Project 1: Markov Chains Draft

Hanzhang Yin

October 21, 2024

1 Summary

The Project focusing on using Markov models to analyze the tonal patterns in classical Chinese poetry, paticularly examining whether sequences of tones align with the strict metrical rules of (lv shī). These rules dictate specific tonal arragnements between *even tone* and *oblique tones* within each line of a poem, similar to the stressed and unstressed syllables arranged in metered English poetry. The goal of this project is to construct a Markov matrix to capture the probabilities of transition between different tones, find the equilibrium distribution, and use it to gain insights into whether the observed tonal patterns match the expected ones.

The tonal pattern in (lv shī) have a direct analogy to metrical patterns in English poetry, such as iamic penameter, trochaic tetrameter, and other structured forms where syllable stress plays a defining role. In classical Chinese poetry, each character (or syllable) is categorized intor either even (stable tone), or oblique (rising / falling tone). This binary cateogization resembles the *stressed* and *unstressed* syllables in English Prosody.

Markov matrices are tools used to model transitions between states, where each element M_{ij} represents the probability of moving from state i to state j. In the context of English poetry meters, we can define states as *stressed* (S) or *unstressed* (U) syllables, or even larger metrical units like *iambs* (U-S) and *trochees* (S-U). For instance, an iambic pentameter line alternates between U and S, which can be represented by a simple Markov matrix:

$$M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Indicating that U is always followed by S, and vice versa. By expanding the matrix to allow for deviations (e.g., occasional trochees or spondees), we can analyze the stylistic flexibility of a poem. Furthermore, computing the *equilibrium distribution* of the matrix reveals the long-term proportion of different metrical units, which helps identify dominant rhythmic patterns.

Lastly, here are some additional linguistic features that might be interesting using Markov methods to analyze under the hood:

- Rhyme Schemes: Could applied to different poetry under different "Dynasty" to study how rhymes evolve and how certain poets deviate from common schemes.

- Sentence Length and Complexity: Study transitions between short and long sentences, or simple vs. complex clauses, especially in prose or in political speeches, to identify style...

2 Introduction

The project aims to distinguish the styles of Zhu Shuzhen and Du Fu using Markov matrices by modeling the tonal patterns in their poetry. Each poem is encoded as a sequence of *even* or *oblique* tonal states, with transition probabilities captured in a matrix M, where M_{ij} denotes the probability of transitioning from state i to state j.

Using Markov Matrices to Distinguish Authors:

For each poet, we construct a separate Markov matrix M_{Zhu} and M_{Du} from their corpus, defined as:

$$M_{\text{Zhu}} = \begin{bmatrix} p_{UU} & p_{UZ} \\ p_{ZU} & p_{ZZ} \end{bmatrix}, \quad M_{\text{Du}} = \begin{bmatrix} q_{UU} & q_{UZ} \\ q_{ZU} & q_{ZZ} \end{bmatrix}$$

where p_{ij} and q_{ij} are the probabilities of transitioning from tone i to tone j for Zhu Shuzhen and Du Fu, respectively. These matrices are expected to differ due to each poet's stylistic choices in adhering to classical tonal rules. The equilibrium vector $\mathbf{v}$ for each matrix, satisfying $M\mathbf{v} = \mathbf{v}$, represents the long-term distribution of tonal states.

Strategy 1:

Given a test poem's tonal sequence $x = (x_1, x_2, \ldots, x_n)$, we compute the *log-likelihood* under each author's matrix:

$$\mathcal{L}_{\text{Zhu}} = \sum_{i=1}^{n-1} \log M_{\text{Zhu}}[x_i, x_{i+1}], \quad \mathcal{L}_{\text{Du}} = \sum_{i=1}^{n-1} \log M_{\text{Du}}[x_i, x_{i+1}]$$

If $\mathcal{L}_{\text{Zhu}} > \mathcal{L}_{\text{Du}}$, we classify the poem as Zhu Shuzhen's.

