Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
D		2103070

The MCM Thesis of Team 2103070

Summary

Music genres have gradually diverged since the middle of the last century. With the development of modern technology and people's awakening of the pursuit of entertainment and spiritual life, music creation and change have gradually become complicated and common. After the inhere knowledge of music theory (pause, rhythm or tone, etc.) becomes the foundation, the evolution and development of genres, the interaction between artists and artists, and the correlation between economic society and music have become more quantifiable. Based on quantifying the past mutual influence mechanism of the music industry and the reverse influence mechanism of human economic and social emotional life on music, we have mainly established 3 models for the 7 questions of this question.

The $\mathbf{1^{st}}$ **model** simply considers the relationship between people's influence over time and domains, and uses a complex network model to establish a music influence model.

The $\mathbf{2^{nd}}$ **model** incorporates the factors of music itself, considers and compares the similarities and differences at the genre level, and designs four main analysis methods in the middle. Finally, what determines the answer to the core proposition of genre through the decision tree.

The $\mathbf{3^{rd}}$ **model** involves more abstract and grand influential factors such as time, economy, society, technology, national policies, and artist thinking. Lengthen the influence cycle and expand the scope of influence under the fully quantitative model mentioned above. Involving five sub-analysis, and discussing the corresponding era characteristics behind.

Finally, we appropriately extend the discussion on more influencing factors,give our own thoughts;and make objective judgments on the advantages and disadvantages of the model apart from practical application explanations.Based on the existing conclusions, suggestions are given for the future, and comprehensive assessment is made for the promotion of genre development, diversified development of music, and improvement of humanistic and artistic perception.

Keywords: Complex Network, Principal component analysis, Decision tree, Euclidean distance, Similarity Analysis, Logistic Analysis

Contents

1 Introduction

1.1 Problem Background

Figure 1: Musicians&Genres&Music&Life

1.2 Restatement of Problem

- Use the influence_data data set or portions of it to create a (multiple) directed network(s) of musical influence,where influencers are connected to followers.Develop parameters that capture 'music influence' in this network.Explore a subset of musical influence by creating a sub network of your directed influencer network.Describe this sub network.What do your 'music influence 'measures reveal in this sub network?

- Use full_music_data and/or the two summary data sets (with artists and years) of music characteristics, to develop measures of music similarity.Using your measure,are artists within genre more similar than artists between genres?

- Compare similarities and influences between and within genres.What distinguishes a genre and how do genres change over time?Are some genres related to others?

- Indicate whether the similarity data,as reported in the data_influence data set,suggest that the identified influencers in fact influence the respective artists.Do the 'influencers' actually affected the music created by the followers?Are some music characteristics more 'contagious' than others,or do they all have similar roles in influencing a particular artist's music?

- Identify if there are characteristics that might signify revolutions (major leaps) in musical evolution from these data? What artists represent revolutionaries (influencers of major change) in your network?

- Analyze the influence processes of musical evolution that occurred over time in one genre. Can your team identify indicators that reveal the dynamic influencers, and explain how the genre(s) or artist(s) changed over time?

- How does your work express information about cultural influence of music in time or circumstances? Alternatively, how can the effects of social, political or technological changes (such as the internet) be identified within the network?

1.3 Literature Review

This quantitative music is not the first, but also has achieved certain results in specific aspects,Judging from the existing research in the field of sociology, artificial intelligence, and even economics at home and abroad.

- for the research on the characteristics of music itself, there have been research on the recognition and generation of music style based on deep learning(quoted from reference[1]etc.), and the analysis of music emotion characteristics based on feature vectors(quoted from reference[13]etc.).
- the genre research of artists also owns its in-depth studies, such as judging genre attribution through early scientific methods like deep attention mechanism(quoted from reference[10]etc.)
- the relevance of music and social life changes based on temporal research(quoted from[7][11]etc.

However, studying the interaction of music for music, music for musicians, musicians for music, musicians for musicians, genres for genres etc is still a systemic and complex subject that needs to be further improved in algorithm optimization,fitting validity and accuracy. It is also the key stage of our research. The meaning lies.

1.4 Our work

Taking into account the complexity and large number of problems, we use the idea of classification to summarize the problems, and then package the model to build.

- We apply Model I,**Complex Network Model(CNM)** in solution to problem1and 2^{nd} part of question 5.By defining the weight of music influence,a comprehensive subnetwork is established to reflect the mapping between musicians.And the following **Cluster K (elbow method)** shows the major changers under the network
- We apply Model II, **Inner Music Measurement Model(IMMM)** in solution to problem 2,2 thirds of problem 3 and the very first of problem 4,which mainly contains **Music Similarity Measurement Model(MSMM)**with **Principal Component Analysis(PCA),Euclidean Distance(WED)** and **Similarity Measurement(SM)**to compare genres,Random Forest(RF) to distinguish genres and **Co-relation Model(CM)** with Pearson Coefficient with to explore the components of music.
- We apply Model III **Logistic Regression(Time Series)**responding to the rest parts of problem 4/5/6/7.Giving insight into music trends and extend to the future

The more detailed models and according results are shown below.

2 General Assumptions and Model Overview

2.1 Assumptions

To simplify the problem,we make the following basic assumptions,each of which is properly justified

- **Assumptions 1:** the definition of "past music" refers to music previously produced by all music in history

  Justification: If only a single musician is considered, it will increase the difficulty of the work and have little effect on the results

- **Assumptions 2:** All factors are considered to be within the normal range. For example, the duration/-ms is based on the 2-10ms of the song market

  Justification: Increase the accuracy and authority of conclusion by defining the domain.For example, the feature of duration_ms and popularity show an obvious negative correlation, but if songs last less than 1 minute,there will be an obvious positive correlation.

- **Assumptions 3:** :Considering the phenomenon of cross-genre, the genre given in the table is the absolute definition of singer genre

  Justification:: Reduce unnecessary work, save time for the construction of important models.There are singers,like Taylor Swift who absorbs a lot of country music,debuts from the country music but eventually becomes a big hit as a popular singer. But the existence of cross-genre is relatively rare and more or less this phenomenon happens on every singe

- **Assumptions 4:** :Other factors which are not considered in this question will not fundamentally influence analysis results.

