to explore interesting scientific applications. Additionally, this Markov testing could be extended to higher-order Markov processes to cover a wider range of applications.

6 Supplemental Materials

To support and extend the findings in this paper, all relevant R codes and datasets are available in a public GitHub repository. This repository, created under the username ryrod724 and titled Testing-for-Markovian-Property-A-Statistical-Approach-to-Transition-Matrix-Homogeneity, provides readers with access to the full suite of resources used in this study. By exploring the code, users can better understand the methodologies employed, including how the estimation of the transition probability matrix, chi-squared tests, and change-point detection analyses were implemented.

Acknowledgments

I would like to express my deepest gratitude to those who supported and guided me throughout this project. To my friends and family, thank you for your unwavering belief in me and constant encouragement. A special thanks to Mr. Maxon, whose passion for the subject has been a source of inspiration and whose guidance prepared me for this work. I am also immensely grateful to my mentor Dr. Olga Korosteleva, a professor of Statistics at California State University, Long Beach, for her invaluable support, insight, and dedication; without her expertise and commitment, this project would not have been possible.

References

[1] Schwarzacher, W. (1969). The use of Markov chains in the study of sedimentary cycles. *Mathematical Geology*, 1: 17-39.

[2] Gingerich, P.D. (1969). Markov analysis of cyclic alluvial sediments. *Journal of Sedimentary Research*, 39 (1): 330–332.

[3] Riegl, B.M. and S.J. Purkis. (2009). Markov models for linking environments and facies in space and time (recent Arabian Gulf, Miocene Paratethys). *International Association of Sedimentologists Specis Publication*, 41: 337–360.

[4] Sharma, A. and N. Adlakha. (2014). Markov chain model to study the gene expression. *Advances in Applied Science Research*, 5(2):387-393.

[5] Zakarczemny, M. and M. Zajecka. (2022). Note on DNA Analysis and Redesigning Using Markov Chain. *Genes*, 13(3): 554.

[6] Djordjevic, I.B. (2015). Markov Chain-Like Quantum Biological Modeling of Mutations, Aging, and Evolution. *Life*, 5(3): 1518–1538.

[7] Pérez-Lechuga, G., Venegas-Martínez, F., and J.F. Martínez-Sánchez. (2021). Mathematical Modeling of Manufacturing Lines with Distribution by Process: A Markov Chain Approach. *Mathematics*, 9, 3269.

[8] Winkovich, T. and D. Eckardt. (2005). Reliability analysis of safety systems using Markov-chain modelling. *2005 European Conference on Power Electronics and Applications*, 1-10.

[9] Almutiri, T. and F. Nadeem. (2022). Markov Models Applications in Natural Language Processing: A Survey. *International Journal of Information Technology and Computer Science*, 2, 1-16.

[10] More, S. and D.r Shuklain. (2019). Review on Internet Traffic Sharing Using Markov Chain Model in Computer Network, in D. K. Mishra et al. (eds.), Data Science and Big Data Analytics, *Lecture Notes on Data Engineering and Communications Technologies 16*, Springer.

[11] Thokala, P., Dodd, P., Baalbaki, H., Brennan, A., Dixon, S. and K. Lowrie. (2020). Developing Markov Models From Real-World Data: A Case Study of Heart Failure Modeling Using Administrative Data. *Value in Health*, 23 (6): 743-750.

[12] Waller, E.N.K., Adablah, P.D., and Q.-A. Kester. (2019). Markov Chain: Forecasting Economic Variables, in *2019 International Conference on Computing, Computational Modelling and Applications.*

[13] Efor, T.E. and J. Chukwu. (2020). Application of Markov Analysis to Consumers Preference. *International Journal of Mathematics and Statistics Invention*, 8 (6): 14-23.

[14] Dennis, L.A., Fu, Y., and M. Slavkovik. (2022). Markov chain model representation of information diffusion in social networks. *Journal of Logic and Computation*, 32 (6): 1195–1211.

[15] Frankel, E.G. (1984). *Systems Reliability and Risk Analysis*, Martinus Nijhoff Publishers.