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

Diego Hernandez, Hope Casey-Allen, Jian Yang Lum

Stanford University, California, United States

1 Introduction and Literature Review

Few composers have been as influential in the Western musical tradition as Johann Sebastian Bach (1685-1750), whose four-part chorales for soprano, alto, tenor, and bass (abbreviated SATB, and listed in order from highest to lowest) have become well-known distillations of important music-theoretical principles. Such principles, dealing with both the construction of individual parts and the interactions between those parts, came to dominate Western music for about 150 years after Bach's time and continue to inform musical understanding today.

Some research in machine learning has been devoted to computerized harmonizations (i.e., given a soprano melody, generate the other three voices in Bach's style). However, previous projects have encountered obstacles in their work due to a lack of rigorous understanding of the nuances in Bach's style. For instance, Allan and Williams (2005) attempted to generate chorale harmonizations in Bach's style using a Hidden Markov Model. The results were suboptimal: although the relationships between voices in the resulting chorales were consistent with harmonic rules seen in textbooks, the individual voices themselves sometimes erratically jumped in pitch instead of flowing in natural, intentional contours.

It is clear that research is needed to more precisely characterize SATB voices in Bach chorales, the insights from which would not only improve algorithms that seek to generate harmonizations but would be useful to musical theorists and performers in understanding and interpreting music.

Our first goal is thus to determine features (independent of relative positions) that distinguish individual SATB voices in Bach chorales. Our second goal is to find and interpret clusterings of chorales that may be artistically useful (e.g., helpful in identifying musically similar chorales for purposes of inspiration, comparison, etc.)

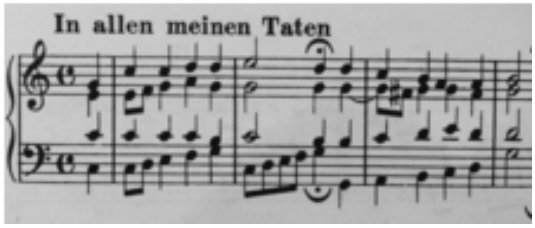
We have 2 inputs, corresponding to the different goals: inputs for the first goal are the features from the voice parts of these chorales, looking at 183×4 SATB parts in total. We used k-means clustering, with $k = 4$, in an attempt to yield clusters that divide the voices into their natural SATB taxonomies, revealing the defining characteristics of each melody; and softmax / multinomial regression for voice prediction, where the output is one of the possible SATB values.

The inputs to our second goal are features we extracted from the scores of 183 of Bach's chorales. We used the elbow method to determine that $k = 4$ for k-means clustering on this input, and outputted distinct clusters for the chorales which reveal insights into Bachs style; and we used EM in an attempt to reveal traits of Bachs style as well.

We believe our task has never been attempted in the literature. The closest task we could find to ours is in [3], in which Quinn and Mavromatis clustered every chord transition in a combined corpus of chorales of two categories: chorales by Bach and modal chorales from about a century before Bach. They then compared each chorale category's coverage of each cluster as a means of identifying distinguishing harmonic characteristics of each category. Although this research provides a window into the overall harmonic landscape of Bach's chorales, it says nothing about the characteristics of SATB voices.

2 Dataset and Features

Our dataset contains 183 Bach chorales and 183×4 SATB voice parts, all accessed from "http://kern.humdrum.org/cgi-bin/ksdata?l=musedata/bach/chorales&file=chor-01.krn&f=musedata" where 01 is the number of the chorale. The dataset is encoded in the Humdrum **kern data format; Humdrum was developed by CCARH (Center for Computer Assisted Research in the Humanities) at CCRMA (within Stanford).



(a) Sample of a Bach Chorale opening. S is the top line, A the second highest, T the third, B the lowest.

```
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
```

(b) The same score, translated into **kern

As seen above, musical information (pitches, metronome markings, measures, metadata, etc) is encoded to represent a full musical score: We used a toolkit developed by MIT called **Music21** to separate the chorales into parts and extract their features. We categorized relevant features into 9 buckets. For some buckets with multiple features, we used PCA and extracted the principal component (which for all cases explained at least 96% of variation), while for others we added or subtracted the feature values. Normalization was performed on all variables to have mean 0 and variance 1, to aid k-means clustering by assigning equal importance to each predictor/variable.

