Data Project

Woo Min Kim

1. Introduction

The data was obtained from the UCI require, and it contains geographical information of traditional songs around the world by providing a longitude and a latitude of each song: 1059 pieces of music from 33 countries/areas exist. The audio features were also extracted by using MARSYAS (Tzanetaks and Cook, 1999) from wave files. MARSYAS generates a vector length 68 to estimate the performance with basic timbre information covering the entire length of each track.

The data is quite intersting because music, as a form of art, was mainly studied based on the subjective judgment. And this data-based approach on music is quite meaningful and desirable to seek the differences and similarities of different cultures around the world. This attempt is to find unique characteristics of traditional musics of each region, and it is really interesting to check whether the cultural differences can be expressed numerically. The data provides crudely approximated longitude and latitude of each music, and by solving regression problem we can expect to approximately predict the locations of the music origins.

Prior to the data analysis, we can select features by using kernel PCA and factor analysis in order for more concise model. After finding the most appropriate number of features, the data analysis results by raw and new data will be compared whether the dimension reduction makes sense and which feature selection methods are the best. With this data, we can focus on both classification and regression problems; as analysis tools, linear regression and gradient boosting machine will be used. For the classification problem, a new variable named region was introduced and it divided the data into 15 different Global Environment Outlook (GEO) subregions.

2. Exploratory Data Analysis

Prior to the data anlaysis, we need to explore the data first; especially, both dependent and independent variables. As can be seen in the figure 1, there exist relatively fewer observations on the map; there are 1059 observations, but there are only 33 unique pairs of latitude and longitude. According to the data source, the geographical location information was manually collected based on the description on the CD covers and gathering the exact locations was not physically possible; thereby rendering most of locations overlapping. This can be problematic because linear regression might not perform properly because the response variable might not have normal distribution. In order to check the normality assumption, residual plots will be investigated in the latter section. Furthurmore, instead of focusing on regression problem, we can try to solve the classification problem by solving multinomial model with multiple levels. In the following section, the extreme gradient boosting machine (XGBoost) will be introduced.

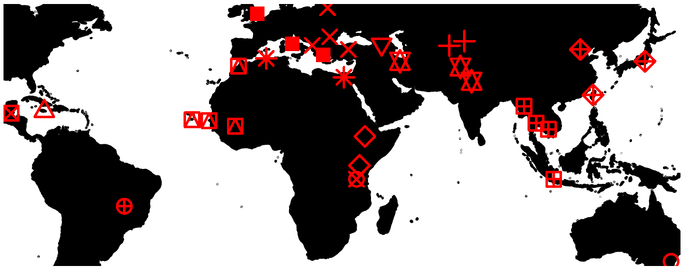

Figure 1: Geographical Location: different shapes infer different regions.

As explained previously, the audio features extracted from MARSYAS are vectors length of 68. All features are numerical and transformed to have a mean of 0 and standard deviation of 1. To check the relationship among variables, pairwise scatter plots can be investigated.