  Justification:: Sometimes music with excellent quantitative indicators still cannot be sold, called"smashing the brand". Only if such phenomena is considered rare in such data sets can the significance of research conclusions be guaranteed

2.2 Model Review

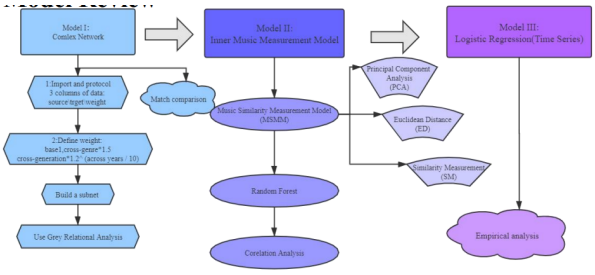


Figure 2: 3 model co-relationship revie

As shown in the figure above, the 3 models we have established are 3 ways to solve problems according to the logical order and analysis depth of the topic:

- model I is model dominated, based on a complete and complex network, and uses the endogenous advantages of the model to explain all problems
- model II is question dominated,combining multiple models,principal component analysis(PCA), Euclidean distance(ED), and decision-making trees to accurately promote process-based problem solving
- model III is analysis dominated,using innovative analysis to fully excavate existing data given by the topic

3 Model Preparation

3.1 Notations

Important notations used in model II are listed in Table.1

Table 1: Symbols and Indicators in model II (MSMM

Symbol	Indicator
a_1	danceability
a_2	energy
a_3	valence
a_4	tempo
a_5	loudness
a_6	mode
a_7	key
a_8	acousticness
a_9	instrumentalness
a_{10}	liveness
a_{11}	speechiness
a_{12}	duration_ms
a_{13}	popularity
a_{14}	explicit

3.2 The Data

3.2.1 Pre-processing

- **Data Collection&Data Cleaning**

  Before data analysis,the availability of data must be guaranteed.Below is the source

Table 2: Data Source collation

Database Names	Database Websites	Data Type
data_by_artist	www.allmusic.com/	Official database
Data_by_year	https://developer.spotify.com/	Commercial website statistics
full_music_data	https://developer.spotify.com/	Commercial website statistics
influence_data	https://developer.spotify.com/	Commercial website statistic

On the basis that there are no missing data this time, and we cut the data differently when analyzing the specific problem, removing noises is not processed uniformly.

- **Data Classification&Add Attribut**

  We not only perform labeled processing on scattered and repeated data in table of data_by_artist, full_influence_data, and full_influence,but also integrate multiple lines including influencer_main_genre,follower_main_genre, etc., to create a total of 20 new data sets of music genres, and name new data sets based on them variable.

3.2.2 Description and Visual Analysis

- There are 98,340 data pieces in the full_music_data.csv and influence_data.csv with combination of different feature variable values under overlapping ID.

Table 3: full_music_data.statistics

Descriptive Statistics	Average	Max	Min	SE	CV(SE /Average)
danceability	0.5263204	0.985	0	0.1640321	0.3116583
energy	0.5342478	1	0	0.2644774	0.4950463
valence	0.5332812	1	0	0.2585376	0.4848054
tempo	118.96853	222.605	0	29.925255	0.2515393
loudness	-10.75594	-0.866	-42.238	5.0676128	0.4711455
key	5.1871771	11	0	3.5067146	0.6760353
acousticness	0.4190569	0.996	0	0.3537099	0.8440617
liveness	0.2070086	1	0	0.1861034	0.8990129
speechiness	0.0638058	0.964	0	0.0761038	1.1927411
duration_ms	238590.92	1415707	11493	108823.18	0.4561078
popularity	35.693329	100	0	17.262527	0.4836345
year	1981.04	2020	1921	19.594702	0.0098911
instrumentalness	0.1255745	0.999	0	0.2724918	2.169961

The reason we choose CV is that compared with purely based on standard deviation, CV makes the degree of dispersion between data sets with different data values more comparable.For instance,year shows great stability,liveness divers,which is in

line with the background.The data collected mainly focuses on the 1950s-1990s (the 1960s stands out).And liveness is a symbol of genre,which we will refer later.

- There are 42770 data pieces in the influence_data.csv,with the overall characteristic of pop music dominated, since Pop /Rock accounted for 56.4%.The total list is showed in the following table.

Table 4: Symbols and Indicators in model II (MSMM

Avant-Garde	Blues
Classical	Comedy/Spoken
Country	Electronic
Folk	International
Jazz	Latin
Pop/Rock	R&B
Reggae	Reggae
Stage & Screen	Vocal
Childrens	Easy listening

- There are 100 pieces of data in the data_by_year.csv,showing feature variables with internal correlations with no repetition of the year as the main key.It is the same with the 5854 pieces of data in

4 Model I Complex Network

4.1 Details about Model

It mainly focus on the question 1(music_influence) and 5(main changer),uses influence_data set to build Directed network graph and the foundation for the following problems

Definition of Network:a network with some or all of the properties of self-organization, self-similarity, small world, and scale-free is called a complex network1

The detail can be described by following 6 characteristics (1-6) and simple v.s. advanced edition in *Fig. 3*

1. The number of nodes is huge, and the network structure presents many different characteristics
2. The main manifestation of network evolution is the generation and disappearance of nodes or connections
3. Connection diversity: The connection weights between nodes are different, and there may be directionality
4. Kinetic complexity

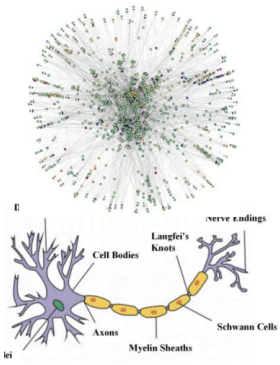


Figure 3:

5. Node diversity
6. Fusion of multiple complexity

4.2 Process and Results

Question 1:

Use the following steps to build a knowledge graph

- Import 3 columns of data: source/target/weight
- Protocol data, construct source/target/weight data set
- Define weight (the influence of musicians is inseparable from genre and time): base1,cross-genre*1.5,cross-generation*1.2^ (across years / 10).The detailed example are shown below

Table 5: Specific weight value of some example genres

Influencer	music_influence
The Beatles	821.25
Bob Dylan	542.89
Hank Williams	444.01
The Rolling Stones	417.18
David Bowie	366.47

- Build a sub net and use Grey Relational Analysis

Then,a directed complex network of **musical influence (Fig. 5)** was drawn below ,where each node (artist) was connected by edge(s) (influencer -> "influence" -> follower) given by influence_data (Supplementary information Section. 1). Artists who have similar influencers are closer.