3 Methods

The two goals we tried to achieve required two different approaches each; determining the difference between individual voices is fundamentally a classification problem, while attempting to observe similarities behind chorales that do not have an observable grouping is an unsupervised problem. To that extent, we employed a mixture of parametric and nonparametric methods for differentiating SATB voices (namely, softmax/multinomial logistic regression, and k-means clustering) and multiple unsupervised methods (k-means and EM) for clustering chorales.

The softmax, or multinomial regression model, posits k (in this case, 4, one corresponding to one voice type) different outcomes/voice types, with total probability summing to one (with a constraint on the last variable). Maximization of the log-likelihood of the model:

$$\sum_{i=1}^{m} \ln \prod_{l=1}^{k} \left(\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right)^{1\{y^{(i)}=l\}}$$

was done using a single-hidden-layer neural network, in which a (hidden) layer of variables is estimated using a vector of weights w with $|w| = nk$ (n being number of features, $k = 4$). The neural net then used backpropagation of errors to the inputs with gradient descent to update the parameters returned. The softmax model was chosen as it did not require breaking down into binary classification problems.

k-means clustering iteratively seeks to find the cluster centroids that minimize the aggregate residual sum of squares of all points in that cluster, $\sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$, by first reassigning both the assignments of each points to the nearest centroid, then moving that centroid to the mean of all points with that same assignment. k-means was used as an exploratory data analysis mechanism, as well as understanding properties of each group by exploring properties behind its centroids.

The EM algorithm, used when a latent random variable is suspected (in this case, the different styles Bach may have used), performs maximum likelihood estimation in a two-step mechanism: it first constructs a lower-bound on the likelihood via Jensen's inequality (by equating the new distribution of the latent variables to be the posterior distribution of the latent variables given the observed variables and the current parameters) (E-step), then optimizes it (by maximizing the lower-bound of the likelihood with respect to the parameters) (M-step). It was another option over the k-means algorithm, in that it allowed us to observe potential distribution properties of the latent variables, besides looking at centroids. All data analysis was done in **R**, with packages **nnet** and **mclust** for softmax and EM respectively.

4 Results and Discussion

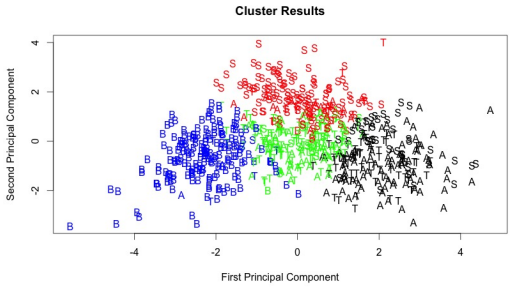
4.1 Individual Voice Clustering/Classification

For clustering, besides attempting to lower the misclassification rate, we hoped to identify variables of interest that strongly influence the classification of voice lines into their voice types, as the goal was to study the properties of different melodic lines.

The k-means approach ($k = 4$, corresponding to SATB voices) yielded a misclassification rate of 32.1%, with most errors misidentifying the SAT voices (only 6% involving both Type I or II errors involving the bass line); this implies that the bass line is most distinct, followed by the soprano line.

Table 1: K-means cluster results versus actual voice type

	One	Two	Three	Four
S	127	39	15	2
A	19	92	70	2
T	13	64	100	6
B	3	0	2	178



(a) Cluster Results, plotted against first two principal components.

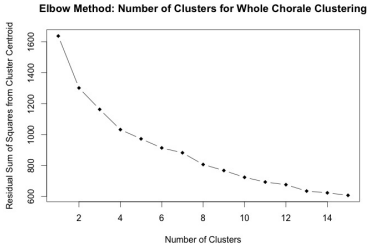
The softmax approach had a 10-fold cross-validation error of 25.5%. With the dataset, the p-values were calculated for each estimator using the test statistic of $Z = \frac{\mathrm{E}(\theta_i)}{\sqrt{\mathrm{Var}(\theta_i)}}$; only the p-values that were statistically significant (i.e. p-value less than 0.05, for the null hypothesis that $\theta_i = 0$) were considered for analysis.