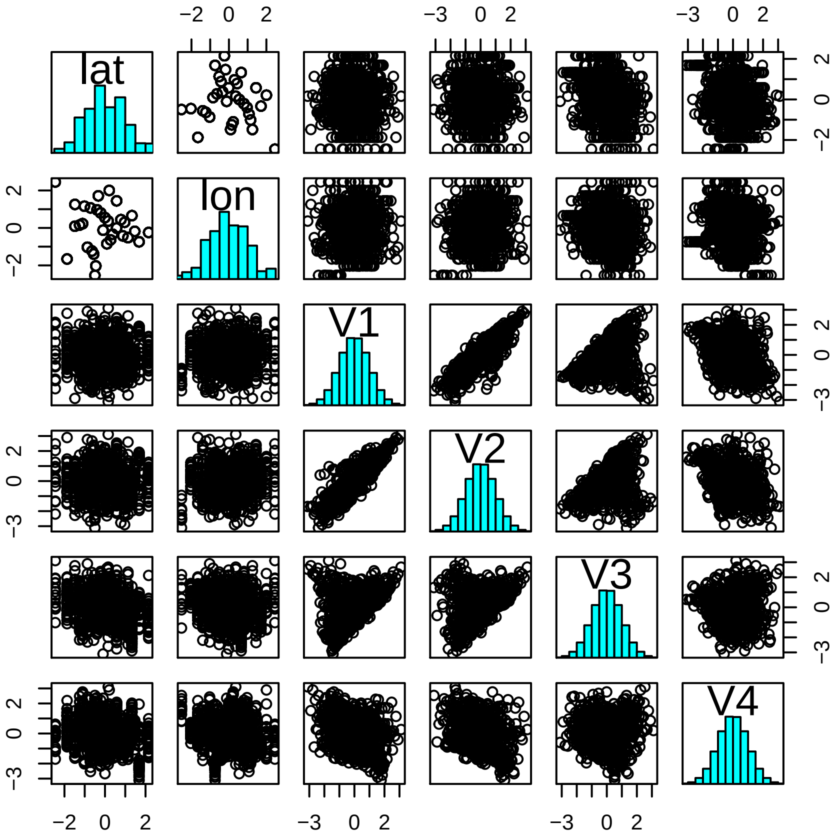

Figure 2: Pairwise Scatter Plots: these are the pairwise scatter plots among location variables and first three music features after Gaussian copula transformation. It is notable that music features seem uncorrelated to location variables. In addition, some of music features are highly correlated, and feature selection can also be considered.

In order for better correlation estimation between response and explanatory variables, I first applied Gaussian copula transformation on the data set. As can be seen in the pair scatter plots (Figure 2) and histograms above, it is hard to say that there exist correlation between response and explanatory variables. The largest absolute values of correlations between response and explanatory variables were computed and none of exmplanatory variable has linear relationship with the location variable. This implies non-linear relationship should also be seeked. This can be done by introducing new interaction terms or polynomial terms; however, kernel PCA can also be great option in this case.

Table 1: Three largest correlations between location and all music features

$\vert$cor$\vert$	0.35	0.31	0.29

Furthermore, in Figure 2, the first and second features are highly correlated and this implies a more concise model is possible by feature selection by various methods such as PCA and factor analysis. Detailed methods will be introduced in the next section.

3. Methods

By Occam's Razor, it is much attractive to do the analysis on the data with fewer variables, and there exist various feature selection methodologies. First, PCA and kernel PCA will be considered so that we can find the pricipal axis that can maximize the variance of the explanatory variables. Factor analysis can also be considered by introducing latent variables. Both methods seem very similar; however factor analysis does not focus on maximizing the variance of the data. So the results of these two methods will be different and the total explained variance with q sources will also not be the same.

3.1. PCA and Kernel PCA

First, in order to find the appropriate number of features, eigendecomposition on the covariance matrix is computed. Then, as in Figure 3, the cumulative variance explained by the number of principal components can be found. With about a half of features, 90% of the variance of the explanatory variables can be explained. In Figure 3, it is notable that the total explained variance increases very smoothly, and this indicates that the explanatory variables have small in-between correlations and the dimension reduction by the principal components analysis may not perform ideally. If there exist explanatory variables with strong discriminative power, a dramatic dimension reduction can be expected.

As discussed previously, the pairwise scatter plots (Figure 2) imply very weak linear relationship, and it is desirable to try find non-linear relationships between response and explanatory variables. One of the kernel options I tried was Gaussian kernel PCA. For the computation I used *kpca* function which is defined in *kernlab* require. As in PCA case, 30 Gaussian kernel PCs were collected.