Several conclusions can drawn from the following figure,**which is about the *music_influence* we need to analyse**

1. Apparently, each genre conquers a part (cluster, group) of the figure. This means that influencers and followers are from the same genre, in most situation.
2. The border of pop and rock music can be seen in the figure, because upper right part of the "Pop/Rock" part (light blue colored) includes most of important rock bands (Metallica, Led Zeppelin, etc.).
3. The Beatles is the dominating influential artist (group) in the music history. In addition, the Beatles is at the center of pop music using conclusion 3.
4. The most influential artists in their genres are always the pioneers and the founder in their genres, such as Marvin Gaye in R&B, Kraftwerk in electronic, Billie Holiday in vocal.
5. Several pop artists influence other genres, especially Bob Dylan's to country music and Brian Eno to electronic music. ("Father of Ambient Music", Brian Eno is almost "buried" in the electronic music group in the network.)
6. Connections of the genre can be seen in the figure:
   (a) Pop music is developed from and effected by most of other genres (country, electronic, R&B, jazz)
   (b) Latin is connected with reggae and jazz. (known as reggae en Español and latin jazz, and later reggaetón)
   (c) The close connection of vocal and R&B confirms that traditional R&B was developed from vocal

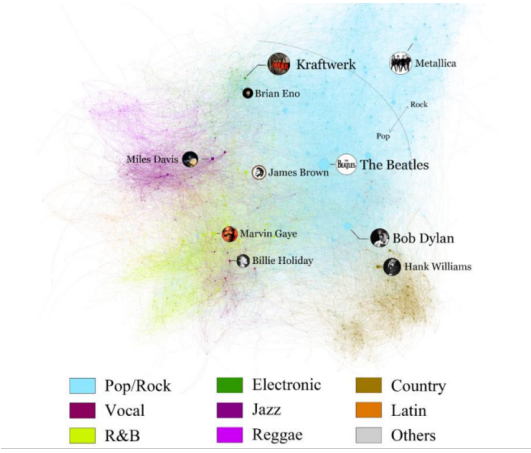


Figure 4: Directed complex network of musical influence

Supplement information: 1. Complex network (Fig. 5) The complex network was drawn in Gephi 0.9.2, using built-in Force Atlas algorithm. All artist icons were fetched from Twitter.

Question 5:: Clustering K (elbow method)

Solution: Compare the differences and similarities of genre styles between A->B,B->C,A->D,A->E in the music_influence network to get the ultimate changer.If the style of A is similar to the genre of D and E musicians, but different from the genre of B musicians, it can indicate that A is a major changer,which is also called elbow method

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{1}$$

- C_i:i-th cluster, p:sample point in Ci, mi=the centroid of

  C_i (mean value), SSE is the clustering error of all samples, representing the quality of the clustering effect.

Model principle:

1. The number of clusters k↑ , the sample division will be more refined, the degree of aggregation in each cluster↑ ,the SSE(Eq.1)↓
2. If $k \leq true$ number of clusters, $k \uparrow$ the degree of aggregation of each cluster greatly ↓ SSE ↓greatly
3. When $k = true$ number of clusters, k↑ the degree of aggregation quickly and the decline ↓of SSE sharply, then level off f as k ↑continues
4. SSE-k ↑is in the shape of an elbow(Fig.5), k value corresponding to elbow= true number of clusters of the data

Step by this:

1. Process the data set: Use excel mining and analysis of the influence_data.csv data, select follower_genre=influence_genre data, and invert the remaining data, get 23% of the overall data, named changer_data.csv
2. Use the elbow method to select the optimal cluster number k for the data in changer_data.csv: Let k start from 1 until the appropriate upper limit is reached.Cluster each value of k and note down the SSE, then draw the relationship between k and SSE), finally select the k corresponding to the elbow as our best Number of clusters
3. Test result: screen on the data in changer_data.csv again, select follower_genre and main influencer_genre with the most different numbers of inflencers, the final result is 95% close to the game changer after K-type aggregation.

Results:

- Final changers are mainly Hank Williams/Muddy Waters/Kraftwerk/Miles Davis/-James Brown/Howlin Wolf/Billie Holiday/Marvin Gaye/Ray Charles/Bob Dylan/The Beatles/Johnny .
- Take the Beatles as an example. It was established in Liverpool, England in 1960. Its music style is derived from the rock music of the 1950s, and it has developed psychedelic rock, pop rock and other genres. So it is a well-deserved changer

5 Model II Inner-Music Measurement Model

It mainly focues on question 2/3 and is composed of Music Similarity Measurement Mode(MSMM),Random Forest and Corelation Model with Pearson Coefficient

5.1 MSMM Construction

5.1.1 Principal Component Analysis(PCA

Index Dimensionality Reduction:

- Use the indicators in full_music_data and data_by_artists to measure musical similarity. which are the characteristics of music. Considering that the dimension of the indicators is too high, we use **PCA** and remove the effects of

The main steps are as follows:Suppose there are n samples and p indicators, then a sample matrix of size n*p can be formed as Eq.(2)

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (x_1, x_2, \cdots, x_p) \tag{2}$$

Step1: Standardize the indicator data by zero-mean normalization. Calculate the mean $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and standard deviation$s_j = \sqrt{\frac{(x_{ij}-\bar{x}_{ij})^2}{n-1}}$ by column,calculate the standardized data . The original sample matrix is standardized to Eq.(3)

$$x = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = (X_1, X_2, \cdots, X_p) \tag{3}$$

Step2: to calculate the covariance matrix of standardized samples.

Step3: to calculate the eigenvalue and the eigenvector of R:the eigenvalue$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$:; the eigenvector:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \quad r_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j) = \frac{1}{n-1}\sum_{k=1}^{n} X_{ki}X_{kj}$$

$$a_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}, a_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{p2} \end{bmatrix}, \cdots, a_p = \begin{bmatrix} a_{1p} \\ a_{2p} \\ \vdots \\ a_{pp} \end{bmatrix}$$

Step4: calculate principal component contribution rate and cumulative contribution rate the component contribution= $\frac{\lambda_i}{\sum_{k=1}^{i}\lambda_K}(I = 1, 2, \cdots, p)$

Step5: to select and express principal components. Take the first, second, ..., m-th ($m \leq p$)principal components corresponding to the eigenvalues whose cumulative contribution rate exceeds 85%, and the i-th principal component:

$$F_i = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{pi}X_p \ (i = 1, 2, \cdots, m)$$

5.1.2 Similarity Measurement based on Euclidean Distance

Considering the standardized data is dense, choose distance to calculate the similarity. Here, use Euclidean Distance(ED) of different samples to measure music similarity. The ED principle are shown in Eq.(3)

$$d(s_1, s_2) = \sqrt{\sum_{n=1}^{m}(s_1 - s_2)^2}$$

- The two samples s_1 and s_2 are m*1 dimensional vectors, storing the values of principal components
- The greater the $d(s_1, s_2)$ value,the longer the ED between the two music samples, which reflects the smaller the similarity between them, so we decide to convert the reverse $d(s_1, s_2)$ index to a positive indicator, which is recorded as $SM(s_1, s_2)$ to indicate the similarity. The forward processing method is shown in Eq.(4)

$$SM(s_1, s_2) = \frac{\max - d(s_1, s_2)}{\max - \min} \tag{4}$$

5.1.3 MSMM Principle

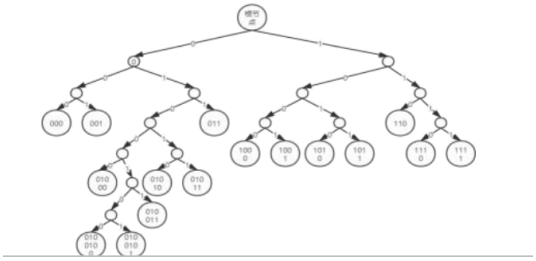
- Here, the object of our music similarity study is an abstract concept. In fact, different sample matrices under the same index can have different practical meanings.

