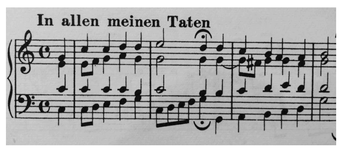
# Clustering Bach Chorales: Insights into SATB and Bach's Style

Diego Hernandez, Hope Casey-Allen, Jian Yang Lum
CS 229, Stanford University

## Goals

- To find and interpret clusterings of Bach chorales => allow musicians to connect musical and contextual features.
- To determine features (independent of relative positions) that distinguish individual SATB voices in Bach chorales.

## Methodology

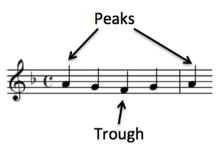



```
C4      2       1 q     d
measure 1
C4      2       1 q     d
C4      2       1 q     d
C4      2       1 q     d
B3      2       1 q     d
measure 2
C4      4       1 h     d
B3      2       1 q     d               F
B3      2       1 q     d
```

- Scraped scores encoded in the Humdrum file format, using kern representation, from CCARH

- Used MIT's *music21* framework to extract features from 183 Bach chorales and their 183 * 4 individual voices

<u>Example features:</u>
Duration of melodic arc
= Average number of notes separating peaks and troughs



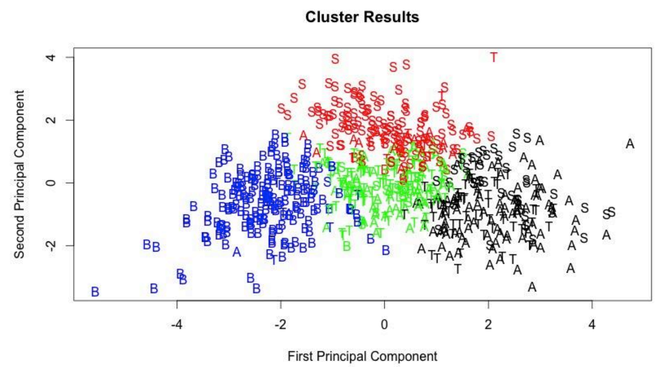

Average melodic interval
= Average # of semitones between consecutive notes

Melodic tempi
= average # of notes per second and average note duration

Categorized features into 8 buckets:
- Dissonance
- Prevalence of major vs. minor chords
- Note range
- Average melodic interval
- Chromaticism vs diatonicism
- Melodic contour
- Melodic tempi
- Phrase structure
- Amount of arpeggiation

## Clustering Voices

Algorithm: K-Means, K *implicitly defined to be 4 (SATB)*




**Error rate: 0.321**

Cluster Assignments:

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| S | 127 | 39 | 15 | 2 |
| A | 19 | 92 | 70 | 2 |
| T | 13 | 64 | 100 | 6 |
| B | 3 | 0 | 2 | 178 |

Results: Bass line most distinct, soprano next

## Voice Prediction

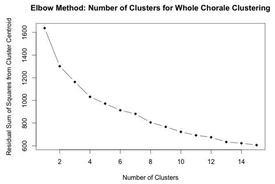
Algorithm: Softmax / Multinomial Logistic Regression, CV = 10
*10-fold CV Error Rate = 0.255*
*p-Values < 0.05*:

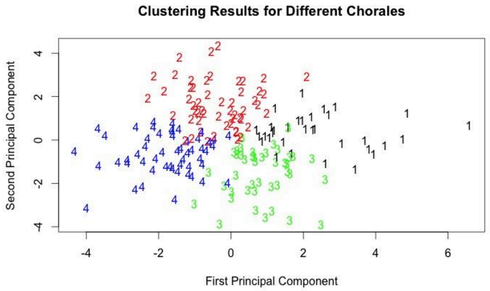
| averageMelodicInterval | rangeIndivVoices |
|---|---|
| 2.118886e-03 | 0.002563396 |
| 1.081896e-06 | 0.002920784 |
| 2.118639e-06 | 0.001877747 |

Results:
a) Most significant predictors of voice type:
   i) Average Melodic Interval (largest for Bass)
   ii) Range of Individual Voices (largest for Bass)
   iii) Number of Repeated Notes (most for inner voices)
b) Interpretation (a basic model of phrasing): To highlight independence of lines, performers can emphasize
   i) Highest and lowest notes
   ii) Transitions between notes

## Clustering Chorales

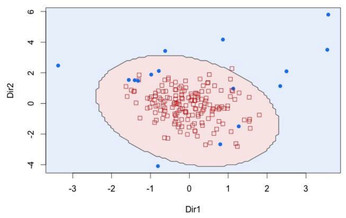
1) *K-Means, k (determined by Elbow Method) = 4:*

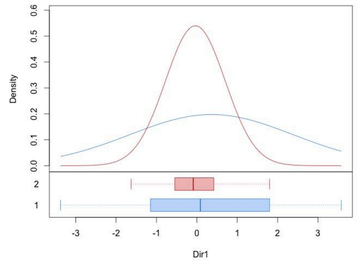



Findings: 4 distinct groups clearly separated:
1 - arpeggiated, repeated notes
2 - small range
3 - big melodic interval, few repeated notes
4 - long melodic arcs, highly consonant, big range

2) *EM algorithm (R package: mclust); k = 2 (lowest BIC)*




Findings: Classification Plot (above) implies no obvious clustering; density seems to overlap. Very few (16) on one cluster

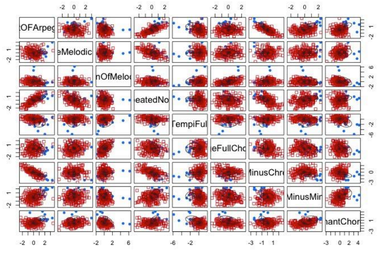
No conclusive results drawn from EM algorithm *or Bach had a uniform style*

```
---------------------------------------------------------
Density estimation via Gaussian finite mixture modeling
---------------------------------------------------------

Mclust EVE (ellipsoidal, equal volume and orientation) model with 2 components:

 log.likelihood   n df       BIC       ICL
      -1868.43 183 72 -4111.942 -4121.038

Clustering table:
  1   2
 16 167
```



## Future Work

In future, can apply this methodology
- with different musical features (reflecting diff. styles of music theory)
- with different composers and types of compositions

In addition, analyze our data with respect to
- contextual features (e.g., map clusters to date of composition, intended occasion of performance, etc.)