# Analyzing Musical Genres and Trends with Topic Models

Team 20:
Sam Lichtenberg, Ryan McNellis, Yolanda (Yo-Tzu) Yeh, Rebecca Zhang

May 13, 2014

## Abstract

Using an LDA topic model and document influence model, we analyzed 856 pieces of music from three distinct genres and nine subgenres. We found that nine topics, defined by their most common three-chord sequences, were able to distinguish the subgenres, identify the most influential pieces, and track the changes in frequency of chord progressions over musical history.

## Motivation

One of the main goals of machine learning is to gain insights from a large data set that are not obvious to the human eye. Occurrences in the physical world (DNA sequences) and human processes (language, document organization) are natural applications of this tool. Less common, however, is to analyze data that are more artistic and exhibit patterns less clearly. Music is often considered the one medium that allows for pure creation, but the study of music theory shows that there can actually be very rigid structures underlying the composition.

The motivation of our project is to study genres in music and look at trends in different musical ideas over time. We believed that topic models are a natural approach to the problem, and had not seen it applied to this medium in class. The way we approached the problem of what data to use from music was to look at hand labelled roman numeral chord progressions as a text representation of songs. This allows us to apply the modeling techniques typically applied to text documents in different types of topic modeling.

## Data

The data comes from the Prosemus project, from the Universitat Pompeu Fabra and the Universidad de Alicante, containing an academic repository of 856 pieces that have hand-labelled roman numeral chords. Roman numeral notation represents chords in relation to the root note and is the clearest way to represent chord progressions in music, and characterize the harmonies that make up a piece of music. One of the advantages is that this notation shows relative chord transitions and allows us to look at all the songs without worrying about the differences in their keys. This collection of songs came with three genre divisions (popular, jazz, academic) and nine subgenres (pop, blues, and celtic within popular, pre-bop, bop, and bossanova within jazz, baroque, classical, and romanticism within academic). We had at least 50 songs in each subgenre, which we believe to be sufficient to produce meaningful insights. For use in dynamic topic modeling, we hand-labelled composition years for all pieces in the academic genre in order to look at trends within topics overtime.

## Genre Analysis

### Methods

Latent Dirichlet Allocation is a generative probabilistic model for discrete data [**?**]. It is typically used to analyze large text corpora. LDA makes the assumption that the documents within a corpus are exchangeable and that the words within a document are exchangeable. This allows LDA to be described by the graphical model in Figure 1a. An equivalent way to describe LDA is to describe the generative, document-creating process it models:

1. Choose $N \sim$Poisson($\eta$)
2. Choose $\theta \sim$Dir($\alpha$)
3. For each of the $N$ words $w_n$:
   (a) Choose a topic $z_n \sim$Multinomial($\theta$)
   (b) Choose a word $w_n$ from $p(w_n|z_n, \beta)$, a multinomial probability given the topic $z_n$ and topic distributions over words $\beta$

We assume, for LDA, that the number of topics $k$ is known. In practice, this is not the case, but we can circumvent this issue by performing cross-validation and choosing the value of $k$ with the highest log-likelihood on the held-out documents.

To make use of LDA, we need to compute the posterior distribution

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}$$

but this distribution is intractable because of the denominator. Therefore we need to find an approximate solution. Two commonly-used methods are Markov Chain Monte Carlo, which approximates via repeated random sampling, and variational inference, which finds a simpler family of distributions, indexed by variational parameters, that lower bound the log likelihood of the distribution of interest and minimize the Kullback-Liebler divergence between it and the simpler distribution. The variational parameters that minimize this are the solution to an optimization problem which can be solved by iterative fixed-point methods.

Then, to compute the model parameters $\alpha$ and $\beta$, an Expectation-Maximization algorithm is used. In the E-step, the lower bound on the posterior is maximized with respect to the variational parameters just described; in the M-step, those parameters are held fixed and the posterior lower bound is maximized with respect to $\alpha$ and $\beta$.