Strategy 2:

Alternatively, we can compare the test poem's equilibrium vector $\mathbf{v}_{\text{test}}$ with $\mathbf{v}_{\text{Zhu}}$ and $\mathbf{v}_{\text{Du}}$ using similarity measures such as *cosine similarity* or *Euclidean distance*.

$$d_{\text{Zhu}} = \|\mathbf{v}_{\text{test}} - \mathbf{v}_{\text{Zhu}}\|, \quad d_{\text{Du}} = \|\mathbf{v}_{\text{test}} - \mathbf{v}_{\text{Du}}\|$$

The author is identified as the one corresponding to the minimum distance and maximum similarity. Moreover, differences in matrix structure or equilibrium distributions may also indicate genre differences, as classical genres often impose distinct tonal patterns.

Baseline Strategy: Lastly, we will introduce method using infinity vector norm to discern authorship. Leveraging the differences between equilibrium vectors between train and test data's markov matrices. The author is identified as the one corresponding to the minimum differences.

Expectations:

Given the strict tonal constraints in classical poetry and known stylistic differences, we expect to distinguish Du Fu and Zhu Shuzhen with reasonable accuracy. I think log-likelihood Strategy might provide robust result, while the other two might show turbulences among

small test-case's discrepancies (Especially for euclidean distance and cosine similarity calculations).

3 Implementation and Data

For my implementation, I used three different strategies to address the author prediction task.

3.1 Algorithm

Algorithm 1 construct_markov_matrix

Require: tones_list: list of tone sequences, num_states: number of possible states
Ensure: A normalized Markov transition matrix with columns summing to 1

Initialize a matrix $transition\_counts$ of size (num_states, num_states)
for each $tones$ in $tones\_list$ **do**
  for $i = 0$ to $length(tones) - 2$ **do**
    $current\_tone \leftarrow tones[i] - 1$
    $next\_tone \leftarrow tones[i + 1] - 1$
    $transition\_counts[current\_tone, next\_tone] \leftarrow$
      $transition\_counts[current\_tone, next\_tone] + 1$
  end for
end for
for $j = 0$ to $num\_states - 1$ **do**
  $col\_sum \leftarrow \sum_{i=0}^{num\_states-1} transition\_counts[i, j]$
  if $col\_sum > 0$ **then**
    for $i = 0$ to $num\_states - 1$ **do**
      $transition\_counts[i, j] \leftarrow transition\_counts[i, j]/col\_sum$
    end for
  end if
end for
return $transition\_counts$

Algorithm 2 equilibrium_vector

Require: matrix: a Markov transition matrix
Ensure: The equilibrium vector of the matrix

Find the index idx of the eigenvalue closest to 1
Set $equilibrium\_vec \leftarrow$ the eigenvector corresponding to idx
Convert $equilibrium\_vec$ to real values by taking the real part of the vector
Normalize $equilibrium\_vec$ by dividing each element by the sum of all elements in $equilibrium\_vec$
return $equilibrium\_vec$

Algorithm 3 cosine_similarity

Require: $vec1$, $vec2$

Ensure: The cosine similarity between $vec1$ and $vec2$

1: Convert $vec1$ and $vec2$ to SageMath vectors
2: Compute the dot product of $vec1$ and $vec2$
3: Compute the norms of $vec1$ and $vec2$
4: Divide the dot product by the product of the norms to compute cosine similarity
5: **return** $cosine_similarity$

Algorithm 4 euclidean_distance

Require: $vec1$, $vec2$

Ensure: The Euclidean distance between $vec1$ and $vec2$

1: Convert $vec1$ and $vec2$ to SageMath vectors
2: Compute the difference between $vec1$ and $vec2$
3: Compute the norm of the difference
4: **return** $norm\_diff$

Algorithm 5 infinity_norm

Require: $test\_vector$, $input\_vec$

Ensure: The infinity norm difference between $test\_vector$ and $input\_vec$

1: Compute the element-wise absolute difference between $test\_vector$ and $input\_vec$
2: Find the maximum value of the absolute differences
3: **return** $norm\_diff$

3.2 Data Results:

Log-likelihood Calculation:

