Where Does Haydn End and Mozart Begin? Composer Classification of String Quartets

Katherine C. Kempfert[1] and Samuel W.K. Wong[2]
[1]University of Florida, Department of Statistics
[2]University of Waterloo, Department of Statistics and Actuarial Science

Abstract

For humans and machines, perceiving differences between string quartets by Joseph Haydn and Wolfgang Amadeus Mozart has been a challenging task, because of stylistic and compositional similarities between the composers. Based on the content of music scores, this study identifies and quantifies distinctions between these string quartets using statistical and machine learning techniques. Our approach develops new musically meaningful summary features based on the sonata form structure. Several of these proposed summary features are found to be important for distinguishing between Haydn and Mozart string quartets. Leave-one-out classification accuracy rates exceed 85%, significantly higher than has been attained for this task in prior work. These results indicate there are identifiable, musically insightful differences between string quartets by Haydn versus Mozart, such as in their low accompanying voices, Cello and Viola. Our quantitative approaches can expand the longstanding dialogue surrounding Haydn and Mozart, offering empirical evidence of claims made by musicologists. Our proposed framework, which interweaves musical scholarship with learning algorithms, can be applied to other composer classification tasks and quantitative studies of classical music in general.

1 Introduction

Music information retrieval (MIR) is an interdisciplinary field that has grown as digitalized music data and computing power have become widely available. Methods have been developed to automatically perform many types of tasks in MIR: composer, genre, and mood classification (Pollastri & Simoncelli, 2001), (Tzanetakis & Cook, 2002), (Laurier, Grivolla, & Herrera, 2008); query, such as matching a sung melody to a song (Kosugi, Nishihara, Sakata, Yamamuro, & Kushima, 2000); generation of novel music (Johanson & Poli, 1998); and recommender systems for consumers, such as Spotify and Pandora (Van den Oord, Dieleman, & Schrauwen, 2013). Thus, MIR has become increasingly relevant to how music is both studied and enjoyed. For a review of MIR and its applications, see Downie (2003) and Schedl, Gómez, and Urbano (2014).

In this MIR study, we focus on composer classification. Specifically, we use the content of music scores to classify Haydn and Mozart string quartets, motivated by the historical and cultural significance and the difficulty of the task. Haydn and Mozart had many similarities: "They were not only contemporaneous composers, using the harmonic vocabulary of the late eighteenth century at a time when its syntax was the most restricted and defined, but they shared the summit in the development of ... the sonata style" (Harutunian, 2005, Foreword). At times, members of royalty commissioned both Haydn and Mozart (for example, King Frederick William II of Prussia), which may have further constrained Mozart's and Haydn's compositions to be similar (Zaslaw, 1990). The two composers had similar patrons and cultural upbringings, both Austrians active in Vienna during periods of their lives (Zaslaw, 1990). In addition to their shared cultural influences, the composers directly influenced each other, with "quartet playing…central to contact between Haydn and Mozart" (Larsen & Feder, 1997, p. 54). In fact, Mozart dedicated his Op. 10 set of six string quartets to Haydn. After hearing a performance of the quartets, Haydn told Mozart's father Leopold, "I tell you before God as an honest man that your son is the greatest composer known to me either in person or by name. He has taste, and what is more, the most profound knowledge of composition" (Zaslaw, 1990, p. 264).

For centuries, the music and history of Haydn and Mozart has been compared by scholars. According to Robert L. Marshall (2005), "The critical and scholarly literature devoted to this repertoire is nothing short of oceanic and includes contributions from some of the most profound musical thinkers of the past two centuries–among them such authorities as Hermann Abert, Friedrich Blume, Wilhelm Fischer, Leonard Ratner, Charles Rosen, and Donald Francis Tovey" (Harutunian, 2005, Preface). More recent comparative analyses include *Metric Manipulations in Haydn and Mozart* (Mirka, 2009) and *Haydn's and Mozart's Sonata Styles: A Comparison* (Harutunian, 2005). Mirka argues that Haydn's music is "artful popularity", "appealing to all kinds of listeners", while Mozart's "overwhelming art", stemming from "harmonic and polyphonic complexity … required greater intellectual involvement of listeners …" (p. 303). Harutunian confirms the overwhelming artistry of Mozart, repeatedly referring to his music as "operatic" (p. 65, 81) and even citing this as a reason for his greater success over Haydn in the opera. These differences between Haydn and Mozart are only a few simple examples of the many complex qualitative comparisons undertaken over the centuries.

Despite music scholars' claims that Haydn and Mozart possess distinctive personal styles, many listeners fail to hear any differences. The difficulty of identifying Haydn versus Mozart string quartets can be exemplified by the results of an informal online quiz (created by Craig Sapp and Yi-Wen Liu of Stanford University and accessed at `http://qq.themefinder.org`). The user is prompted to answer a series of questions (including number of years in classical music training, instruments one can play, and familiarity with Haydn and Mozart), then to identify randomly selected Haydn and Mozart string quartets. Even the users with maximal music experience have not achieved more than 67% accuracy on average. Although this quiz is not a random and representative survey, the results still evidence the difficulty of the Haydn-Mozart classification task.

Over the years, statistical and machine learning methods have been applied to many tasks with which humans have struggled. Such methods use probabilistic models to describe data; for the task of *classification*, where each observation belongs to one of several classes, any type of model for a categorical response variable can be used. A fitted classification model then determines the most probable class to which an input observation belongs. Variables used to classify observations are also known as *features*, and the calculation of features from data is referred to as *feature extraction*. Feature extraction techniques can range from fully automatic (e.g., a matrix representation of an image) to manual (e.g., calculating specific summary measures). An advantage of manual definition and encoding of variables is in their interpretability. The interested reader may refer to Hastie, Tibshirani, and Friedman (2001) for an excellent overview of the main tasks, methods, and issues in statistical and machine learning.

The Haydn-Mozart string quartet classification problem is one such area that has benefited from these statistical and machine learning methods. However, to date, classification accuracies have been surprisingly low for this task. Prior to our study, the highest classification accuracy was 80.40%, with a predictive model that used pixel-related features automatically extracted from images of piano roll scores (Velarde, Weyde, Chacón, Meredith, & Grachten, 2016). However, the computer vision techniques lacked musical interpretability, and that model contributed little insight to the musicological aspects of Haydn-Mozart comparative studies. Thus, we are motivated to develop a classifier using features that are both musically interpretable and lead to high classification accuracies. As in many other prior studies, we use features manually extracted from the musical scores of Haydn and Mozart string quartets. These include summary statistics calculated for individual voices, such as the mean and standard deviation of pitch in the cello voice. The novelty in our approach is that we leverage musical scholarship to extract more sophisticated features based on the structure of Mozart and Haydn compositions, where the classical *sonata form* has a key role.