Specifically, based on our model, if we extract complete indicators data for the sample matrices, we could study the similarity of two music works, two artists, and two years, and even two genres.

- Here, we sample in pairs randomly from full_music_data, and finally 250,000 samples are extracted. We use the method above to calculate the value of d(a,b) within the samples. We find that the maximum value of the 250,000 d(a,b) is approximately 11.0019, and the min is approximately 0.1758. Taking into account sampling error, we increase max to 11.5000, and reduce min to 0.0000.
- Finally, we use SM(a,b) to measure the similarity between music. The greater the SM(a,b) value of two pieces of music, the greater their similarity, and furtherly, the value of SM(a,b) generally varies between 0 and 1, so, when SM(a,b) value is closer to 1, the music similarity is greater. Besides, if SM(a,b) value is negative, it means that there is very little similarity between the two piece of music

5.2 Random Forest Construction

Develop a random forest which is shown in *Fig.6* below as a classifier model using *data_by_artist*, if accuracy of the classfier is high(finally reaches 89.8%), the characteristic of each genre will be clearly explained The model was built using following methods:



1. 12 columns of the original dataset (['danceability', 'energy', 'valence'etc.]) were chosen as independent variables, and ['genre'] is thedependent variable
2. Use -Fold method (in this situation), which split the dataset in two parts (9/10 for training, 1/10 for testing) 10 times.
3. More precise parameters: number of decision trees: 100; max depth of each tree: 30; min samples of each leaf: 3

Results: The accuracy is up to 70.2% (63.1% mean), compared with only a 1/19 (5.3%) mean accuracy when classifying randomly.Table.6 below shows genre's performance index on musical characteristics.1, 2, 3 represent the most prominent, -1, -2, -3represent the least prominent.

	danceability	energy	valence	loudness	key	acousticness	liveness	speechiness	popularity
Classical		-1	-1	-1		1			18
Country					1				8
Childrens	3		1						12
Comedy/Spoken							1	1	15
Electronic		1		1		-1			1
Reggae	1	2	3	4		-2			4

- Classical music is outstanding in acousticness; Country is outstanding in key; children is outstanding in valence; comedy/spoken is outstanding in livenss and speechiness; electronic is outstanding in energy, loudness/popularity; Reggae is in danceabilitys Outstanding performance; new age has outstanding performance on instrumentalness and duration_ms

5.3 Correlation Model Construction

Our solution: Analyze some typical musicians in the full_music_data data set,use correlation analysis to test the Pearson correlation coefficient(PCC) and p-value,exploring the relationship between various indicators and popularity. The higher the correlation, the more Contagious.

Data selection: select musicians whose appearances were more than 100,like Wolfgang Amadeus Mozart, Billie Holiday.,who has sufficient data size(874 times) and proportion in the subset table.And themselves are influential and representative .

this process can reduce the complexity of training and also reduce the occurrence of over-fitting.The results are as follows

Results shows:

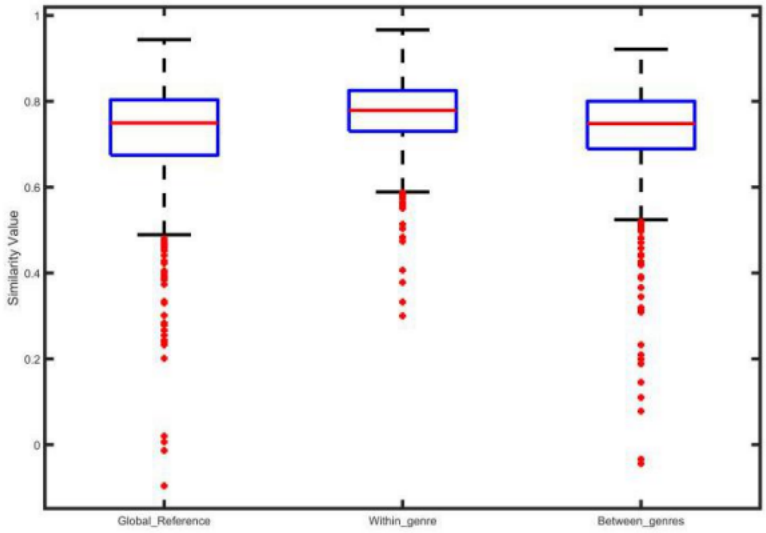
1. There is no correlation between popularity and danceability, valence, tempo, key, liveness, and explict
2. There is a significant positive correlation between popularity and energy, loudness, mode, and year
3. There is a significant positive correlation between popularity and acousticness, duration_ms, speechiness, and instrumentalness Significant negative correlation

5.4 IMMM Application

5.4.1 Artists within genre more similar than those who not

1^{st} , we take 'artist' as the sample object and use the data in data_by_artist to construct a sample matrix to determine the principal components. The principal component expression is shown in Eq.(5):

3^{rd}, use the *influence_data* to classify artists according to genres, and then we take the RS method to take out 500 sets of experimental data from each genre and between genres to calculate their similarity. Then, we make a comparison among the three groups of data. The results are shown in *Fig.7*:



Finally, draw our conclusion from the Fig.7:

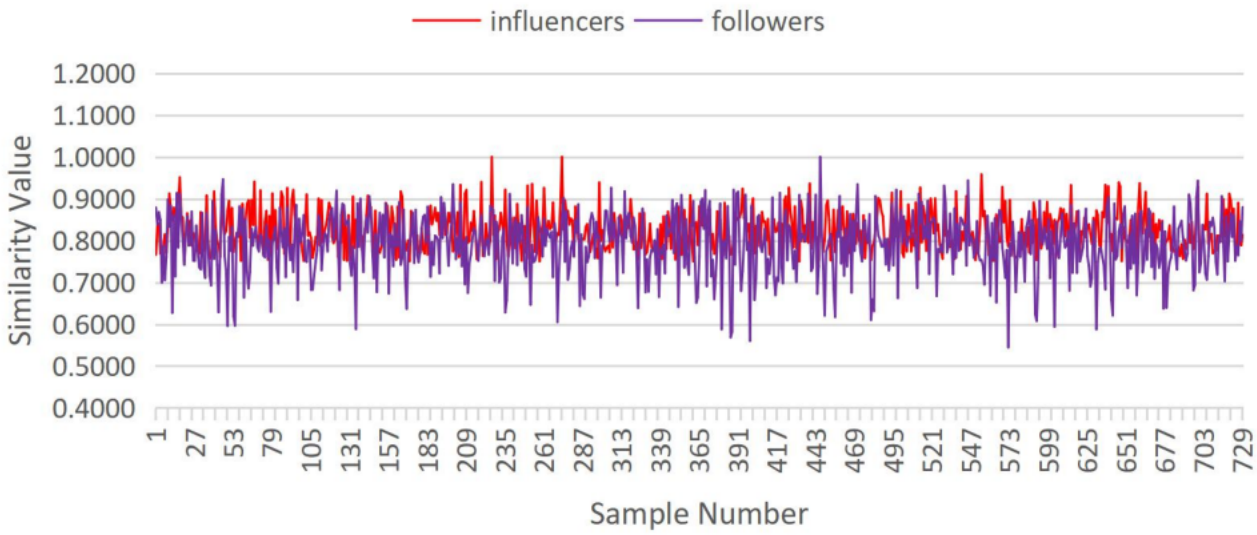
- Artists within genre are more similar than artists between genres. As reflected in Fig.7, the median of the similarity value, which means the average similarity between artists within genre are much higher than between genres. For data distribution, the similarity between artists within genre are centralized higher, and there are many extreme values with low similarity between genres.
- Besides, as global reference data represents the average within and between genres, we think that the box diagram of it should be located in the middle of within genre and between genres. Our predictions are the same as the experimental results, so this can also verify the rationality of the model.

