Background & Data

Data: Million Song Dataset is a collection of audio features and metadata for a million contemporary popular music track, collected by music intelligence platform the Echo Nest.

~500,000 songs from the

* Year
* 12 mean timbre features
* 78 covariances

No variables labeled

(insert table of data distribution by decade)

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

Approach: Principal components analysis

WHY?

HOW?

1. Take smaller random subset of 10,000
2. Split into 7,500 training, 2,500 test set
3. Run PCA on correlation matrix of 90 predictors
4. Used first 8 species

Findings:

Assumptions & Limitations:

Scaling:

Sub-training set pulled out the same # of PCs as the bigger training set: 8 PCs of the correlation matrix predict 80%.

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.