|  |
| --- |
| Oregon State University – ST 599: Statistical Computing & Big Data |
| Predicting with the Million Song Dataset |
|  |

|  |
| --- |
| Sarah Guermond  Andy Olstad  Jasmine Pahukula  Timothy Skalland  6/2/2014 |

**Background & Data:**

The Million Song Dataset is a collection of audio features and metadata for a million contemporary popular music tracks, collected by music intelligence platform The Echo Nest. The data set was created for research purposes under a grant from the National Science Foundation.

*Question of interest: Can we use song feature data to predict the year of a song?*

This is fundamentally a prediction problem: we want to predict the year of release of new songs based on information we already have about other songs. While the Million Song Dataset has a great many variables including the start time for each beat, the “danceability”, “loudness”, “timbre” and more, our dataset includes 515,345 rows and 91 columns including: year, 12 timbre averages, and 78 covariance terms. Since the columns were not labeled, it was difficult to interpret what these terms meant, except that they were output from “The Echo Nest”. Because exactly 12 columns (13 through 24) were strictly positive, we believe that these are the variance terms. Although we did not thoroughly understand “timbre”, we understood that it described the quality of a song beyond simple tone and pitch. The 12 values each summarized some aspect of the timbre of the entire song: “flatness”, “brightness” and so on.

We decided to think of year as a continuous response and hoped to find predictors which would either correctly guess the year, or err only a little on one side or another. The dataset consists of a large number of songs, the vast majority of them from 1990-2010.

The dataset was partitioned into a “training set” and “test set”, which we respected.

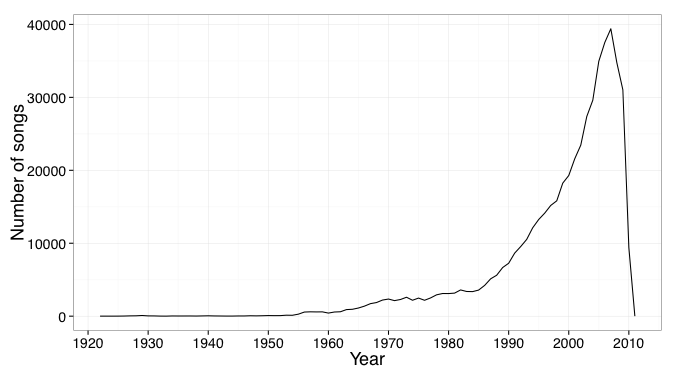
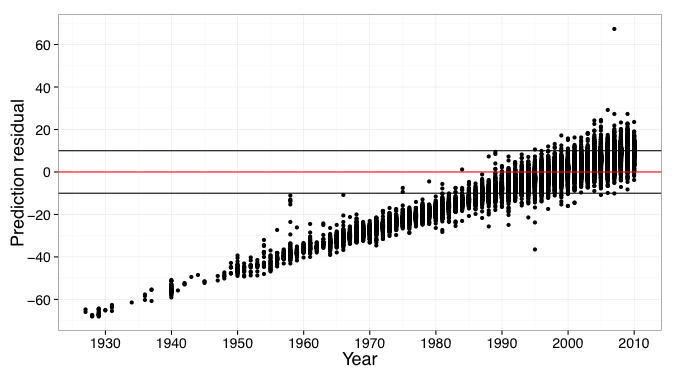
**Machine Learning Method:**

Principle Component Analysis develops linear combinations of variables that explain a desired percent of the overall variation in those variables. The ultimate goal of PCA is dimension reduction. The Pareto Principle, or ‘80/20’ rule, was used as a threshold to identify the principle components to be used. The ‘80/20’ rule states that 80% of the variation will be explained by 20% of the variables.

We created a training set (463,715 songs) and test set (51,630 songs), as recommended. We performed PCA to reduce the dimensionality of our training data from 90 predictors to 8 principal components since we found that ~80% of the variation was explained by these principal components. We then performed multiple linear regression on these 8 principal components, treating the year of the song as a continuous response as well as considering a grouping the years into 3 categories (pre-1960, 1960-1989 and post-1990) on which we performed multinomial multiple linear regression to try to predict the era of the song.

We also performed simple and multiple linear regressions on the original data (before the PCA reduction) as a check against the assumptions used in PCA. Furthermore, we tried building a model just using the timbre means (ignoring the covariance terms), again with the goal of providing a check against our more complicated models.

**Findings:**



The data was heavily skewed in favor of the 1990s and 2000s, which made accurate predictions outside of these decades nearly impossible with our techniques. The model overestimated the year of release for earlier songs (1920 to 1980). As expected, 1990s and 2000s had the best year predictions likely due to the large number of songs for those years. We have created an estimator with a strong bias towards a particular range of years.

Performing a multiple linear regression against the raw data showed that while nearly every variable was “statistically significant”, all were of little practical significance since all coefficients were near zero. This regression (R squared .24) consistently guessed years in the 1990s and 2000s, as we saw with the PCA. A further approach involved guessing that the first 12 columns were our mean values and performing a regression on them (R squared about 0.15). We again saw the same strong favoring of the 1990s and 2000s. And finally the multinomial regression favored these years as well, wanting to put all predictions in the “Post-1990” category with about 80% probability. Overall we find highly statistically significant predictors (of little practical significance) using the different methods.

**Assumptions & Limitations:**

We chose to consider year as a continuous response because we want the best approximation of the year as possible: if we “miss” a 2004 song we'd like to end up somewhere in the same decade. We did look at categorizing year as described once we found that treating is as a continuous predictor was not working. One of the main concerns with PCA—difficulty in interpreting results-was of little concern as our task was purely prediction and because we have very little information about the content of our dataset which renders interpretation difficult anyway.

**Scaling**:

A sub-training set pulled out the same number of PCs as the bigger training set. PCA scales well to very large datasets, although dimensional reduction may become computationally intensive.

**References:**

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.