5.4.2 Identified influencers in fact influence the respective artists?

we still take **artist similarity** as the sample object. Therefore, the result of the PA selection is the same as the previous question. Here, our goal is to study when the influencers are identified, if their followers are all similar.

To quantity this question, we select a pair of influencers whose artist similarity value is greater than 0.75, which means although the influencers are not the same, but are very similar, then we select one follower from this pair of influencers to form a matching pair as object of comparison. The results are shown in *Fig.8*

1. The similarity between followers matched with influencers is less than the similarity between influencer.



2. Under this condition, the similarity between followers has its instability, which proves different artists are influenced respectively by the same artists. So, we think that the identified influencers do in fact influence the respective artists.

5.4.3 'Influencers actually affect the music created by the followers?'

For this part, we use the full_music_data to study the similarity of music works, our goal is to compare the similarity of the music works created by an influencer's followers, with by an influencer's followers

1^{st} , we take 'song_title (censored)' as the sample object and use the data in *data_by_artist* to construct a sample matrix to determine the principal component, the results of PRC is shown as follows:

Table 6: Results of Principal component analysis

No./Symbol	Characteristic root	Variance contribution rate	Comulative variance contribution rate
a_1	-0.11	24.76%	24.76%
a_2	0.06	10.97%	35.73%
a_3	0.06	8.90%	4.63%
a_4	-0.04	8.00%	52.63%
a_5	0.19	7.46%	60.09%
a_6	-0.01	6.62%	66.71%
a_7	0.02	6.25%	72.96%
a_8	-0.05	5.53%	78.49%
a_9	-0.41	5.32%	83.81%
a_{10}	-0.18	4.94%	88.74%
a_{11}	-0.35	4.42%	93.16%
a_{12}	0.56	2.49%	95.65%
a_{13}	0.31	2.20%	97.84%
a_{14}	-0.35	1.36%	99.20%
a_{15}	-0.30	0.80%	100%

The eigenvector matrix corresponding to eigenvalues is:

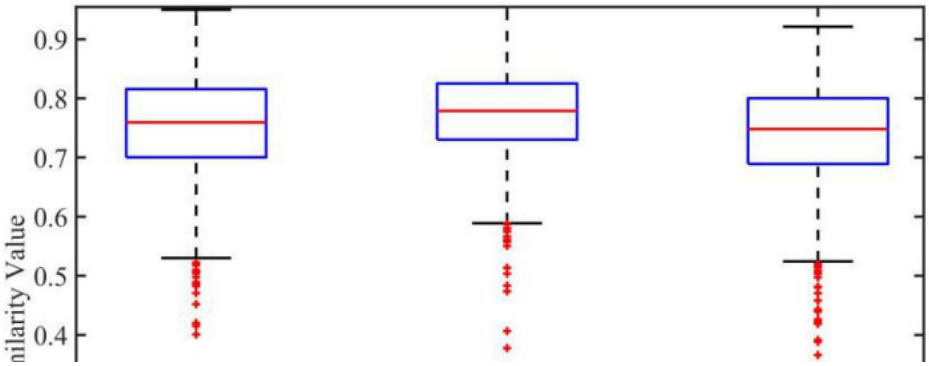
[0.19 0.50 0.21 0.33 0.01 0.28 0.03 0.14 0.16 -0.11;
0.44 0.00 -0.21 -0.15 0.16 0.07 0.02 0.05 -0.26 0.06;
0.19 0.60 -0.06 0.00 0.19 0.16 0.13 0.13 -0.03 0.00;
0.14 0.00 -0.25 -0.38 0.17 -0.39 0.52 0.00 0.45 -0.04;
0.43 0.03 -0.08 -0.14 0.03 -0.02 -0.09 -0.08 -0.18 0.19;
-0.02 0.11 0.00 -0.45 -0.53 -0.01 -0.20 0.68 0.07 -0.01;
0.02 -0.01 -0.03 0.30 0.47 -0.57 -0.43 0.43 0.04 0.02;
-0.42 0.09 0.01 0.12 -0.14 -0.09 -0.01 -0.07 0.14 -0.05;
-0.22 -0.22 0.00 0.01 0.27 0.18 0.42 0.42 -0.47 -0.41;
0.03 -0.10 -0.63 -0.03 -0.05 0.20 -0.39 -0.11 -0.14 -0.18;
0.07 0.01 -0.52 0.43 -0.23 0.02 0.13 0.08 0.30 -0.35;
0.14 -0.18 -0.11 0.45 -0.35 -0.12 0.34 0.21 -0.20 0.56;
0.01 -0.38 -0.01 0.01 0.32 0.56 -0.06 0.26 0.50 0.31;
0.36 -0.24 0.32 0.08 -0.14 -0.05 -0.05 -0.02 0.12 -0.35;
0.39 -0.28 0.25 0.09 -0.14 -0.04 -0.03 0.01 0.06 -0.30]

Before we start to finish the question2, there is some information can be disclosed here. On one hand, the contribution rates of these principal components decrease sequen-

tially, on the other hand, in the principal component expression, the greater the absolute value of the correlation coefficient before each indicator, the greater the correlation between the principal component and the indicator. Therefore, we can obtaining indicators a1,a2,a5,a8,a10,a11 are more decisive in explaining the similarity between musical works. In other words, these indicators can better distinguish between different musical works.

2^{nd}, we take **Random Sampling(RS)** method to take out 5000 sets of musical work sample from full_music_data, so as to understand the overall similarity level. Especially, the max of the data is 14.274 and the min 0.826. Taking sampling error into account, we increase max to 15.000, and reduce min to 0.0000 .