Our contribution in this study is an approach that combines musical expertise with statistical learning, to improve understanding of the compositional differences between Haydn and Mozart string quartets. Our results show that Haydn and Mozart string quartets are discriminable, as evidenced by high classification accuracy rates that are attainable using only musical features extracted from the scores. Overall, we recommend our approach as a general framework for composer classification tasks (and other topics in MIR) that prioritizes both musical interpretability and quantitative validation.

In the next section, we present our dataset. The motivation and calculation of musically meaningful features are discussed in Section 3. In Section 4, we explain the feature selection approach and classification model for discriminating between Haydn and Mozart string quartets. The results are presented, compared to prior studies, and musically interpreted in Section 5. We conclude our paper and suggest further directions of research in Section 6. Finally, the dataset and source code for our methods are publicly available at `https://github.com/wongswk/haydn-mozart`.

2 Data

Music data can be expressed in the form of auditory or symbolic information. Audio representations include live performances and recordings, such as MP3 files, CDs and tapes, while symbolic representations include scores, text, and computer encodings like Musical Instrument Digital Interface (MIDI) and **kern (Downie, 2003). Though auditory formats capture pitch, rhythm, and other musical information, they fundamentally rely on a certain performance or performer's interpretation of the music, which can vary substantially for classical music. In contrast, symbolic formats transcribe the musical score itself and thus more closely reflect the intention of the original composer. Our motivation is to identify differences between Mozart and Haydn as composers, so a symbolic format is preferred. To our knowledge, all other Haydn-Mozart classification studies have also used symbolic formats: MIDI (Kaliakatsos-Papakostas, Epitropakis, & Vrahatis, 2011), (Herlands, Der, Greenberg, & Levin, 2014), (Hontanilla, Pérez-Sancho, & Inesta, 2013); **kern (Van Kranenburg & Backer, 2005), (Hillewaere, Manderick, & Conklin, 2010), (Taminau et al., 2010); and piano rolls (Velarde et al., 2016).

We opt to use the **kern symbolic format of music. Its specification permits the encoding of not only pitch and duration, but also of accidentals, articulation, ornamentation, ties, slurs, phrasing, glissandi, barlines, stem-direction, and beaming. Quantitative analysis is facilitated by **kern's ASCII (plaintext) format. A discussion of **kern, as well as other symbolic formats beyond MIDI, can be found in Selfridge-Field (1997).

We obtain the **kern representation of Haydn and Mozart string quartet scores from the KernScores website (`http://kern.humdrum.org/`), which is maintained by the Center for Computer Assisted Research in the Humanities at Stanford University. Each string quartet has one to five movements, with each movement containing the four standard voices (or parts), Violin 1, Violin 2, Viola, and Cello. Together, there are 82 Mozart string quartet movements and 210 Haydn string quartet movements available on the website, representing the majority of known string quartet movements by these composers: 86 movements authored by Mozart and 280 by Haydn. There are 7 **kern files with errors in the encoding of scores, so we omit the corresponding movements from our analysis. Thus, our dataset consists of 82 Mozart movements and 203 Haydn movements.

We process the data in the statistical programming environment R (R Core Team, 2017). For each voice in each movement, pitch and duration information are extracted from the **kern files. Hence, each movement is represented with 8 tracks: pitch and duration tracks for all 4 voices. As an example, Figure 1 displays our pitch and duration encodings for several bars of the Violin 1 part of a Mozart string quartet movement, as we now describe.

Each voice generally only plays one note at a time, such as seen in Figure 1. Chords and harmonic intervals in a single voice (known as *multiple stopping*) occur very infrequently, so for simplicity we retain only the highest of simultaneous notes in those cases. Rests are encoded as 0. The pitch of each note is encoded as an integer between 1 and 12 (except when intervals are calculated, as in Section 3.2.2), following the order of the chromatic scale (with $1, 2, 3, \ldots, 12$ corresponding to C, C-sharp, D, …, B respectively). Thus, octave information is discarded; for example, middle C is encoded as 1, as are any higher or lower Cs. Our reduced representation facilitates analysis by capturing only the most meaningful aspect of pitch; some studies have shown that listeners mostly perceive the pitch of a note relative to the pitches of nearby notes, rather than in terms of absolute frequency (Levitin & Rogers, 2005).

The duration of each note is encoded as the fraction of time it makes up in a bar. For example, in common time, a quarter note is encoded as 0.25. Therefore, the time signature of the movement is implicitly encoded in the duration information we extract.

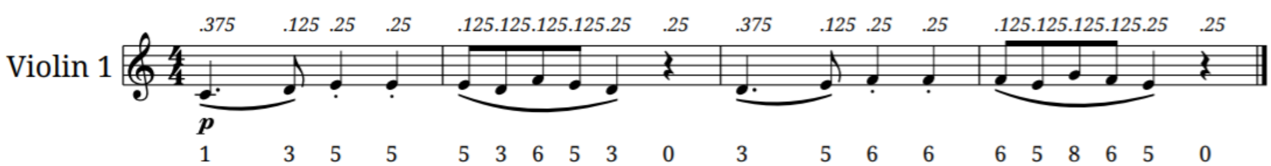

Figure 1: Encoding of an excerpt from Mozart's String Quartet No. 4, Mvmt. 1, in C Major (K. 157). Encoded pitch values and duration values are displayed below and above the score, respectively.

3 Feature Development and Extraction

Feature development involves proposing a litany of summary measures that may help to discriminate between Haydn and Mozart string quartets. The novelty in our approach to feature development is in quantifying the qualitative differences that have been discussed at length in scholarly Haydn-Mozart comparisons. A concise subset of the most important features for classification will be subsequently selected by statistical methods, as discussed in Section 4. Therefore, we can gain insights from both selected and unselected features: selected features suggest areas in which Haydn and Mozart string quartets differ, while unselected features might point to similarities between the composers.

3.1 Review of the Sonata Form

In the exhaustive qualitative analysis *Haydn's and Mozart's Sonata Styles: A Comparison*, musicologist John Harutunian states, "Central to the music of Haydn and Mozart is the concept of sonata style" (p. 1). Hence, it is natural to use the sonata form as a basis for developing new quantitative features. As the sonata form is essential to understanding these features, we provide a brief summary based on Harutunian (2005, p. 1-2).