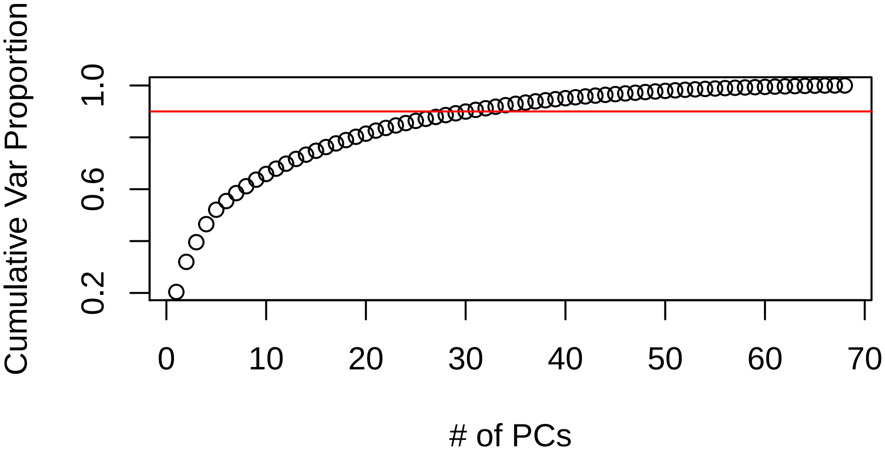

Figure 3: Proportion of Variance Explained by PCs. The red solid line is 0.9 implying 0.9 of variance of design matrix can be explained with around 30 PCs.

3.2. Factor Analysis

The feature selection can also be done by the factor analysis. By introducing latent variables for common and unique factors, we can rewrite the design matrix with those new latent features. The number of features can be determined by hypothesis testing based on the Bayesian Information Criteria (BIC). As in PCA case, Table 2 indicates that the appropriate number of features are 30.

Table 2: BIC comparison for different number of features

	q = 5	q = 10	q = 20	q = 30	q = 35
BIC	32641.75	23095.24	13841.21	11803.6	12033.27

The proportion of total variance explained by 30 features is $\frac{tr(A^TA)}{tr(A^TA+\Psi)} = 0.79$, where A and Ψ are factor loading matrix and unique variance, respectively. It is much less than that by PCA in the previous part. This is because factor analysis does not select features based on the variantion maximization. With only 79% of variation of the explanatory variables, it might be hard to obtain the robust result as close to as the result by using all 68 variables. The results will be compared in section 4.