3^{rd}, we compare the similarity of music works by selecting three sets of samples. The three sets of samples are: random sampling of the overall data, and the similarity data between the works of musicians with and without influencing relationships. 1000 sets of data are selected in each, and the box plot is as *Fig.9*



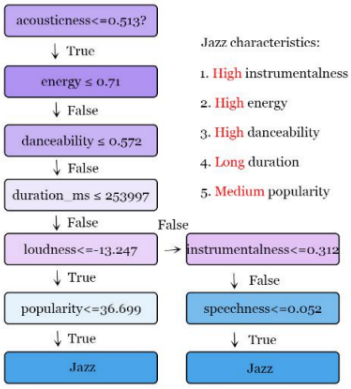
We can draw the conclusion that the influencers actually affect the music created by the followers. The conclusion is similar to the conclusion of the question 2. For the reason, we have talked of that the similarity of music is an abstract indicator. If two samples have the indicators data, no matter they are musical works, artists, or even the genre. Their similarity could be calculate using our model.

5.4.4 What distinguishes a genre?Are some genres related to others?Are some music characteristics more contagious than others?

Result I: Characteristic of all 19 genres can all be disclosed from the random forest. Fig.10 only shows a few branches of one tree from the random forest

We take 'Jazz' as an example

In music history, jazz changes rapidly as time grows. Since the 1920s, jazz has been recognized as a major form of musical expression in traditional and popular music. [1] As jazz spread around the world, it drew on national, regional, and local musical cultures, which gave rise to different styles. Bebop and later hard bop emerged in the 1940s and 50s, their fast tempo and rapid chord changes shifting jazz to be danceable music. And 1980s' successful smooth jazz gains a lot of popularity.



The history of jazz tells us a lot about its high instrumentalness, high energy and high danceability, which is confirmed in the random forest model.[1]

Correlation anaysis between music characteristics and/with popularity

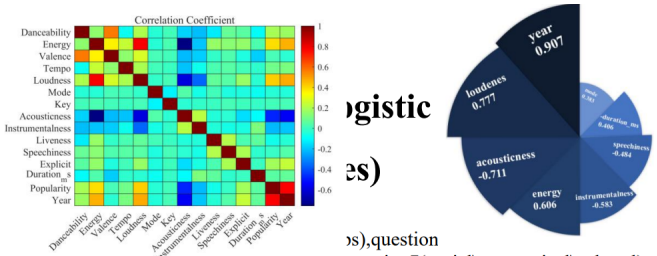
Result II:Obviously, different music indicators play different roles in infectiousness

- For example,explicit have a correlation coefficient of zero when the p value is 1,while year have a correlation coefficient is as high as 0.907.In order to comparethe differences in the contributions by different musical characteristics moreclearly, we made a Nightingale rose chart, as shown on the Fig.11.2.

In order to answer more concisely about the roles played in influencing a particular artists or music,we need to combine another question which focuses on the inner connection

Result III: Fig.11.1-heat map shows inner co-relationship fig.10.

- From which,we find that the strongest relationship lies between energy and loudness,then accousticness and loudness,valence and danceability,year and loudness,accousticness and popularity(from high to low)



ɔgistic

ɛs)

ɔs),question

question7(social\economical\cultural)

[1][1] Hennessey, Thomas (1973). From Jazz to Swing: Black Jazz Musicians and Their Music, 1917–1935 (Ph.D. dissertation). Northwestern University. pp. 470–473.

6 Model III Logistic Regression(Time Series)

6.1 Empirical analysis

Solution I:Regarding the possible revolutionary (major leap) features in the evolution of music, we drew a distribution map of each feature of music over time according to time sequence.

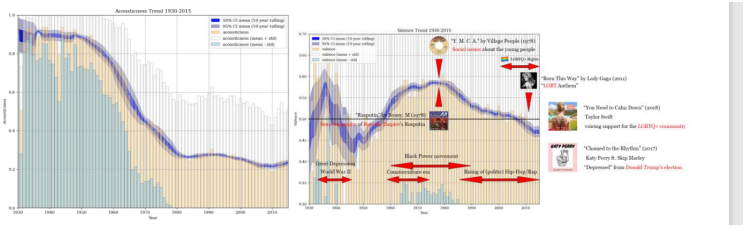
Results I:

- In response to major changes, we have observed qualitative changes in characteristic values such as popularity, accousticness, explicit, instrumentness, valence, etc.
- Characteristics such as energy\loudness\slowly develop when time goes by
- Characteristics such as liveness, speechiness\tempo even seems like no big change over time
- popularity, accousticness, explicit, instrumentness, valence as the characteristics that signify revolutions.Among which popularity is one of the most significant indicator that will show the enthusiasm of creators,ups and downs of musical genres.Besides,the conclusion can also be proved by our findings under the social,political,technologic changes over time.We show them below.

Solution II:(social\economical\cultural)Select the peak moment with sudden change in growth rate for analysis, including abrupt increase and abrupt slowdown (since the size of music characteristics, showing all isn't realistic

Result II:

- **Significant changes in accousticness caused by technological iteratio**



From 1930 to 2010, it has shown a significant decline over time, and its credibility has dropped from 1.0 to only 1/5 of about 0.2; The trend of slow first and then fast declined the fastest from 1950 to 1980. (*Fig.12 above*)

The reason behind it is obvious: technological iteration and progress. After entering the 1950s, the rapid development of computer technology and the continuous improvement of the level of electronic music and related production companies have led to the use of technology to process sound and artificially amplify the sound more and more common. In today's electronic information age, completely real-sound records have been It's

so rare, I don't think it's a rare thing. The abrupt slowdown after 1980 is the result of the combined effect of the popularity of the retro wave and the development of technology to a certain stage.

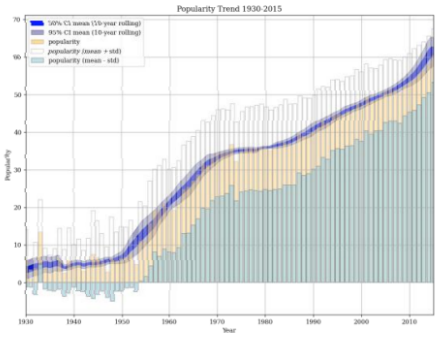
- **The rise of rock music led to the confrontation between music and vocals and the final introduction of music (instrumentalness)**

As shown in the figure below, from the turbulent 1930s to the 1950s, the value of instrumentalness fluctuated significantly (from the initial higher value). It oscillates to a state below 0.5 through drastic changes), which actually shows that interjections such as "oh" and "ah" have gradually become accepted by people through the rise and change of rock music.

As more policy like European Convention on Human Rights were published in 20 th century,political atmosphere leads to ideological progress.At the same time, introducing more instrumental melodies such as violin, piano and even suona into songs has become the norm after the rock and roll revolution. This trend can also be reflected by the drastic changes in valence from the 1940s to the 1950s.