A piece of music in sonata form has three sections: the exposition, development, and recapitulation.

1. In the **exposition**, the basic thematic material of the sonata is presented. The beginning key is known as the tonic. As the exposition ends, the key modulates, so that it generally ends in a different key from which it started.

2. In the **development**, one or more themes from the exposition are altered, and some new material may be introduced. The development often contains the greatest amount of change.

3. In the **recapitulation**, the opening material is revisited, but it is all in the home key, giving a "sense of resolution and completion" (Harutunian, 2005, p. 1). In general, the recapitulation begins with the opening material in the tonic.

The sonata is the most common structural form for Haydn and Mozart string quartet movements, containing the basic A-B-A structure. Though not all movements strictly follow the sonata form, they often contain similar structure. For example, movements in the Rondo form follow the pattern A-B-A-C-A-B-A (or a variation) and thus have similar elements of an exposition, a development, and a recapitulation. Therefore, sonata-related features are expected to extract meaningful information from nearly all Haydn and Mozart string quartet movements.

3.2 Feature Extraction

This section presents the list of quantitative features that we compute for each Haydn and Mozart string quartet movement, along with descriptions of their musical significance. Many of the features we propose are entirely novel and designed for this specific problem. We incorporate expert musicological knowledge drawn from Haydn-Mozart comparative studies, in particular the aspects of sonata form discussed in (Harutunian, 2005). Other than a study classifying Baroque style composers using contrapuntal features (Mearns, Tidhar, & Dixon, 2010), we are unaware of any prior MIR studies on classical music that have relied on musically sophisticated features.

We complete our feature set by including some that have worked well in previous studies. We may organize our features into five main categories: basic summary, interval, exposition, development, and recapitulation. As appropriate to each category, *monophonic* and *polyphonic* features are considered. Monophonic features are intended to measure the specific melodic and rhythmic role of each separate voice, while polyphonic features capture the interaction between voices. These features are summarized in Table 1 and discussed in depth in the following subsections.

Many of the higher-order *segment* features described in what follows utilize sliding windows, so we describe them here. Let M denote the total number of notes in a voice of a movement and m the desired length of the sliding window (or segment). Then a segment feature is calculated $M-m$ times; for all $i \in \{1, 2, \ldots, M-m+1\}$, the feature is calculated for notes $i, i+1, \ldots, i+m-1$ in order. We need not consider all segment lengths; e.g., segment lengths 8 and 9 would yield essentially the same information, so including one of the lengths should suffice. In our study, we

Table 1: Features for Mozart and Haydn String Quartet Scores

Feature Category	Duration	Pitch
Basic Summary		Number of notes
	Mean and standard deviation of duration	Mean and standard deviation of pitch
		Proportion of simultaneous rests
		Proportion of simultaneous notes
Interval		Proportion of each pairwise interval type
		Voicepair differences in proportion of pairwise interval types
		Proportion of each pairwise interval mode
		Proportion of each pairwise interval sign
	Mean and standard deviation of interval distances	Mean and standard deviation of interval distances
		Voicepair differences of mean and standard deviation of interval distances
		Summary statistics for proportion of minor third intervals in each segment
		Minimum, first quartile, median, third quartile, maximum
		Mean and standard deviation
		Count of segments with proportion 0 and at or above 0.6
Exposition	*Maximum fraction of overlap with opening material within first half of movement*	*Maximum fraction of overlap with opening material within first half of movement*
	Percentile of maximum fraction of overlap match	*Percentile of maximum fraction of overlap match*
	Fraction of overlap counts at thresholds 0.7, 0.9, and 1	*Fraction of overlap counts at thresholds 0.7, 0.9, and 1*
Development	*Maximum standard deviation over all segments of fixed length*	*Maximum standard deviation over all segments of fixed length*
	Percentile of maximum standard deviation segment	*Percentile of maximum standard deviation segment*
	Count of standard deviations at thresholds 0.7, 0.8, 0.9, and 0.95	*Count of standard deviations at thresholds 0.7, 0.8, 0.9, and 0.95*
Recapitulation	*Maximum fraction of overlap with opening material*	*Maximum fraction of overlap with opening material*
	Percentile of maximum fraction of overlap	*Percentile of maximum fraction of overlap*
	Fraction of overlap counts at thresholds 0.7, 0.9, and 1	*Fraction of overlap counts at thresholds 0.7, 0.9, and 1*

Our novel proposed features are marked with italics.

choose segment lengths $m = 8, 10, 12, 14, 16, 18$ for all segment features. This range of lengths is expected to capture musical motifs in the string quartet genre. Segment features are applied to both pitch and duration tracks.

For pitch, each segment is transposed to either C major or A minor. As mentioned previously, most listeners perceive pitch relatively, rather than in terms of absolute frequency (Levitin & Rogers, 2005). By transposing all segments to a common major or minor key, we can better detect musical phrases that sound the same to most listeners, even if the phrases are in different keys. Since key is perceived by comparing nearby pitches, some of which do not lie perfectly on the diatonic scale, the entire segment is used to transpose the key. Fixing a segment length m, for all ordered segments of such length in a voice of a movement, the segment is transposed with respect to the first note of the segment. For example, suppose a segment is in a major key, and the first note is an A. Then A would be encoded 1, and a C-sharp in the segment would be encoded as 5.

To compare two segments (for duration or pitch), we often calculate the *fraction of overlap*, defined as the proportion of notes in the segment pair that match. In addition, we define the *fraction of overlap count at threshold* t, the number of segment pairs with a fraction of overlap at or above t.

3.2.1 Basic Summary Features

For each voice, we calculate several basic features from the *Alicante* set: the number of notes, mean and standard deviation of the duration of all notes, and mean and standard deviation of the pitch of all notes (De Leon & Inesta, 2007). Similarly to Herlands et al. (2014), we also calculate the proportion of notes and rests played simultaneously by all four voices. These features can indicate whether the voices interact differently in Mozart's versus Haydn's compositions. The interplay of voices is an important consideration in the string quartet genre, famously described by Johann Wolfgang von Goethe in 1829 as "a conversation among four intelligent people" (Klorman, 2016). Although these basic summary features are not the most interesting qualities of music, they may work together with more sophisticated features to help reveal differences between Haydn and Mozart string quartets.

3.2.2 Interval and Rhythm Features

In music, an *interval* refers to the distance between two notes. Intervals have a "special status" in the pitch of music, serving as the basis of the diatonic scale, harmony, and melody (Krumhansl, 2000, p. 165). To calculate intervals, pitch is considered on a full scale from 1 to 132 (with 1 corresponding to the lowest note and 132 to the highest, in chromatic order), since the octave of a note is necessary for this purpose.