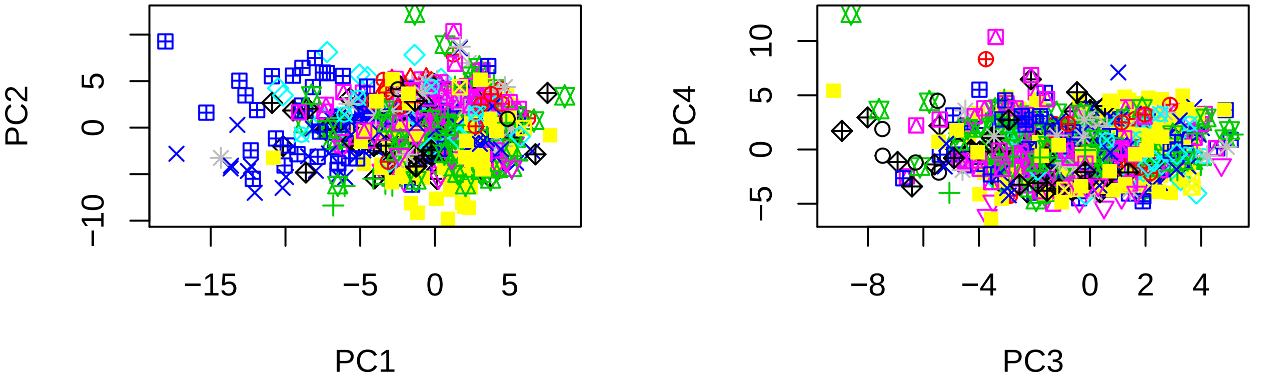

Figure 4: Scatter Plots of PCs

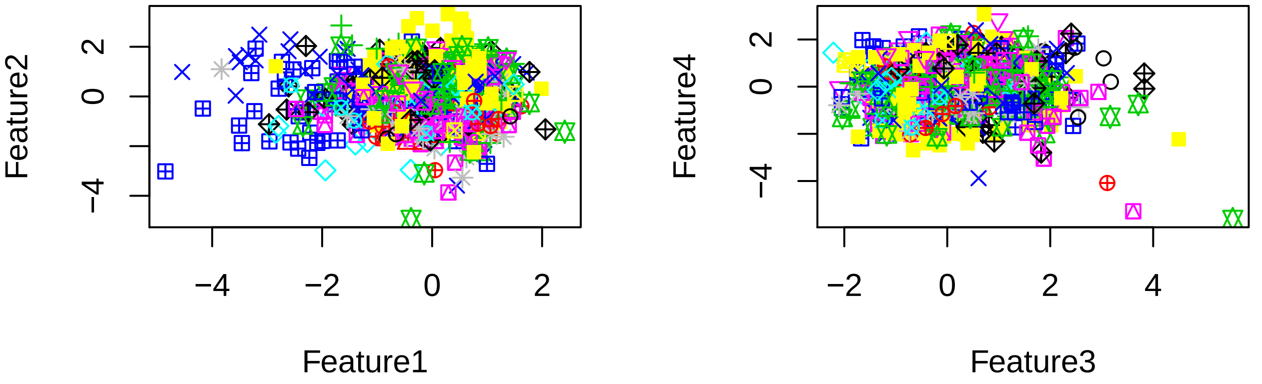

Figure 5: Scatter Plots of Features by FA

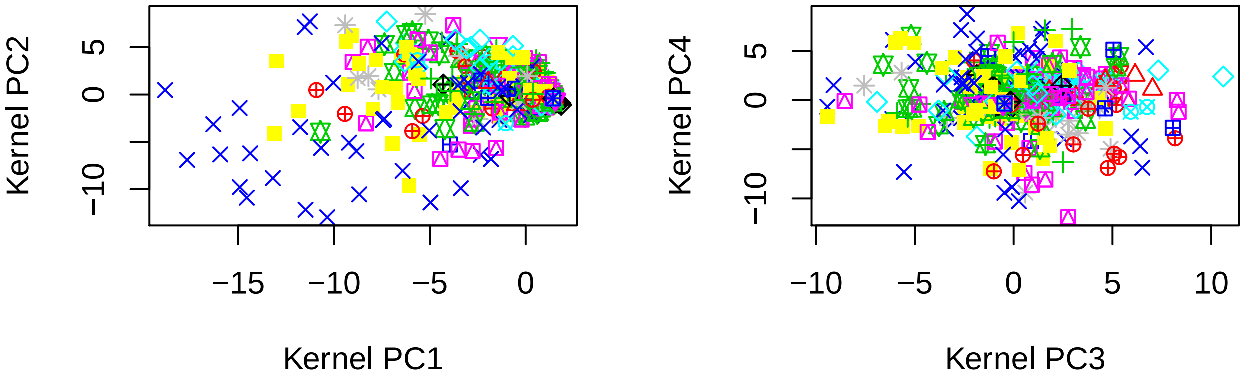

Figure 6: Scatter Plots of kernel PCs

Figure 4 through 6 display the performance of new features. However, none of these three methods overcome the others; this can be expected because most of the explanatory variables (music features) were originally uncorrelated.

3.4. Linear Regression and Extreme Gradient Boosting Machine (XGBoost)

The diagnostic plots (Figure 7) of the linear regression fittings implies normal assumption on the response variable is possible. This may not true that the origin location of musics has normal distribution because it directly depends on the data collection. Nevertheless, with this data and responses, we can at least make predictions with linear regression model. On the other hand, we can also perform classifications with the data by implicating machine learning algorithms. The performance of linear regression can be improved in terms of increment in R^2 by adding interaction terms stepwise and choose new terms based on BIC; however, since there are two response variables – longitude and latitude – and two BICs it becomes another optimization problem when it comes to BIC comparison among models. It is not reasonable to sum both BICs of models of longitude and latitude because the scales of BICs are different. In this project, instead of stepwise variable selection for regression model, XGBoost, an algorithm under the gradient boosting framework, will be used to perform the regression and classification with the selected features by three methods above.