- **People's increasing attention to entertainment life has led to an increase in the overall popularity of music**

As shown below, the growth rate suddenly increased from the 1950s to the 1970s, and the popularity value suddenly changed from a single digit to the median level of today's popularity value. One is because of the period when the official European genre was renewed and musical styles and creations flourished; the second was because of the rapid increase in productivity in Europe and America (the British Industrial Revolution), and ordinary people were one step closer to the pursuit of entertainment and spiritual life; and the third was technological progress. In a sense, the production of music has been increased, and commercial operations have continued to stimulate the audience, resulting in better and better paying.



6.2 How genres change over time?

Solution III:(dynamic influencers)Link two tables of influence_data and full_music_data through My SQL,create a foreign key through artist_name/artist_id, organize these two into one, named genre_by_year,then draw the distribution diagram of genre over time accordingly.

Result III:

Pop/Rock: The Beatles bring the "spring" of pop music into music industry, as the most influential artist (group) in the world and music history, they active from 1960 to 1970, leading to the fact that the share of pop music less than 20% in 1960 to over 70% in 1970.

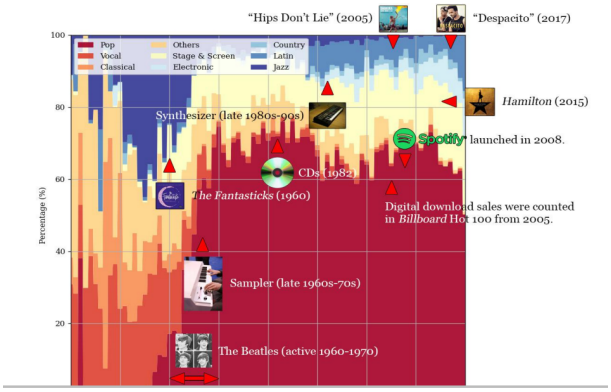
In late 1960s, samplers were invented. Samplers give producers opportunity to reproduce music based on old music. With sampler, the share of pop music increases rapidly. For example, when the famous producer Mark Ronson talks about the history of music, he is talking about sampler and one of the most sampled song, Doug E. Fresh and MC Ricky D's 1985 song "La Di Da Di". [1] [2]

In 1982, digital audio compact disc, later known as CD, was released. It was much cheapier and affordable than vinyl record. The invention of CD gives room for pop music to grow.

With the invention of the Internet, online music store like iTunes store emerged in early 2000s, it helped the popularity of the pop music continues.

Then on-demand streaming platform appear in the music industry, the most influential ones are YouTube (launched in 2005), Spotify (launched in 2008) and Apple Music (launched in 2015). On-demand streaming changes the industry that people can listen to their favorite music low or free of charge. People won't be dominated by pop songs' annoying hooks and turn to their favorite genre.

Recent years, we are glad to see many pop artists try to shift their pop career to other genres, such as Dua Lipa's 80s' nu-disco style, Grammy nominated hit "Don't Start Now" (2020); The Weeknd's 80s' synthwave style, Super Bowl LV performed hit "Blinding Lights" (2020). This phenomenon is a reason of the descending of pop/rock these years.



Stage & screen:

Stage music refers to musicals. The "Broadway prosperity" started in early 1950s, with the famous musicals of West Side Story (1957) and The Fantasticks (1960). Since then, musicals were played everywhere in the world, from Broadway, to Off-Broadway even Off-Off-Broadway. Even today, musicals are still popular, such as Lin-Manuel Miranda's Hamilton (2015). Screen music refers to film score and soundtracks. Film score and soundtracks are original music written specifically to accompany a film. With the popularity of film increasing, the popularity of screen music is still

Electronic:

The history of electronic music is hugely depending on the invention of synthesizer. With synthesizers, producers can change the voice of instruments, vocals, even voice of animals, using built-in filters, envelopes, and low-frequency oscillators. These characteristics of the synthesizer defines the evolution of electronic music.

Jazz:

As mentioned in the random forest model part, jazz changes rapidly as time grows. After the dominating era (1950s), it was divided and combined with other genres like pop, rock even electronic.

Latin:

Because of the population of hispanic and latino Americans, latin/latino music is also popular. In this streaming era, especially after the release and Justin Bieber's remix of "Despacito" (2017), latin music grows rapidly, resulting that "Despacito"'s music video is the most viewed music video on YouTube (nearly 7.2 billion, dated Feb. 8, 2021).3

7 Conclusion and further Discussion

7.1 Summary of Result

- Music _influence shows the cross-person relationship in the field of music.Artists within genre are more similar than artists between genres.The inner characteristics the genre owns distinguishes them but it may change as time goes by since there is definitely relationship between the characteristics.the identified influencers in fact influence the respective artists.The 'influencers' like Beatles(influencers of major Change)actually affect the music created by the followers.

- Some music characteristics more 'contagious' than others like the year music was produced Characters that might signify revolutions (major leaps) in musical evolution are popularity, acousticness etc which change greatly when times differ Different genres change in their own way,but show the same reaction to technology.And popularity is the most accessible indicator to know the trend

- The development of music is a complicated question which is under the influence of social, political or technological

7.2 Sensitivity Analysis

7.3 Strengths

Innovative model method, simple construction of the most effective model

- Initially, in Model II (MSMM), we considered the idea of combining variance filtering and principal components, and later adopted weighted Euclidean distance (rather than direct Euclidean distance), directly removing the variance ratio small indicators, this approach does not take into account that each indicator must contribute a part of the value when measured by distance. Only using principal component analysis can achieve the effect of screening indicators to only 8 remaining, which serves the purpose of contribution differentiation, increases accuracy, reduces model iteration complexity, and kills two birds with one stone.

Complete, logical, and truthful

- Data visualization technology is applied to interpret the original data,and the results are presented intuitively and concisely

Have a wide range of applicability

- Through the analysis ,the influence in the field of music can be predicted clearly,which is helpful for the mordern development of music

7.4 Possible Improvement

Some detailed work need to refine better

7.5 Future development enlightenment

Distribute attention appropriately:

- There is a big difference between music characteristics and popularity, which inspires artists when creating music. They should focus on more prominent features, not only for utilitarian considerations, but also because some musical characteristics are closer to the essential characteristics of music creation. For example, the performance of musical instruments can indeed increase the level of music, and develop emotions and intelligence when inspiring resonance with the audience, and harvest growth.

The road to quantify music:

- Under the current trend of increasingly close integration of music technology, Quantitative music is no longer limited to the traditional teaching of music theory through notes, rests, and singing. It is hoped that more interdisciplinary applications will be involved to move towards multiple experiences and creations

The datasets do not cover all of the data of music industry, if more data is given, we will continue our study both macroscopically (for the music industry and culture) and microscopically (for an artist).