Both *pairwise* and *contour* intervals are considered in each voice of a movement:

1. *Pairwise* intervals are defined by each pair of notes, in order. For example, the segment G, A, B, B, B, G has the pairwise intervals G-A, A-B, B-B, B-B, and B-G. These intervals are meant to identify local patterns, summarizing the relationships only between consecutive notes. Intervals defined by successive notes are included in (De Leon & Inesta, 2007) and often have been used for this task, e.g., in (Kaliakatsos-Papakostas et al., 2011), (Herlands et al., 2014), and (Hontanilla et al., 2013).

2. *Contour* intervals are defined by the first note of a segment and each subsequent note in the segment. The example segment from above has the contour intervals G-A, G-B, G-B, G-B, and G-G. More global than the pairwise intervals, contour intervals more effectively capture melodic context. To our knowledge, these intervals have never been used for this task.

With *pairwise* intervals, we compute summary statistics of the following interval aspects of pitch:

- The interval's *type* refers to its distance in semitone on the chromatic scale (equivalently, encoded pitch($mod 12$)). Figure 2 displays the 12 interval types on the C chromatic scale. Summary statistics of interval types are frequently used as features, as in (Kaliakatsos-Papakostas et al., 2011) and (Herlands et al., 2014).

- The *sign* specifies whether the interval is ascending, descending, or constant. For example, if the interval is middle C then the next E above, the interval would be labeled with ascending sign. Interval signs are incorporated in the *Jesser* feature set (Jesser, 1991), among others.

- The interval's *mode* refers to whether it is diminished/augmented, major, minor, or perfect. Summary statistics of nondiatonic intervals are included in the *Alicante* feature set (De Leon & Inesta, 2007) and have been used in (Herlands et al., 2014), (Hillewaere et al., 2010), and (Taminau et al., 2010).

For each interval aspect, our features are the proportions of intervals belonging to each category.

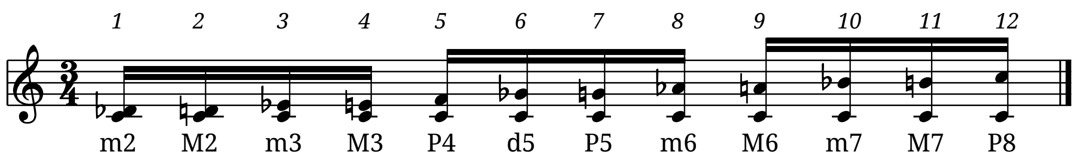

Figure 2: Interval types for chromatic scale in C. Distance in semitone and interval type are printed above and below the staff, respectively. Enharmonic equivalents are represented with the same distances and types.

Fixing a segment length m, *contour* intervals are computed for each segment of pitches in the voice of a movement. Within each segment, the proportion of minor third contour intervals is calculated. The features are summary statistics of the proportions: minimum, first quartile, median, third quartile, maximum, mean, and standard deviation. Many segments contain no minor third intervals, while few segments contain mostly minor third intervals. Therefore, we include as features the count of segments with a low proportion (0) and a high proportion (at or above 0.6). For each voice, 0.6 is approximately the mean (over all movements) of the maximum proportion of minor third intervals.

Emotional response in music listeners is affected by the interval aspects, motivating their use as features. Interval sign has been linked to interval size. Large intervals create discontinuity in the melody, and ascending intervals heighten tension (Vos & Troost, 1989). Therefore, large, ascending intervals are a frequent combination for drama, while small, descending intervals are combined for calm (Vos & Troost, 1989). Meanwhile, perception of happiness or sadness in music is related to mode (Temperley & Tan, 2013). The music scholar Harutunian argues that Haydn exhibits a "keener sense of surface drama" than Mozart (p. 270); in these composers' string quartets, interval type and sign may reveal a difference in surface tension, while interval mode may expose a contrast in "happy" or "sad" sounds.

Minor third intervals are of special interest, contributing significantly to the perception of minor mode and a "sad" sound. Indeed, Temperley and Tan (2013) found that listeners rate melodies containing a minor tonic triad (a type of chord containing a minor third) as sounding less happy than those containing a major tonic triad. The minor third is commonly used when modulating from a major key to a minor key. By tracking minor thirds, we can identify key modulations and offer quantitative evidence for whether Mozart's string quartets are more "emotional" than Haydn's.

Analogous to how intervals refer to differences in pitch, rhythm measures differences in duration between notes. For both pitch and duration, the mean and standard deviation of pairwise interval distances are computed, as in (De Leon & Inesta, 2007). For each pair of voices in a movement, the difference of those pitch interval means and the difference of those pitch interval standard deviations are calculated. In addition, voicepair differences in proportion of interval types are calculated. Voicepair differences are natural generalizations of monophonic features to polyphonic features and have been used in some studies, e.g., (Herlands et al., 2014) and (Van Kranenburg & Backer, 2005). These features, though simple, may reveal tendencies in Haydn's and Mozart's use of intervals and rhythm, particularly across voices.

3.2.3 Exposition Features

The exposition section of a sonata often contains an initial theme, the *opening material*, followed by a secondary theme, the *secondary material*. Occasionally, this convention is broken through *monothematic* expositions. Harutunian claims Haydn's sonatas are more often monothematic than Mozart's sonatas (p. 201, 270), motivating our proposition of exposition features.

To quantify this notion, we search for close repetitions of the opening material within the first half of each voice of a movement. This avoids detection of the recapitulation, which typically witnesses a repetition of the opening theme. Fixing a segment length m, we compare the opening segment to all subsequent segments within the first half of the movement. For all such pairs of segments, we compute the maximum fraction of overlap. We also calculate the percentile (i.e., the ordered location of the segment divided by the total number of segments) corresponding to the segment with maximum fraction of overlap. (If there are multiple segments with the same maximum fraction, then the percentile is defined by the last instance.) The fraction of overlap count is computed for thresholds $0.7, 0.9$, and 1. Besides exact matches (i.e., with threshold 1), segments with a high degree of similarity (i.e., with thresholds at or above 0.7 or 0.9) are of interest, since listeners would likely perceive the segments as sounding approximately the same. These exposition features are calculated for both pitch and duration.

If Haydn is more likely than Mozart to have monothematic expositions, then we would expect his sonatas to yield higher maximum fractions of overlap, percentiles, and threshold counts than Mozart. A fraction of overlap equal to 1 indicates a perfect repetition of the opening material within the exposition, so a high count at threshold 1 suggests one recurring theme. A high percentile may reflect a theme sustained throughout the exposition, corresponding to monothematicism.