We first looked at individual chords of a song, building an unsupervised topic model based on the distribution of chords observed. What we found was that common chords such as the tonic and dominant (I, V) were prevalent in any genre of music, and so looking at chords independently did not provide enough information to distinguish between different styles of music. Based on what we knew about common rules in music theory, we hypothesized that looking at chord transitions or series of chord progressions would yield more information specific to each genre.

To select the optimal number of topics for our unsupervised topic model, we wrote a script to run five-fold cross validation on bigram chord transitions. We computed estimated log-likelihood per selection of number of topics by running inference on held-out data from one fold, based on the topic model created by the other folds. We tried this with the number of topics ranging from three (for the genres) to nine (for the total subgenres).

Looking at the estimated likelihoods versus number of topics seen in Figure 2, the differences in mean likelihood were negligible in comparison to the amount of standard error between likelihoods found for each combination of folds. We observed that choosing nine topics yielded the least amount of standard error, suggesting running our unsupervised topic model with nine topics would produce the most reliable and consistent results.

Using this, we ran our topic model for nine topics, and made several notable observations by comparing the proportions of the different topics for each genre and looking at the chord transitions associated with topics. At first, we chose to run this on transitions between two chords because we had done cross validation on these, but later saw that three-chord transitions were more distinct across genres and more meaningful than transitions between two chords. As for four-chord transitions, these have very low frequency counts and the common ones are often permutations of two-chord pairs (I-V-I-V) which would yield less meaningful insights.

**Results**

Looking at Figure 3 and referencing Figure 4 for the topic descriptions, we notice that songs under the broader academic genre have high proportions in topic 1. This makes sense as the most common chords in topic 1 are transitions between tonic and dominant chords, which we would expect in classical music - the most often in Mozart, less so in Bach, and least in Chopin as the romantic era became more free-form. On the other hand, topic 1 is the least prevalent across all subgenres of jazz, with a proportion near zero, showing that jazz music differentiates itself by avoiding these standard chord progressions.

In fact, what we notice about the distribution of jazz songs into the different topics is that it is slightly more uniform than classical or pop songs. This demonstrates that jazz songs use a greater variety of chord progressions, which might stem from the nature of improvisation in jazz and the general style. From what we know about jazz music, in later jazz, musicians disregarded standard chord progressions altogether, which further coincides with what we observe about the trough in topic 1. Instead, topic 6 is dominant in the jazz genre and rare in the others. The chord sequence most common in topic 6 is the II-V-I progression, using both major and dominant 7ths; both the progression and the use of 7ths are highly associated with jazz music.

If we look specifically at the distribution of proportion of topics across the subgenres of pop, we find a particular distinction between blues and celtic. For example, celtic songs overwhelmingly exhibit progressions of topic 1, whereas blues songs have their highest proportion at topic 0. Looking at the chord transitions associated with these topics, we find that topic 1 includes more transitions between dominant and tonic chords, which sound resolved and are used as cadences to end phrases. At the same time, it doesn't exhibit as much of the II chord as some other topics, which means there is not much modulation to the dominant key (the II chord is the V chord in the dominant key). Both these observations make sense as celtic music is influenced by folk music that is highly repetitive in terms of harmony, repeating simple non-modulating phrases while varying other characteristics such as rhythm. On the other hand, what's distinctive about topic 0 are the many IV7-to-tonic chord transitions, a progression with a dissonant half-step that is typical of blues music. The distribution for celtic music resembles those of the academic music subgenres, which is unsurprising; even though the researchers who prepared this data set chose to label celtic as popular music, it is much more commonly associated with instrumental, classical music.

Overall, within the three genres, we see that the distributions are similar - an indication that the hidden topics are strongly related to key genre characteristics, and that topic modeling was successful. At the same time, there are interesting unique features specific to each subgenre, evidence that the topics are enough to distinguish between different types of music.