Zhu Shuzhen's Markov Matrix:

$$\begin{bmatrix} \frac{3}{10} & \frac{47}{190} & \frac{29}{190} & \frac{3}{10} \\ \frac{53}{156} & \frac{1}{4} & \frac{11}{78} & \frac{7}{26} \\ \frac{35}{103} & \frac{27}{103} & \frac{15}{103} & \frac{26}{103} \\ \frac{55}{199} & \frac{56}{199} & \frac{27}{199} & \frac{61}{199} \end{bmatrix}$$

Du Fu's Markov Matrix:

$$\begin{bmatrix} \frac{1}{3} & \frac{29}{103} & \frac{46}{309} & \frac{73}{309} \\ \frac{88}{279} & \frac{73}{279} & \frac{49}{279} & \frac{23}{93} \\ \frac{27}{80} & \frac{41}{160} & \frac{1}{10} & \frac{49}{160} \\ \frac{69}{236} & \frac{37}{118} & \frac{21}{118} & \frac{51}{236} \end{bmatrix}$$

Test Case for Zhu Shuzhen:

Total Log-Likelihood for Zhu Shuzhen:

$$\begin{aligned} &3\log\left(\frac{35}{103}\right) + 5\log\left(\frac{53}{156}\right) + 4\log\left(\frac{61}{199}\right) + 3\log\left(\frac{3}{10}\right) \\ &+ 6\log\left(\frac{56}{199}\right) + 3\log\left(\frac{55}{199}\right) + 2\log\left(\frac{7}{26}\right) + 5\log\left(\frac{26}{103}\right) \\ &+ 5\log\left(\frac{1}{4}\right) + 3\log\left(\frac{47}{190}\right) + 4\log\left(\frac{29}{190}\right) + 2\log\left(\frac{15}{103}\right) \\ &+ 2\log\left(\frac{11}{78}\right) + \log\left(\frac{27}{199}\right) \end{aligned}$$

Total Log-Likelihood for Du Fu:

$$\begin{aligned} &3\log\left(\frac{27}{80}\right) + \log\left(\frac{1}{3}\right) + 5\log\left(\frac{88}{279}\right) + 6\log\left(\frac{37}{118}\right) \\ &+ 5\log\left(\frac{49}{160}\right) + 3\log\left(\frac{69}{236}\right) + 3\log\left(\frac{29}{103}\right) + 5\log\left(\frac{73}{279}\right) \\ &+ 2\log\left(\frac{23}{93}\right) + 2\log\left(\frac{73}{309}\right) + 4\log\left(\frac{51}{236}\right) + \log\left(\frac{21}{118}\right) \\ &+ 2\log\left(\frac{49}{279}\right) + 4\log\left(\frac{46}{309}\right) + 2\log\left(\frac{1}{10}\right) \end{aligned}$$

Predicted Author for Zhu Shuzhen's Test Tones:

Zhu Shuzhen

Test Case for Du Fu:

Total Log-Likelihood for Zhu Shuzhen:

$$\begin{aligned}
&3\log\left(\frac{56}{169}\right) + 4\log\left(\frac{61}{186}\right) + \log\left(\frac{29}{93}\right) + 3\log\left(\frac{19}{62}\right) \\
&+ 3\log\left(\frac{9}{31}\right) + 2\log\left(\frac{57}{200}\right) + 5\log\left(\frac{47}{169}\right) + 2\log\left(\frac{11}{40}\right) \\
&+ 4\log\left(\frac{53}{200}\right) + 2\log\left(\frac{22}{93}\right) + 6\log\left(\frac{3}{13}\right) + 5\log\left(\frac{7}{31}\right) \\
&+ 6\log\left(\frac{7}{40}\right) + \log\left(\frac{27}{169}\right)
\end{aligned}$$

Total Log-Likelihood for Du Fu:

$$\begin{aligned}
&2\log\left(\frac{103}{314}\right) + 2\log\left(\frac{49}{153}\right) + 5\log\left(\frac{87}{275}\right) + 3\log\left(\frac{73}{242}\right) \\
&+ \log\left(\frac{46}{153}\right) + 5\log\left(\frac{69}{242}\right) + 4\log\left(\frac{44}{157}\right) + 3\log\left(\frac{14}{51}\right) \\
&+ 3\log\left(\frac{74}{275}\right) + 6\log\left(\frac{73}{275}\right) + 2\log\left(\frac{69}{314}\right) + 4\log\left(\frac{51}{242}\right) \\
&+ 6\log\left(\frac{27}{157}\right) + \log\left(\frac{41}{275}\right)
\end{aligned}$$