3.2.4 Development Features

The exposition section of a sonata leads into the development section, which contains exploration and contrast of the beginning themes. Haydn and Mozart may differ in their development styles: Harutunian asserts that Mozart exhibits more "continuous flow" from the exposition into the development, while Haydn possesses "an immediate formal delineation" between the two sections (p. 199). To identify such differences, we propose features related to musical turbulence.

Capturing variations of thematic material, we search for the area of greatest variability in each voice of a movement. For a fixed segment length m, we compute the standard deviation of notes within each segment of the voice. The maximum of all such standard deviations and its percentile are calculated. (If multiple segments have the same maximum standard deviation, the percentile is determined by the first occurrence.) We also count the number of segments with standard deviations greater than or equal to s. For each segment length and voice combination, we set thresholds for s as the weighted $0.70, 0.80, 0.90$, and 0.95 quantiles of the movements' standard deviations. Accounting for differing movement lengths, we define the weight

$$w_{ijm} = \frac{1}{l_{ijm}},$$

for all movements $i = 1, 2, \ldots, 285$, voices $j = 1, 2, 3, 4$, and segment lengths $m = 8, 10, 12, 14, 16, 18$, where l_{ijm} is the number of segments of length m in voice j of movement i.

If Haydn's developments consist of more "organic construction" and "greater sectionalization" (Harutunian, 2005, p. 273-4), then these aspects may translate to, on average, Haydn string quartets having a higher maximum standard deviation and count. The percentiles represent locations of great change within a movement; differences between Haydn's and Mozart's percentiles may suggest distinct placements of tumultuous material.

3.2.5 Recapitulation Features

In the recapitulation, the material from the exposition is often reiterated. Harutunian claims, "Mozart's recapitulations mirror his expositions far more closely than do Haydn's" (p. 212); his changes are often "ornamental," unlike Haydn's "sweeping changes" (Harutunian, 2005, p. 270). Therefore, we identify the recapitulation and determine how closely it matches the exposition.

Fixing a segment length m, we compare the opening segment to all subsequent segments in the voice of a movement. For each segment, we calculate the fraction of overlap. The maximum fraction of overlap and its associated percentile become our features. (In the case of multiple segments with the same maximal fraction, the percentile is determined by the final occurrence.) The fraction of overlap count at thresholds $0.7, 0.9$, and 1 are computed. Our incentive for choosing these thresholds is similar to that for the exposition thresholds.

The maximum fraction of overlap and counts can measure similarity between the exposition and recapitulation sections. Higher values for these features in Mozart compositions, on average, may verify Mozart's exposition-recapitulation symmetry. The percentile is the location of the last closest repetition of opening material within the voice of the movement; as such, it may indicate differences in Haydn's versus Mozart's approach to concluding a piece.

4 Statistical Methods

Using the musical features from the previous section, we apply statistical methods to analyze the differences between Haydn and Mozart string quartets. In 4.1, we propose our classification model. In 4.2, we discuss feature selection.

4.1 Classification Model

Logistic regression is used as the classification model. Advantages of this model include its ease of interpretation (i.e., the effect of each feature on the composer probability can be clearly explained) and the availability of well-understood inference procedures. We assume the usual additive effects, so that the model is of the form

$$\pi(X) = \frac{e^{\beta_0 + X\beta}}{1 + e^{\beta_0 + X\beta}}, \tag{1}$$

where π is the probability of a movement belonging to the Haydn versus Mozart class, X is the $n \times p$ data matrix containing the n movements and p features, β_0 is the intercept, and β is a $p \times 1$ vector of coefficients for the features. For improved numerical stability in parameter estimation, Bayesian logistic regression is used from (Gelman, Jakulin, Pittau, & Su, 2008). To each coefficient except the intercept, independent Cauchy prior distributions with mean 0 and scale $\frac{2}{2.5S}$ (where S is the standard deviation of the associated feature) are applied; for the intercept, a more conservative Cauchy prior distribution with mean 0 and scale 10 is used. Implementation is provided through the *bayesglm* function from the R package *arm* (Gelman et al., 2016).

To use a logistic regression model for classification, the estimated probabilities $\hat{\pi}(X)$ must be converted to binary classes. In datasets with balanced classes, it is customary to use 0.5 as the cutoff: assign observations with greater than 0.5 (estimated) probability to one class, and the remaining observations to the other class. In a dataset with imbalanced classes, a cutoff 0.5 may not be optimal for classification accuracy; instead the cutoff may be treated as a tuning parameter, which is a type of approach that has been explored for such binary classification problems (Zou, Xie, Lin, Wu, & Ju, 2016). We test the sequence of cutoff values $0, 0.01, 0.02, \ldots, 0.98, 0.99, 1$ and choose the "best" cutoff value as the one that maximizes classification accuracy within the training data.

Since the total number of proposed features 1115 exceeds the number of observations ($n = 285$), logistic regression cannot be applied directly to the full feature set. Further, a fitted logistic regression model that contains a large number of features could be difficult to interpret and suffer from reduced classification accuracy due to overfitting. In particular, highly collinear features would not have meaningful interpretations for their estimated coefficients, due to inflated standard errors: for example, each sonata-style feature is computed 6 times, for segment lengths $m = 8, 10, 12, 14, 16, 18$, and these are strongly correlated amongst themselves. Thus, we perform feature selection before fitting the logistic regression model. Intuitively, the important musical differences between Haydn and Mozart string quartets might be expressed in a concise subset of features.

4.2 Feature Selection

The goal of feature selection is to determine the appropriate features to include in the final model. From a practical perspective, feature selection helps identify a succinct subset of variables representing meaningful differences between Haydn and Mozart string quartets. There are many feature selection approaches from the statistical and machine learning literature, including methods that transform the features to reduce their dimensionality (e.g., factor analysis, principal component analysis, and discriminant analysis) and algorithms to search for optimal subsets of variables (e.g., stepwise regression) (Guyon & Elisseeff, 2003). Our proposed features have musical meaning that would be lost in a transformation, so the latter category of feature selection methods is more pertinent.

For very high-dimensional problems, a two-scale feature selection process is commonly adopted: (i) a crude large scale screening followed by (ii) a moderate scale selection (Fan & Lv, 2010, 2008). The goal of (i) is to quickly, efficiently, and substantially reduce the feature set; in (ii) a more traditional feature selection method can be applied to the reduced feature set. Potential advantages of this process are reduced computational cost, improved accuracy, and model sparsity (Fan & Lv, 2010), motivating its application to our very high-dimensional feature set.