As an entire predictor, only **averageMelodicInterval** and **rangeIndivVoices** were statistically significant for all three voice types (one of the four categories, the bass, was arbitrarily used as the baseline, hence only three predictors were needed). For both predictors, the bass had the highest average melodic interval and the largest range of individual voices, agreeing with music-theoretical approaches of wide movements in the bass line. Amongst other variables that were statistically significant, the alto and tenor both showed increased percentages of repeated notes, the soprano generally had faster melodic tempi and a lot more

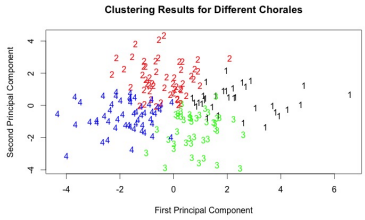
diatonic movement (as opposed to chromatic movement). For all data, duration of melodic intervals was not a useful predictor.

4.2 Chorale clustering: Unsupervised Learning

In picking the appropriate k to use for the initial k-means exploration, we elected to use the elbow method (b), a visual inspection method that attempts to find the maximum number of k for which the improvements by adding one more are minimal (in this case, looking at the smallest angle for each point, when plotted against residual sum of squares (with respect to centroids)). While there was no clear-cut change in gradient, an observable bend at $k = 4$ was noted, and as such k was initialized to be 4.



(b) Elbow Method Plot: RSS against k. We pick the point with sharpest change of gradient, here 4.

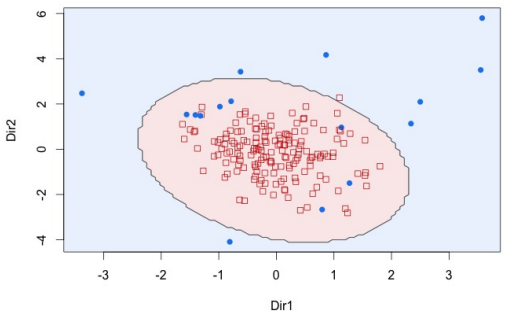


(c) Cluster Visualization, plotted on first two principal components.

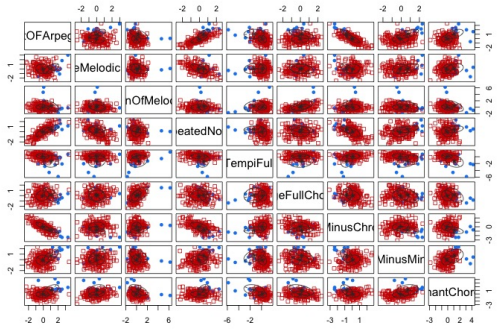
The initial round of clustering (c) produced an equal-sized grouping of clusters. Observing the cluster centroids, the key findings were that the four groups had properties corresponding to:

1. arpeggiated, repeated notes
2. small range
3. big melodic interval, few repeated notes
4. long melodic arcs, highly consonant, big range

The EM algorithm (which proceeded by a Gaussian mixture modelling), by contrast, returned a cluster number of 2, based off the calculation of the Bayes Information Criterion (BIC), $-2\ln(\hat{L}) + k\ln(n)$; the lowest BIC for all models returned a model that had an ellipsoidal distribution, had equal volume and orientation, and had 2 clusters.



(d) Classification Boundary, 2-cluster model.