Predicted Author for Du Fu's Test Tones:

Du Fu

Test Case for Mystery Test Tones:

Total Log-Likelihood for Zhu Shuzhen:

$$\begin{aligned}
&\log\left(\frac{56}{169}\right) + 3\log\left(\frac{61}{186}\right) + 2\log\left(\frac{29}{93}\right) + \log\left(\frac{19}{62}\right) \\
&+ 3\log\left(\frac{9}{31}\right) + 3\log\left(\frac{57}{200}\right) + 7\log\left(\frac{47}{169}\right) + 7\log\left(\frac{11}{40}\right) \\
&+ 2\log\left(\frac{53}{200}\right) + \log\left(\frac{22}{93}\right) + 6\log\left(\frac{3}{13}\right) + 6\log\left(\frac{7}{31}\right) \\
&+ \log\left(\frac{27}{169}\right) + 5\log\left(\frac{13}{93}\right)
\end{aligned}$$

Total Log-Likelihood for Du Fu:

$$
\begin{aligned}
&3\log\left(\frac{103}{314}\right) + \log\left(\frac{49}{153}\right) + 7\log\left(\frac{87}{275}\right) + \log\left(\frac{73}{242}\right) \\
&+ 2\log\left(\frac{46}{153}\right) + 6\log\left(\frac{69}{242}\right) + 2\log\left(\frac{44}{157}\right) + 3\log\left(\frac{14}{51}\right) \\
&+ \log\left(\frac{74}{275}\right) + 6\log\left(\frac{73}{275}\right) + 7\log\left(\frac{69}{314}\right) + 3\log\left(\frac{51}{242}\right) \\
&+ 5\log\left(\frac{49}{242}\right) + \log\left(\frac{41}{275}\right)
\end{aligned}
$$

Predicted Author for Mystery Test Tones:

Du Fu

Euclidean Distance and Consine Similarity Calculation:

Predicted Author for Zhu Shuzhen's Test Tones:

- Cosine Similarity with Zhu Shuzhen: 0.9789041921911098
- Cosine Similarity with Du Fu: 0.9729092716043023
- Euclidean Distance to Zhu Shuzhen: 0.10495126147082444
- Euclidean Distance to Du Fu: 0.11872876556336236
- Cosine Similarity Prediction: Zhu Shuzhen
- Euclidean Distance Prediction: Zhu Shuzhen

Predicted Author for Du Fu's Test Tones:

- Cosine Similarity with Zhu Shuzhen: 0.9598323728518315
- Cosine Similarity with Du Fu: 0.9755859048660054
- Euclidean Distance to Zhu Shuzhen: 0.1482684467244978
- Euclidean Distance to Du Fu: 0.11595078252667476
- Cosine Similarity Prediction: Du Fu
- Euclidean Distance Prediction: Du Fu

Predicted Author for Mystery Test Tones:

- Cosine Similarity with Zhu Shuzhen: 0.9877493067500825
- Cosine Similarity with Du Fu: 0.9860787754087518
- Euclidean Distance to Zhu Shuzhen: 0.08160079380597142
- Euclidean Distance to Du Fu: 0.08693730036230193
- Cosine Similarity Prediction: Zhu Shuzhen
- Euclidean Distance Prediction: Zhu Shuzhen

Baseline Method - Infinity Vector Norm Calculation:

Baseline Infinity Norm Prediction for Zhu Shuzhen's test tones:

- Infinity Norm Difference with Zhu Shuzhen: 0.07692068226362686
- Infinity Norm Difference with Du Fu: 0.09668192553969868
- Baseline Infinity Norm Prediction: Zhu Shuzhen

Baseline Infinity Norm Prediction for Du Fu's test tones:

- Infinity Norm Difference with Zhu Shuzhen: 0.12762972269326903
- Infinity Norm Difference with Du Fu: 0.08415719396641286
- Baseline Infinity Norm Prediction: Du Fu

Baseline Infinity Norm Prediction for Mystery test tones:

- Infinity Norm Difference with Zhu Shuzhen: 0.12762972269326903
- Infinity Norm Difference with Du Fu: 0.08415719396641286
- Baseline Infinity Norm Prediction: Du Fu

4 Analysis

This project distinguish the tonal styles of Zhu Shuzhen and Du Fu using Markov matrices by using various computational methods. The test sequence was analyzed using the **log-likelihood method**, equilibrium vector comparison via **Cosine Similarity** and **Euclidean Distance**, as well as a baseline approach using the **infinity norm**.