For the large scale screening method, we choose correlation ranking, a well-accepted approach that ranks features by magnitude of correlation with the response (Fan & Lv, 2010). Several other Haydn-Mozart classification studies have also involved correlation ranking (Hillewaere et al., 2010; Herlands et al., 2014). Intuitively, music features that independently have strong correlations with the composer might be good predictors. After ranking our features by magnitude of correlation with the composer, we choose a subset of them that have low pairwise correlations among themselves. At the first step, we remove any variables that are strongly correlated (Pearson's $R \geq 0.5$) with the first ranked variable. At the second step, if the second ranked variable has not already been excluded, we remove any variables that are strongly correlated with it. The algorithm continues until the last ranked variable has been considered. While the algorithm prioritizes variables strongly correlated with the composer, variables weakly correlated with the composer are not barred from inclusion in the feature set, since they could be significant in a multiple regression model (as in Equation (1)).

After coarsely reducing the feature set from 1115 variables to dimension $d \leq 1115$, we conduct moderate scale selection. Here, moderate scale selection specifically involves determining which of the d features should be included as predictors to yield the "best" logistic regression model in Equation (1). For any given subset of the d reduced features, the fitted logistic regression model is used to compute the Bayesian information criterion (BIC) (Schwarz, 1978), which may be expressed here as

$$BIC = -2\mathcal{L} + 2(p+1)\log(n), \tag{2}$$

where n is the number of observations in the dataset, p is the number of features included in the model, and $\mathcal{L}$ is the maximized value of the log-likelihood of the model fitted with those features. We adopt BIC, as it is a standard criterion used for model selection in statistics; here then, the subset of features that leads to the lowest BIC value in the fitted model would be considered the "best" subset of features.

However, it is not computationally feasible to exhaustively test all possible subsets of features to find the one with the lowest BIC; we note there are on the order of 2^d such combinations for our reduced feature set. In practice then, one can only test a limited number of subsets and choose the model with the lowest BIC value found. We use the method of Iterative Conditional Minimization (ICM) to search for the minimum BIC, which is discussed in Zhang, Lin, Liu, and Chen (2007) as a simple but substantively more effective alternative to stepwise regression methods.

We summarize ICM as applied here. First, define V to be an empty subset. Variables will be iteratively added to V, representing the best subset of features found thus far. When a logistic regression model is fit with all variables in V as predictors of composer, denote the resulting BIC as BIC_V. The algorithm is presented in pseudocode as follows:

Initialize:

1. Set V to be an empty subset and $\text{BIC}_V = +\infty$.
2. Randomly order the d features (from the reduced feature set) as $1, 2, \ldots, d$.

For j in $1, 2, \ldots, d$:

1. **If** x_j is not in V, **then**
   (a) Fit a logistic regression model with predictors x_j and all variables from V.
   (b) **If** the BIC from the fitted model is less than BIC_V, **then** add x_j to V.
2. **Else if** x_j is in V, **then**
   (a) Fit a logistic regression model to predict composer from all variables in V, excluding x_j.
   (b) **If** the BIC from the fitted model is less than BIC_V, **then** remove x_j from V.

Repeat **For** loop until two successive passes yield no further additions or deletions of variables.

Observe that the final subset of features will depend on the order in which the d features are tested, which is randomized when initializing the algorithm. Thus, in practice we may run this algorithm repeatedly with different random seeds and select the lowest BIC model among the repetitions.

5 Results and Discussion

In 5.1, we present our cross-validated composer classification results from applying the statistical methods from Section 4; our results are then compared to prior studies. In 5.2, we summarize one comprehensive model of composer. In 5.3, we discuss the musical meaning and insights gained from our model.

5.1 Accuracy Comparisons with Previous Studies

Following the statistical approach outlined in the previous section, we classify the composer of Haydn and Mozart string quartets. The resulting classification accuracy can quantify the extent to which these compositions can be discriminated.

As highlighted in (Hillewaere et al., 2010), the dataset is small, motivating the use of cross-validation (CV) to assess classification ability. We choose the leave-one-out (LOO) CV approach, which estimates the "true" classification accuracy for unseen observations using observations from the dataset. In past studies, LOO was the most common CV approach for the Haydn-Mozart string quartet task, so our use of LOO facilitates comparison. The evaluation criterion is the classification accuracy.

For each movement $i = 1, 2, \ldots, 285$, we form training fold i and testing fold i. Training fold i contains all movements except i, and testing fold i contains only movement i. Feature selection and model fitting are performed on training fold i, and the composer is subsequently predicted for movement i. That is, within training fold i, the feature set is reduced to size d using our modified correlation ranking; random ICM is run for ten random seeds using the reduced feature set; the "best" model is chosen as the one achieving minimal BIC out of the ten; and the "best" classification cutoff is selected as the one that maximizes training classification accuracy. That "best" model and cutoff from training fold i are subsequently used to classify the composer of movement i in testing fold i. The LOO classification accuracy is computed here as the proportion of movements correctly classified in their respective testing folds. Hence, our approach involves fitting and predicting from 285 models, one per fold.

The robustness of our approach is evidenced by the similarities seen across the 285 folds. After applying the modified correlation ranking to each training fold, in all cases the reduced feature set contains between 210 to 217 variables. The variables are well-represented in the reduced feature sets, with all feature categories (basic, interval, exposition, development, and recapitulation) present. The subsets of variables selected from random ICM are also very similar across the folds. In total, there are 25 variables selected in at least one fold of LOO, summarized in Figure 3. As listed in Table 2, nine of those variables are included in more than 200 folds, with five included in all 285 folds. These commonly selected variables, together with the composition of the reduced feature sets, confirm the stability of our methods across LOO folds.