## Trend Analysis

### Background

The Dynamic Influence Model (DIM) is an extension to LDA that also models both how topics changes over time and how certain documents influence those topical shifts more than others. It was originally developed to model influential academic papers, as a different approach than simple citation counts. DIM is represented by the graphical model of Figure 1b and by the generative process, for a given time slice $t$, as follows:

1. Draw topics $\beta_t|\beta_{t-1},(w,l,z)_{t,1:D} \sim N(\beta_{k,t}+\exp(-\beta_{k,t})\sum_d l_{d,k}\sum_n w_{d,n}z_{d,n,k}\sigma^2 I)$
2. Draw $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1},\delta^2 I)$
3. For each document
   (a) Draw $\eta \sim N(\alpha_t, a^2 I)$
   (b) For each word:
      i. Draw $Z \sim Mult(\pi(\eta))$
      ii. Draw $W_{t,d,n} \sim Mult(\pi(\beta_{t,z}))$

Where

$$\pi(\beta_{k,t})_w = \frac{exp(\beta_{k,t,w})}{\sum_w \beta_{k,t,w})}$$

The purpose of these adjustments, such as swapping the Dirichlet distribution with a Gaussian noise model, is to model the fact that the documents are now considered sequential data, not exchangeable. The posterior is approximated by treating the variational parameters as outputs of a Kalman filter.

**Results**

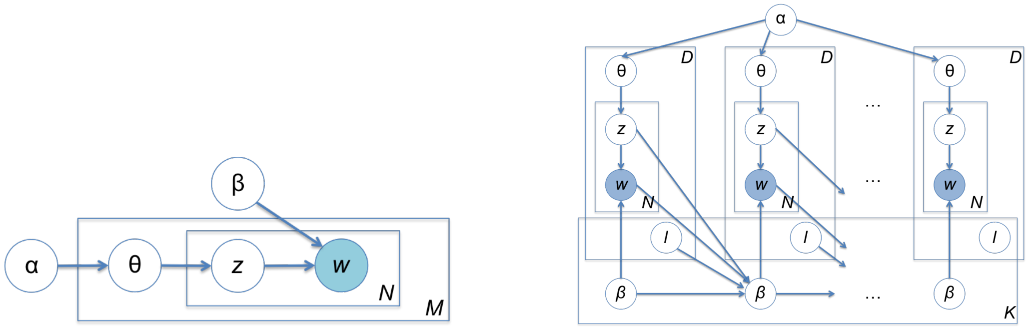
After labeling the composition years for pieces in baroque, classical, and romanticism subgenres, we ran a dynamic topic model on three topics to capture the changes in frequency of three-chord progressions across time-slices, each with 25 pieces. As seen in Figure 5, we found interesting trends in the topics over time, both expected ones based on our musical knowledge as well as not so obvious ones. Consider topic 1: the progression V7-I-IV indicates a plagal cadence or "Amen cadence" (IV-I) which decreased in frequency as fewer composers wrote for the church. Progressions like III-VII7-III appear in minor pieces that have modulated into the relative major, seen in the late classical works of Beethoven and Schubert as they experienced a turning inward, contemplative style of composition. On the other hand, the I-V-I progressions, which have always been the most common, appeared less in the topic over time; meanwhile the I-V7-I progression, essentially the same except with an additional note, became more frequent. This is an interesting result as it suggests the more complex V7 replaced the V over time.

We then fit an influence model to pinpoint musical works that have had a significant impact on chord transitions over time. For each topic, the model assigns an "influence score" to musical works that corresponds to how influential the work is on that topic. Figure 6 displays boxplots corresponding to the influence score of all works in each time slice. If a composition had a significant impact on a particular topic, we would expect its influence score to be significantly higher than that of the rest of the works from that era. Thus, we consider a composition to be significant if its influence score is displayed as a high outlier in our boxplot. In the boxplots, one can see that there are very high outliers in the earliest time slice for topic 0 and topic 1. The top three outliers correspond to Vivaldi's The Four Seasons Autumn movement 2 and Handel's Water Music suite, works known for their influence, as well as yielding surprising results like Bach's Lute prelude.