Log-Likelihood Method

The log-likelihood method correctly predicted the corresponding authors refer to two given test cases.
Supposingly, the reason this method worked well is that it directly uses the transition probabilities from the Markov matrix, which are based on observed tonal patterns in each poet's work. Even with a small test case, the log-likelihood method effectively quantifies how likely the transitions are for each poet, resulting in an accurate prediction. This approach is particularly robust when dealing with small datasets, as it inherently accounts for the probabilities of each transition in a cumulative manner.
Thus, for the additional Mystery Test Case, I will use its result as baseline to come up with potential reasoning.

Cosine Similarity and Euclidean Distance Methods

The equilibrium vector comparison using cosine similarity and Euclidean distance also predicted the correct authorship for both test cases. The cosine similarity values for Zhu Shuzhen's test tones were higher when compared to her own matrix (0.9789) versus Du Fu's (0.9729), and the Euclidean distance was lower for Zhu Shuzhen (0.1049) than Du Fu (0.1187), leading to the correct prediction for Zhu Shuzhen. Similarly, Du Fu's test tones showed higher cosine similarity with his own matrix (0.9756) than Zhu Shuzhen's (0.9598), and a lower Euclidean distance (0.1159) compared to Zhu Shuzhen (0.1483), resulting in the correct prediction for Du Fu.
While these methods performed well, slight discrepancies in the similarity scores and distances suggest that vector-based methods may be less robust when dealing with small datasets. Thus, such characteristics causes seemingly incorrect prediction in the additional Mystery Test Case.
Here are several potential reasons that might render prediction inconsistency:

- **Small Test Case Size:** The limited size of the test sequence means that the calculated Markov matrix and equilibrium vector do not fully capture the general stylistic information. Base on my observation, the additional Mystery Test Case is smaller than the other two.
- **Overfitting to Local Patterns:** The equilibrium vector derived from the small dataset may overemphasize specific transitions that are not representative, leading to closer matches incorrect author's matrix.
- **Sensitivity of Similarity Measures:** Cosine similarity and Euclidean distance are sensitive to the direction and magnitude of the vectors. Given the sparse data, these measures may incorrectly favor the equilibrium vector.

Baseline Infinity Norm Method

The baseline infinity norm method also correctly predicted Zhu Shuzhen and Du Fu as the authors of their respective test tones. The infinity norm differences were smaller for the

correct author in both test cases, with Zhu Shuzhen's test tones showing an infinity norm difference of 0.0769 versus 0.0967 for Du Fu, and Du Fu's test tones showing a difference of 0.0842 versus 0.1276 for Zhu Shuzhen.
Similar to second strategy, while these "distance" methods performed well, slight discrepancies in the similarity scores and distances suggest they are sensitive to small test case.
However, probably due to the simplicity within the caluclation, the baseline method "captured" somewhat partly of the generic tonal behavior of the Mystery Test Tone, giving out a correct prediction consistent with the log-likelihood method. (This is surprising!)

5 Conclusion

The log-likelihood method provided highly accurate results, leveraging the probabilistic nature of the transition matrices and proving particularly reliable with limited data. The cosine similarity, Euclidean distance, and infinity norm methods also correctly predicted the authorship in both cases, though minor discrepancies in their values suggest that these methods may be more sensitive to small datasets.
Overall, the results demonstrate that while all methods performed well in this case, the log-likelihood method stands out for its robustness. The vector-based methods, including cosine similarity, Euclidean distance, and the infinity norm, provide correct predictions but may benefit from larger datasets or further refinement to more accurately capture subtle tonal transitions and stylistic nuances in general.