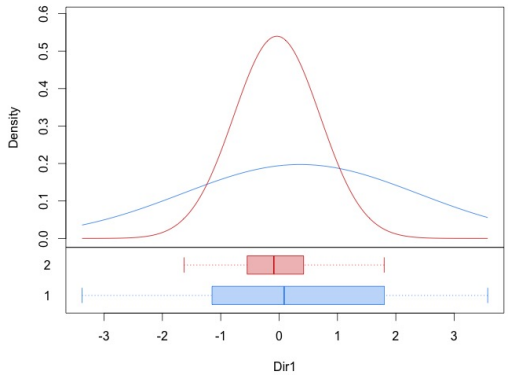


(e) Biplot of variables, 2-cluster model.

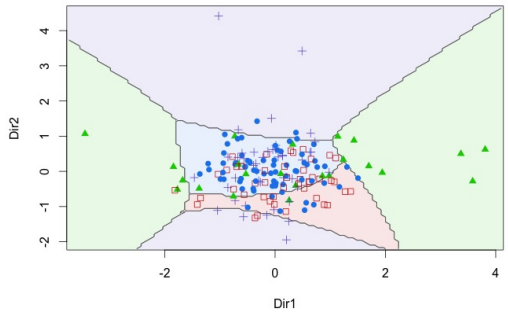
However, the EM algorithm has one cluster with only 9% of data points (16), while the other cluster had 91% of the data points. Furthermore, a classification decision boundary (on a two-dimensional feature space

comprising the first two principal components of the data) (d) reveal a circular inner, coupled with an outer and scattered cluster.

Similarly, a biplot of all variables (e) do not reveal two distinct cluster groups, while a density model of both latent variables (cluster groups) (f) show that both means overlap within a high-density region, hence making it hard to separate the clusters into linearly separable groups.



(f) Density plots, 2-cluster model.



(g) Classification Boundary, 4-cluster model.

Another round of EM was conducted, this time with $k = 4$ (in conformance to the elbow result). The results were not as clear as that of k-means, featuring a substantial proportion of outliers (g). The findings of the EM demonstrate that perhaps more variables might be required to make a more conclusive stance on cluster assignments, and that perhaps Bach did not actually have multiple styles used. Furthermore, the EM algorithm's failure to detect any substantial/clear-cut clustering may have to do with the assumption required that the model was Gaussian; some of the variables (e.g. dissonance) was more likely than not drawn from either a Bernoulli model or some non-Gaussian distribution.

5 Conclusions and Future Work

Our first goal, investigating the differences in musical lines through structural elements, proved useful in constructing a typology of voice type; the conclusions (mostly through softmax), which include bass lines having widest jumps and largest range and soprano lines having faster melodic tempi etc, conform to theoretical part-writing rules.

However, our second objective - to observe clustering within different Bach chorales - proved mixed; a potential failing of the EM algorithm, as pointed out, could have been the Gaussian assumption, while the k-means centroid analysis could see further research done on it, not only to verify that such groupings exist, but to see how it may have correlated with his other works across time (his cantatas, large-scale masses).

It follows that future work would include investigating our second objective even further. We could gather data for other composers and run EM/other algorithms to clarify the potential failing of our EM algorithm and see if style differences manifest. We would also perform further contextual analysis by extracting features for date of composition (only available for some pieces). This would enable us to investigate not only if a composer's style changed over time, but also how their composition style corresponds to the liturgical calendar, which was a major influence for much of western musical tradition. Overall, the insights gleaned from more research in this area would shape musicians' understanding of these pieces and the subtleties that accompany them.

References

[1] Allan, Moray, and Christopher KI Williams. "Harmonising chorales by probabilistic inference." Advances in neural information processing systems 17 (2005): 25-32.

[2] Morris, Robert D. "New directions in the theory and analysis of musical contour." Music Theory Spectrum 15.2 (1993): 205-228.

[3] Quinn, Ian, and Panayotis Mavromatis. "Voice-Leading Prototypes and Harmonic Function in Two Chorale Corpora." MCM. 2011.

[4] Temperley, David. "Probabilistic Models of Melodic Interval." Music Perception: An Interdisciplinary Journal 32.1 (2014): 85-99.

[5] Witten, Ian H., Leonard C. Manzara, and Darrell Conklin. "Comparing human and computational models of music prediction." Computer Music Journal (1994): 70-80.