## Further Work

One possible extension of our work is to find a good way to measure the influence of a piece in order to compare this to our results. This would be similar to using the number of citations as a reference for how influential a scientific paper is.

Another possible extension of our work is to build a music recommendation system based on the underlying topics fitted. We implemented a basic system by looking at the Euclidean distance between topic proportions of songs, and we think that a recommendation system based on this could be successful given more data per genre.




(a) (b)

Figure 1: Graphical Model of (a) Latent Dirichlet Allocation (b) Dynamic Topic Model

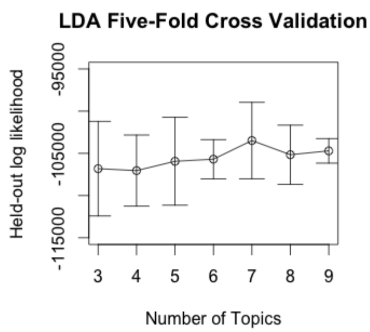



Figure 2: LDA Five-fold Cross Validation

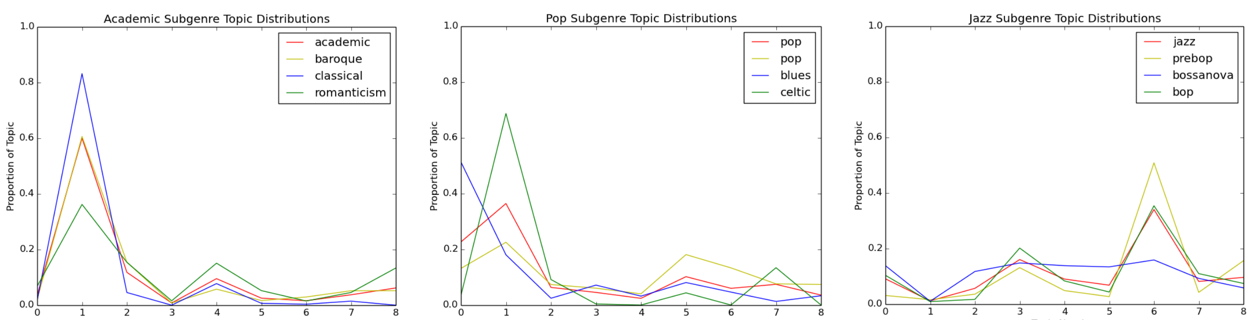



Figure 3: Topic Proportions in Genres

| Topic Number | Top Three-Chord Sequences | | |
|---|---|---|---|
| 0 | I7-IV7-I7 | I-VIIb-I | I-I7-IV7 |
| 1 | I-V7-I | I-V-I | I-IV-I |
| 2 | Im-V7-Im | I-IIm-I | III-VII7-III |
| 3 | IIm7-V7-I | VIm-II-VIm | IIIm7-VI7-IIm7 |
| 4 | VIm-III7-Vim | I-Idim-I | IIm7b5-V7-Im7 |
| 5 | I-I7-IV | V7-I-Vim | IIm-V7-I |
| 6 | IIm7-V7-I | IIm7-V7-Imaj7 | VIm7, IIm7, V7 |
| 7 | Im-VII-Im | IV-V-IV | Im7-IV7-Im7 |
| 8 | IIm7-V7-I | Im-V7-Im | IIIm7-VI7-IIm7 |