Macroscopically (for the music industry and culture):

Some important genres are missing in the two datasets, especially genres which will have further effect on culture (Hip-Hop/Rap, K-Pop (subgenre of "International"), Electronic) If more data about missing genres is given, we will do these stuff:

1. Hip-Hop/Rap is also based in black community like R&B, and Hip-Hop/Rap music will contribute lots of speechness and explicit index
2. International subgenres like K-Pop and J-Pop are not based in the United States, but they do a huge effect and shock to the music industry of the US, especially Psy's "Gangnam Style", BTS and Blackpink. These data will help more about forecasting music's
3. Electronic tracks mainly refer to eurodance subgenre here, although eurodance makes little difference to the music culture of America, yet it did a lot to affect European culture (for example, O-zone's "Dragostea Din Tei", Alexandra Stan's "Mr. Saxobeat"), we would like to use more data to explore the music's effect on global culture.

The *full_music_data* dataset was from Spotify API, which means that the popularity of the tracks is limited in the streaming platform and era (Spotify was launched in 2008).

1. We cannot get the precise popularity data only depending on streaming data. Nowadays, music charts like Billboard consider three parts of songs as total popularity: **streaming, digital download** and **radio audience**. While considering **all three components**, we will get more precise popularity.
2. Pure sales of songs (physical sales, vinyl sales, digital sales), without streaming data should be given as popularity. If these data is given, we will dive more into and compare the difference of different eras of music, which are album era [1], digital download era [2] and streaming era.

Microscopically (for an artist):

If more data given, we would like to develop a model, which can tell us a story about the music style transition of an artist, taking Rihanna as an example:

Rihanna's debut studio album *Music of the Sun* (2005) mainly focuses on Reggae, her fifth album *Loud* (2010) turns to dance-pop, but her latest album *ANTI* (2018) tries R&B and hip hop soul. This can be reflected on the plot of track's energy against Rihanna's studio album timeline (discography) (Fig. ?)

Not only Rihanna, most artists have a transition like this (Taylor Swift shifting Country "Tim McGraw" (2005) to Pop/Rock "Shake It Off" (2014), back to Country "Cardigan" (2020)). Therefore, more data for a model telling this story about transition is needed.

MCM Team #2103070[4]

Feb.9th 2021

References

[1] Xiao Fan Liu,Chi K.Tse,Michael Small.Complex network structure of musical compositions: Algorithmic generation of appealing music[J]. Science Direct,2009.08-035(publish).

[2] Pedro Cano.The emergence of complex network patterns in music networks.Conference Paper[N].2004-2

[3] Liu Zhengwei. Comparing the two "leaps" of Chinese and western music in the 16th century[J]. Journal of Nanjing Academy of Arts (music and performance Edition), 2012 (04):

[4] Zhang Wenbo, Liu Qingying. The development of Chinese contemporary pop music from the perspective of communication [J]. Contemporary music, 2020 (04): 147-149

[5] Liu Danni. The theory and practice of darhaus' music history research [D]. Shanghai Conservatory of music, 2014

[6] Liu Zhenyin. Comparative study on the changes of modern music in three East Asian countries [D]. Shanghai Conservatory of music, 2007

[7] Luo Qiqing. Social communication of Internet plus era drama music [D]. Nanchang University, 2019

[8] He Rong. Music recommendation system based on convolutional neural network [D]. Nanjing University of Posts and telecommunications, 2019

[9] Tang Xiaowen. Music recommendation method based on audio features [D]. Liaoning University of science and technology, 2019

[10] Yao Jianing. Research on music genre classification based on deep attention mechanism [D]. Dalian University of technology, 2018

[11] Zhang Yurong. The influence of Arabic music on European music in the Middle Ages: a case study of the Crusade [D]. Northwest Normal University, 2018

[12] Li Jian. Research on personalized music recommendation system based on similar styles [D]. Huazhong University of science and technology, 2015

[13] Hu Bingjie. Research on music sentiment analysis based on feature vectors [D]. Xidian University, 2014.

Appendices

Appendix A Tools and software

Paper written and generated via LaTeX
Graph generated and calculation using Python3.7\SPSS\Powerpoint\Matlab
Data mining using MySQL\Excel

Appendix B The codes(Part of IMMM code)

Here are simulation programms we used in our model as follow

```
load full_music_data.mat % PCA
fullmusicdata = fullmusicdata(2:98341,:);
[n,p] = size(fullmusicdata);
%% NORMALIZE
X=zscore(fullmusicdata);
%% covariance matrix
R = cov(X);
%% eigenvalues and eigenvectors of R
[V,D] = eig(R);
%% contribution rate
lambda = diag(D);
lambda = lambda(end:-1:1);
contribution_rate = lambda / sum(lambda);
cum_contribution_rate = cumsum(lambda)/ sum(lambda);
disp('Characteristic value:')
disp(lambda')
disp('Contribution rate:')
disp(contribution_rate')
disp('Cumulative contribution rate:')
disp(cum_contribution_rate')
disp('The eigenvector matrix corresponding to the eigenvalue is:')
V=rot90(V)';
m =input('Please enter the number of principal components to be saved:');
F = zeros(n,m);
for i = 1:m
    ai = V(:,i)';
    Ai = repmat(ai,n,1);
    F(:, i) = sum(Ai .* X, 2);
End
j = 1;
W=[];
while j < 1001
    O = randsample(n,2,false);
    num1 = O(1,1);
    num2 = O(2,1);
    B = F(num1, :);
    U = F(num2, :);
    cj = sqrt((B-U)*(B-U)');
    W(j,1) = cj;
```

```
    j= j+1;
end
%% sampling
clear,clc
path = 'Sheet3.xlsx';
table = readcell(path);
ind = cell2mat(table(2:5603,4:11));
id = cell2mat(table(2:5603,1));
for i = 1:2000
    A = randsample(98340,2,false);
    num1 = A(1,1);
    num2 = A(2,1);
    D(num1,:) = [n1,m1];
    D(num2,:) = [n2,m2];
    A = randsample(5602,2,false);
    num1 = A(1,1);
    num2 = A(2,1);
    a = ind(num1,:);
    b = ind(num2,:);
    e = sqrt((a - b)*(a - b)');
    max = 17.000;
    min = 0.000;
    f = (max - e)/(max-min);
        for i = 1:2790
            if D(i,1) = n
                n1 = D(i,2);
            end
            if D(i,1) = m
                m1 = D(i,2);
            end
    end
        for i = 1:5603
            if id(i,1) = n1;
                s1 = ind(i,:);
        end
        if id(i,1) = m1;
            s2 = ind(i,:);
        end
    end
    e1 = sqrt((s1 - s2)*(s1 - s2)');
    f1 = (max - e1)/(max-min);
    Q = [];
    Q(i,1:2)=[f,f1];
end
clear,clc
load data_by_artist_match_picture.mat
X = databyartistmatchpicture1;
figure('units','normalized','Position',[0 0 0.8 0.8]);
data = X;

boxplot(data,'Labels',{'global_average', 'within_influence', 'without_influence'});
ylabel('Similarity Value','fontsize',12);
```