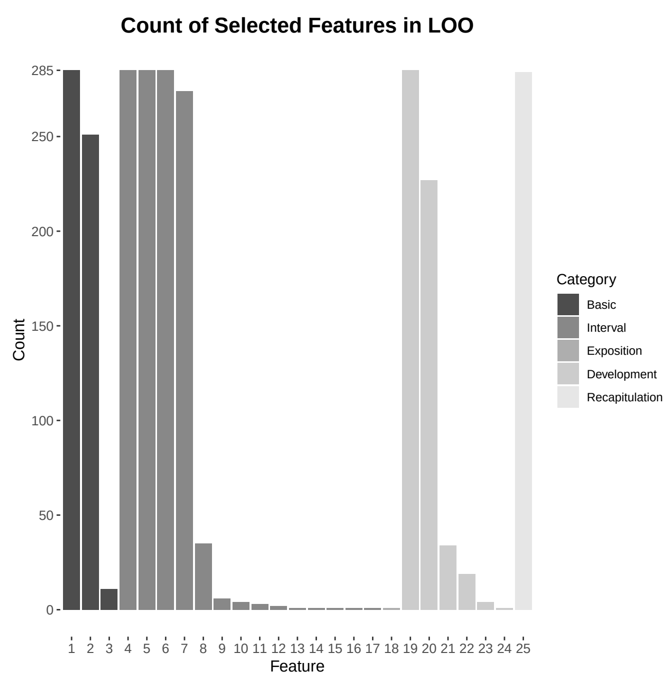

Figure 3: We plot the number of times a variable (out of the 25 variables represented in at least one fold) is selected in LOO.

Our approach achieves 85.26% LOO classification accuracy, higher than attained in prior studies. The LOO Haydn accuracy, which we define here as the proportion of Haydn movements correctly classified in LOO, is very high at 92.12%. The LOO Mozart accuracy, defined in an analogous manner, is 68.29%. A lower accuracy rate for the Mozart class is not surprising, since there are more than twice as many Haydn as Mozart movements in our dataset. Still, as summarized in

Table 2: Commonly Selected Features in LOO

Index	Category	Feature	Count
1	Basic	Standard deviation of duration for Violin 1	285
2	Basic	Standard deviation of pitch for Viola	251
4	Interval	Proportion of pairwise descending intervals for Violin 1	285
5	Interval	Proportion of pairwise intervals with semitone distance 3 for Cello	285
6	Interval	Mean proportion of minor third intervals for $m = 18$ and Viola	285
7	Interval	Mean proportion of minor third intervals for $m = 8$ and Cello	274
19	Development	Standard deviation count at threshold $4.024^{(B)}$ for pitch, $m = 8$, and Cello	285
20	Development	Standard deviation count at threshold $4.829^{(A)}$ for pitch, $m = 16$, and Viola	227
25	Recapitulation	Maximum fraction of overlap for duration, $m = 8$, and Viola	284

We present the features that are represented in 200 or more LOO folds. The index corresponds to the feature label in Figure 3. For the development features, the thresholds are the following weighted quantiles: $(A)0.95$ and $(B)0.70$.

Table 3, relatively few movements are misclassified overall: 26 Mozart and 16 Haydn. The fitted probabilities from LOO are plotted in Figure 4. Generally, the composers are well-separated, with many probabilities clustering around 0 or 1. The well-separated probabilities and high accuracy suggest discernible differences between Mozart and Haydn string quartets, despite human listeners' struggles to detect them.

To understand the role of our novel musical features introduced in Section 3, we repeat the analysis on a "simple" subset of the 1115 features. This subset contains only the basic and interval features and excludes the sonata-style features (exposition, development, and recapitulation). The resulting LOO classification accuracy is 80.35%, while the LOO Haydn and Mozart accuracy rates are about 90.64% and 54.88%, respectively. This represents close to a 5% decrease in classification accuracy compared to using all features. While the LOO Haydn accuracy only drops by about 1.5%, the LOO Mozart accuracy drops by over 13%. Such changes in results emphasize the importance of our musically sophisticated features based on the sonata form, particularly for detecting Mozart's movements.

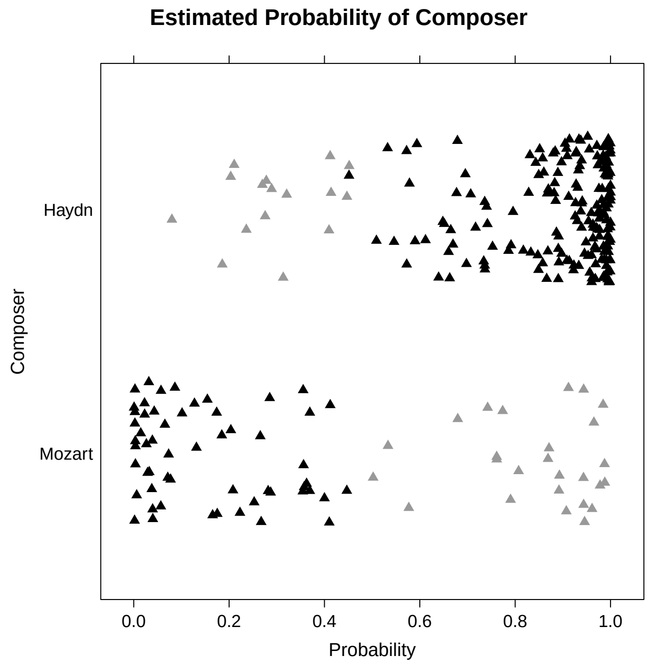

Figure 4: For each movement, the true composer is displayed on the vertical axis (with added jitter for visual readability), and the fitted probability of composer is on the horizontal axis. Gray markings indicate misclassified movements, while black markings correspond to correctly classified movements.

Table 3: Confusion Matrix for Haydn-Mozart String Quartet Classification with Bayesian Logistic Regression and LOO CV

	Predicted Mozart	Predicted Haydn
Observed Mozart	56	26
Observed Haydn	16	187

Our results are compared to all existing studies (of which we are aware). In Table 4, the methods and results from each study are reported. In the pioneering work, the authors implement a 3-nearest neighbor classifier with 20 lower-level musical features transformed through Fisher's discriminant analysis (Van Kranenburg & Backer, 2005). Next, Hillewaere et al. (2010), Hontanilla et al. (2013), and Kaliakatsos-Papakostas et al. (2011) use n-gram (or $(n-1)$th order Markov models) from language analysis, modeling the probability of a musical event given the context of past musical events. Herlands et al. (2014) classify composer with either linear SVM or the Naive Bayes classifier. As discussed before, Velarde et al. (2016) attain the highest accuracy prior to our study, but they use a computer vision approach that is difficult to interpret musically: they apply a Gaussian filter to images of piano roll scores, transform the resulting pixel data through linear discriminant analysis, and classify with a linear SVM. In follow-up work, Velarde, Cancino Chacón, Meredith, Weyde, and Grachten (2018) extend their approach to include image analysis of spectrograms, as well as classification with a k-nearest neighbour classifier; however, as before, their study differs in scope from our musicological investigation. Finally, Taminau et al. (2010) deploy subgroup discovery, a descriptive rule learning technique that involves both predictive and descriptive induction. Evidently, a diverse range of approaches have been applied to the Haydn-Mozart classification problem.