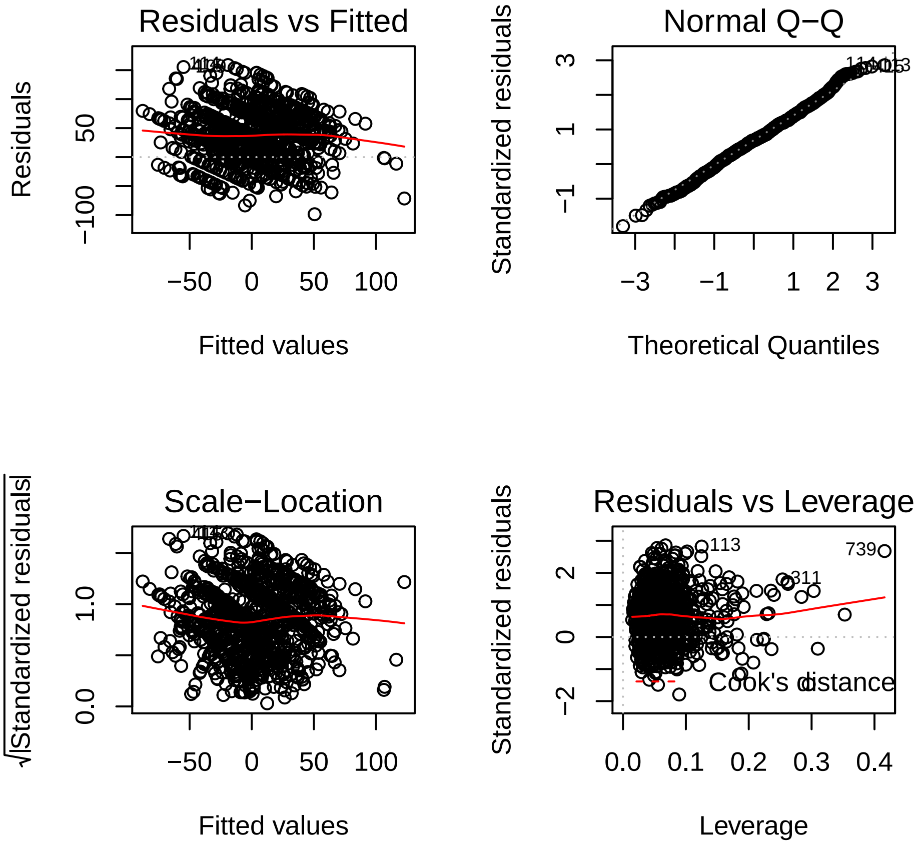

Figure 7: Diagnostic Plot for Linear Regression : lm(latitude .)

Results and Summary

The gradient boosting machine typically use decision trees and make the prediction in the form of an ensemble of the trees. Given an objective function which usually consists of training loss and regularization terms, this algorithm solves the optimization problem by exploiting a gradient descent algorithm to minimize the loss. And XGBoost is a name of a software and one type of gradient boosting machine.

First, the data set was divided into 80% of training set and 20% of test set. In order to solve both classification and regression problems, I used two different models. For the regression, I used longitude and latitude information as the response variable and fitted them into linear regression model. For the classification, I used sub-region information as the only response, and fitted the data into the multinomial model. As the output of the algorithm, Tables 3 and 4 are provided.

Table 3: Prediction Accuracy with Test Set

	PCA 30	FA 30	KPCA 30	RawData
Prediction Accuracy	0.3932039	0.4029126	0.3883495	0.5339806

Table 4: Root Squared Error with Test Set

	PCA 30	FA 30	KPCA 30	RawData
latitude Root Squared Error	239.224	244.8428	251.8993	225.8260

	PCA 30	FA 30	KPCA 30	RawData
longitude Root Squred Error	632.547	666.9952	698.0288	598.0429

