R&B Chord Generator Outline

Randa Ampah, Stephanie Delgadillo-Cartagena, Isabel Delgado, Mohini Gupta, Yani Iben, Asmita Kadam, and Aniyah McWilliams
DS 5030: Understanding Uncertainty
Fall 2025

The dataset that we used for our analysis was the Chordonomicon dataset. It is a dataset that consists of 679,807 songs containing information about each song's chord progressions, release date information, main and subgenre information, and Spotify song and artist ids. The data was collected by a group of scholars at the National Technical University of Athens, Greece for the purposes of furthering music exploration and experimentation by providing a large, widely available, and well structured dataset for the general public to access, something they viewed as lacking in the music domain. The focus of the data is the chord progressions within each song, as they serve as the backbone of compositions. The chord data itself was obtained by scraping the Ultimate Guitar platform, while the metadata was obtained from the Spotify API.

The variables in the original dataset were as follows:

id: a song's id within the original Chordonomicon dataset
chords: the chords present in the song, including what line of the music the chords are in
release_date: the date the song was released, formatted as YYYY-MM-DD
genres: the more specific subgenres that song falls under
decade: the decade the song was released in
rock_genre: the type of rock subgenre the song falls under (Null if the song is not a rock song)
artist_id: the id of the artist within the Chordonomicon dataset
main_genre: the general genre the song falls under
spotify_song_id: the id for the song within the Spotify app
spotify_artist_id: the id for the song's artist within the Spotify app

For our analysis, we decided to focus on songs within the R&B genre. After subsetting the genres variable to only include songs with "r&b", the subsetted dataset contained 5,581 songs. We then chose to focus on songs from the past 10 years, which limited our final dataset to 3,420 songs.

The phenomenon that we are modeling is the structure and progression of chords in a song via a Markov Chain. A markov chain is a probabilistic model that describes that in a sequence of events, the probability of the next event depends only on the current state. It uses a state space and the probabilities of moving from one state to the other. This probabilistic model captures the idea that the probability of future events is conditionally independent of past states

(only depends on what the current state is, not what all the past states have been), given the current state.

The objective of this project is to build a chord generator that intakes the chords of R&B songs, and outputs a generated song. In the context of this project, Markov Chains provide a great framework for modeling chord transitions. The state space for this Markov Chain are the chords that regularly occur in R&B songs. Based on these states we create a transition matrix to calculate the probability of each chord transition. By analyzing the chord progression in R&B songs, we are able to estimate the probabilities of moving from one chord to the next. These estimated probabilities are then utilized to generate an R&B song. For example, based on our transition matrix the probability of the Emaj7/A chord transitioning to the F chord is 0.964. This means that 96.4% of the time the F chord is the next chord to play after the Emaj7/A chord. Using the entirety of the transition matrix we can theoretically generate our own RnB song based on the probability of the sequence transitions.

The non-parametric model is the First-Order Markov Transition Model. The model is suitable for fitting sequential data,like time-series, user behavior paths etc, where the probability of the next state only depends on the current state and not the entire history. The model consists of a finite set of states and an NxN Transition Probability Matrix, where P represents the probability of transitioning from one state to the next.

We fit the model by using a Maximum Likelihood Estimation directly from the observed data. The process is entirely empirical, allowing the model to be non-parametric.
First, we used discretization. This means that if the phenomenon's data is continuous, we set the data into a finite set of meaningful states. Next, we counted transitions. This meant that we iterated through the training sequence and counted the number of times a transition occurred from one state to the next. This gives us a count matrix. Finally, we estimated the transition probability by normalizing the counts across the rows. This model captures the short-term temporal dependencies within the data, replicating the observed one-step conditional probabilities. This allows for realistic properties related to the immediate successor of any given state.

In terms of challenges, it was difficult to get the final music output to sound appealing. This challenge came about in two forms. First, we used a smaller sample set of only 80 songs, meaning that there wasn't enough data to draw upon and yield a proper result. To address this first challenge, our group expanded our sample set to 5,581 songs, focused on the r&b category. We also ran into a second problem of the music output not having a cohesive sound. Our team suspected that this issue might be because of varying song trends over the years. To address the problem we tried to standardize the subset more by filtering upon the last 10 years. Then, adding

the diagonal made it more likely to repeat chords, a technique that is used frequently within actual songs, causing the end result to sound more melodic.

Our generated chord sequences demonstrate some of the same harmonic properties as the training data, however, they do not fully capture the structure or phrasing of the original songs. This limitation arises because our model is a first-order Markov model, meaning it predicts each chord based only on the previous one. As a result, it cannot model longer musical patterns such as recurring verse chorus structures causing the generated sequences to sometimes sound more random or disconnected.

In terms of reliability and uncertainty, the model's estimates are credible for common chord transitions, particularly those that occur frequently across the dataset. However, there is substantial uncertainty for rare or complex chords, whose transition probabilities are less stable due to limited training examples. To evaluate this uncertainty, we conducted a log-likelihood test, which measures how well the model predicts unseen chord progressions based on the probabilities it learned. Our result of –2.7 indicates that the model assigns roughly a 6% probability to the correct next chord on average, showing that it has learned meaningful harmonic tendencies, though not with high confidence.

Overall, the model captures some of the musical characteristics and progression tendencies of the training data but still lacks the long term structure and musical predictability found in real songs.

Our current model is limited by the state space, since generating a song or chord progression based on these transitions can lead to predictability and lack of nuance in a song. Typically, when creating a song, other instrumental tools are used such as synthesizers and mixers which aid in defining a song as part of one music genre versus another. Since we do not utilize music21's music manipulation tools in our model, such as pitch or note manipulation, the generated song is limited to producing a basic R&B chord progression.

For future research, we could consider how our model predictions would change if we manipulated the rhythm of the song. Our current model uses a randomly generated rhythmic pattern, which does not yield an output that resembles an authentic R&B instrumental. For future work, we can incorporate rhythmic state spaces into our model, which could produce a less mechanical output and generate a song that sounds closer to the R&B genre.

With these improvements, our model could be used to help R&B artists generate a rough demo or a more realistic chord structure for their songs. The model could also be used to assist artists who are experiencing creative block by providing common progressions based on an initial chord.

References

Kantarelis, S., Thomas, K., Lyberatos, V., Dervakos, E., & Stamou, G. (2024). CHORDONOMICON: A Dataset of 666,000 Songs and their Chord Progressions. arXiv preprint arXiv:2410.22046.