Figure 4: Most Common Chords for Each Topic

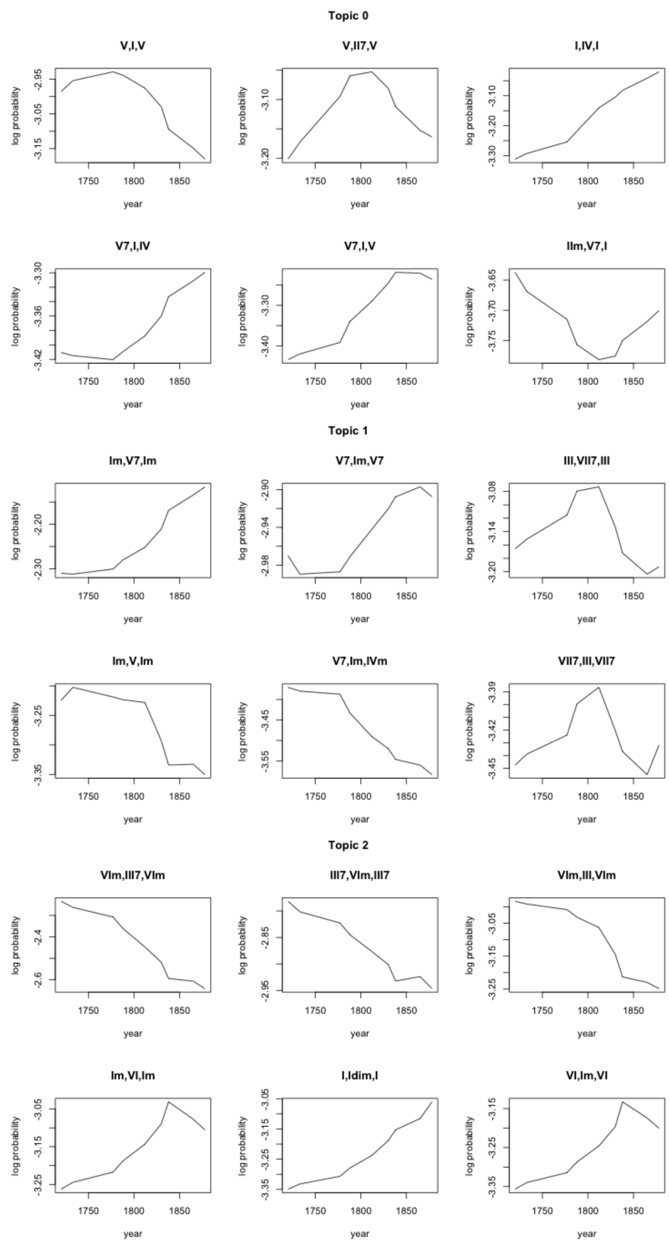



Figure 5: Trends in Topics Over Time (Academic)

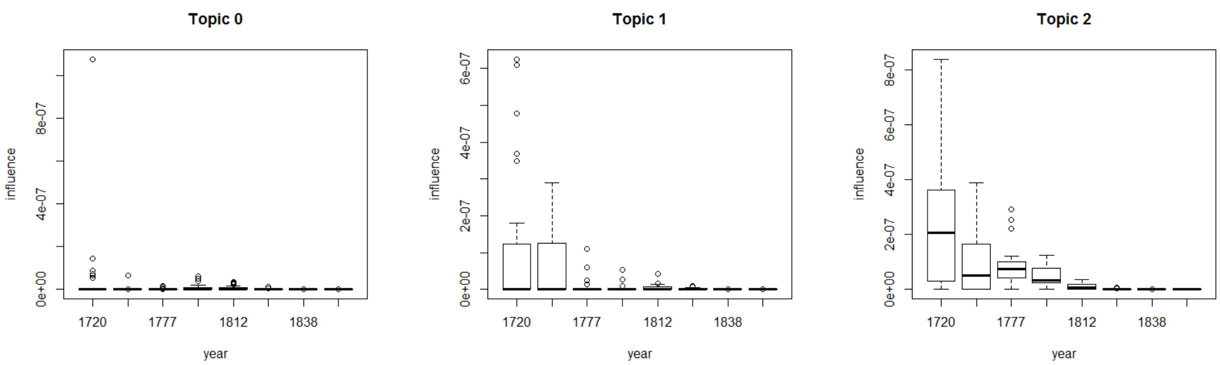



Figure 6: Boxplots of Song Influence in Academic Music