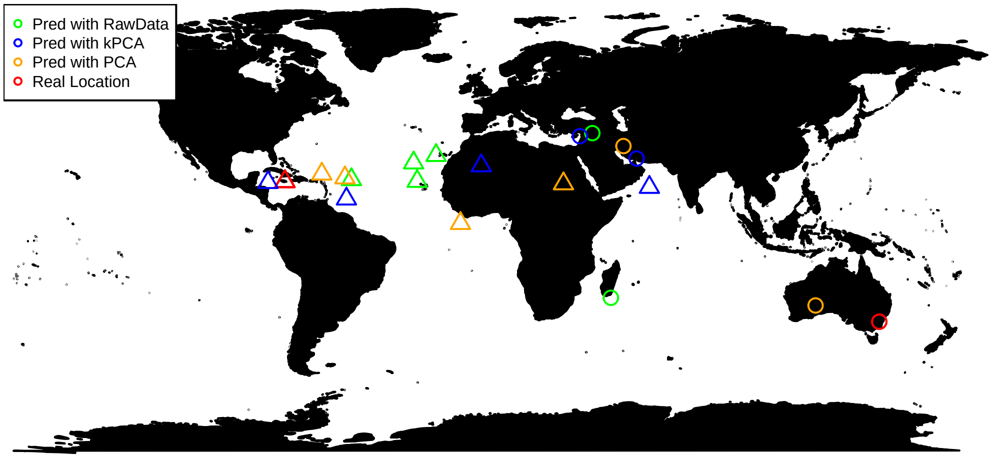

Figure 8: Prediction Plot for Australia and Caribbean

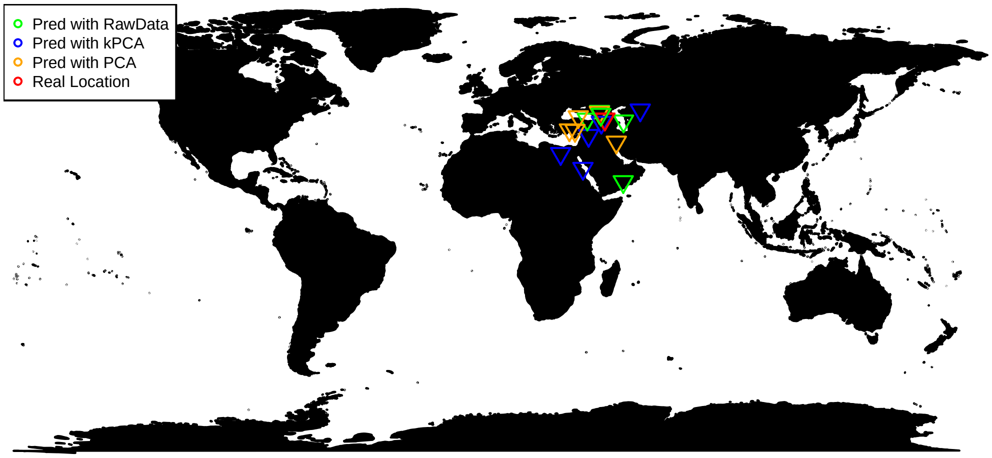

Figure 9: Prediction Plot for Eastern Europe

In Figures 8 and 9, location predictions based on the regression model were plotted. Especially, in Figure 8, it displays the results of Australia and Caribbean subregions with predictions by various features extracted by different methods. Every data set, even the raw data, does not predict the true subregion properly. Only prediction with kernel PCA for Caribbean subregion may work better than others but it is hard to say it is better in overall. In Figure 9, it displays the results of Eastern Europe subregion. In this case, the prediction with the raw data performs the best; however the difference in performence with other features is quite small. In Table 4, the root squared errors were computed for each longitude and latitude and as we might expect the overall performance of the raw data is the best; however, it is also notable that the new features also worked properly.

Table 3 shows the accuracies of the predictions by each feature sets. The classification was done by multinomial family, and none of longitude and latitude was used for the prediction; only subregion information was used. In this classification problem, the raw data performs the best according to the Table 3, and the rest of features by PCA, kernel PCA and factor Analysis perform quite similarly.

As expected, the results of classification and regression were not satisfiable. The explanatory variables do not have enough discriminatory power, and even with all 68 features the results cannot be good enough. Thus, it is quite natural that the feature selection did not perform well with this example. However, if there exist some explanatory variables with more discriminative power our data, the feature selection might make more sense; the PCA or factor analysis would capture the most important features and dramatic dimensional reduction would be expected. Zhou et al. (2014) also did a regression on the same data by using random forest tree algorithm. The mean circle distance error was 3,113 km which was still large. For better prediction, other audio feature extraction methods should be considered in order for features with more discriminative power.

Citations

1. Tzanetakis, G., & Cook, P. (2000). MARSYAS: A framework for audio analysis. Organised Sound, 4(3), 169-175.
2. Zhou, Fang & Claire, Q & D. King, Ross. (2015). Predicting the Geographical Origin of Music. Proceedings - IEEE International Conference on Data Mining, ICDM. 2015. 1115-1120. 10.1109/ICDM.2014.73.
3. Chen, Tianqi & Guestrin, Carlos (2016). XGBoost: Scalable Tree Boosting System. CoRR. abs/1603.02754.