We have achieved 85.26% LOO accuracy for almost all known Haydn and Mozart string quartet movements. The previous benchmark of 80.4% was set by (Velarde et al., 2016), and our LOO accuracy is almost 5% higher. When we used only basic and interval features, our LOO accuracy was about 80.35% and very close to the previous benchmark. These results support the importance of musically meaningful features for this task. We conclude there are significant musical differences between Haydn and Mozart string quartets, enabling less than 15% LOO error and the selection of similar models across folds.

Table 4: Comparing Accuracy Rates and Methods with Prior Studies

Method	Cross-validation (CV)	Accuracy
FDA + k-means clustering (Van Kran. et al., 2005)	LOO	0.7944
3-grams model (only Cello) (Hillewaere et al., 2010)	LOO	0.754
Weighted-Markov chain model + SVM (Kal. et al., 2011)	30 simulations	0.70
Linear SVM or Naive Bayes (Herlands et al., 2014)	CV trials	0.80
3-grams model (Hontanilla et al., 2013)	LOO	0.747
LDA + Linear SVM (Velarde et al., 2016)	LOO	0.804
KNN + SVM ensemble (Velarde et al., 2018)	LOO	0.748
Subgroup discovery (Taminau et al., 2010)	LOO	0.730
Bayesian Logistic Regression (ours)	LOO	**0.8526**

5.2 Model of Composer on Musical Features

Having demonstrated classification accuracy of our approach via leave-one-out cross-validation, we now fit a single descriptive model of composer to the full dataset. The statistical methods of Section 4 are applied to all 285 movements together, and we summarize the resulting model in what follows.

For each variable j (including the intercept) in the model, the estimated effect $\hat{\beta}_j$ and its standard error $\sqrt{\hat{Var}(\hat{\beta}_j)}$ are given in Table 5. Each effect corresponds to a change in probability of composer, controlling for all other variables in the model. For effects with positive sign, increases in the predictor correspond to a greater probability the movement is composed by Haydn, adjusting for other model variables. For example, $\hat{\beta}_4 = 65.23$, so Haydn is more likely than Mozart to have higher mean proportions of minor third intervals in the viola voice, controlling for the other variables. In contrast, we interpret predictors with negative effects as negatively associated with Haydn. For example, Haydn movements are less likely than Mozart movements to have high proportions of minor third intervals in the cello voice (since $\hat{\beta}_7 = -46.19$), adjusting for other variables.

By the assumption of additivity, an effect is constant for each value of the feature, even as other features' values change. The effect of each variable (including, trivially, the intercept) on composer is tested by the hypotheses

$$\begin{aligned} H_0 &: \beta_j = 0 \text{ when all other variables are in the model} \\ H_A &: \beta_j \neq 0 \text{ when all other variables are in the model,} \end{aligned} \tag{3}$$

for $j = 1, \ldots, 10$. The Wald p-values for these tests are listed in Table 5. For each coefficient, the p-value is less than 0.02. Strongly significant p-values are a natural consequence of the use of BIC as the model selection criterion. For example, the "standard deviation counts at thresholds 4.829 and 4.024" have p-values below 10^{-6}, indicating these counts are significant predictors of composer.

Most commonly, a logistic regression model's goodness of fit is assessed through deviance, a generalization of analysis of variance (Nelder & Baker, 2004). Here, the deviance would compare the maximized log-likelihood for the fitted model and for the saturated model (which contains as many parameters as observations). This is handled in our case by using BIC for variable selection, since BIC is a function of the maximized log-likelihood of the fitted model. Tests based on residuals can also be used, and here we apply the Hosmer-Lemeshow test.

In the Hosmer-Lemeshow test (Hosmer & Lemeshow, 1980), the estimated probabilities from the model are divided into g groups, in which the observed outcomes are compared to the expected outcomes from the model. When the model fits the data well and g is chosen such that $g > p + 1$, the test statistic has an approximate χ^2 distribution. We test values of g ranging from 20 to 100. All tests yield p-values greater than 0.1, and the median p-value is 0.9242. With generally large p-values over g, there is no significant evidence of lack of fit.

We note that the 9 predictors selected for the full model (listed in Table 5) have been consistently selected, with or without cross-validation. Indeed, these predictors are exactly the same variables that appeared 200 or more times in LOO (in Table 2 of Section 5.1). Thus, the same sparse subset of musical variables emerges, even when applying our approach to different groups of movements. This consistency strongly substantiates the robustness of our approach.

Table 5: Additive Bayesian Logistic Regression Model of Composer on Musical Features

Category	Feature	$\hat{\beta}_j$	$\sqrt{\hat{Var}(\hat{\beta}_j)}$	p-value
	(Intercept)	8.98	3.89	0.0209
Development	Standard deviation count at threshold $4.829^{(A)}$ for pitch, $m = 16$, and Viola	0.15	0.03	2.09×10^{-8}
Development	Standard deviation count at threshold $4.024^{(B)}$ for pitch, $m = 8$, and Cello	-0.02	0.0040	2.06×10^{-7}
Interval	Mean proportion of minor third intervals for $m = 18$ and Viola	65.23	13.06	5.87×10^{-7}
Interval	Proportion of descending pairwise intervals for Violin 1	16.04	4.15	0.00011
Interval	Proportion of pairwise intervals with semitone distance 3 for Cello	24.21	6.12	7.53×10^{-5}
Interval	Mean proportion of minor third intervals for $m = 8$ and Cello	-46.19	11.27	4.13×10^{-5}
Basic	Standard deviation of duration for Violin 1	-21.78	5.31	4.13×10^{-5}
Recapitulation	Maximum fraction of overlap for duration, $m = 8$, and Viola	-5.76	2.03	0.0047
Basic	Standard deviation of pitch for Viola	-2.19	0.74	0.0033

For the development features, the thresholds are the following weighted quantiles: $(A)0.95$ and $(B)0.70$.

5.3 Musical Interpretation

We now provide musical interpretations of the differences between Haydn and Mozart string quartets, based on the features from the model in the previous section. That model was identified from feature selection as a "good" discriminator of composer, suggesting differences in the features for Haydn versus Mozart. The features were also frequently selected in LOO CV in Section 5.1, further motivating their analysis. Results for the sonata-style features generally agree with the music scholar Harutunian's claims regarding Haydn's versus Mozart's sonata styles. The inclusion of other variables in the model yield additional insights into these composers' string quartets. We discuss these features in detail, starting with basic and interval features, followed by sonata-style features. The distribution of each variable is plotted by composer in Figure 5.

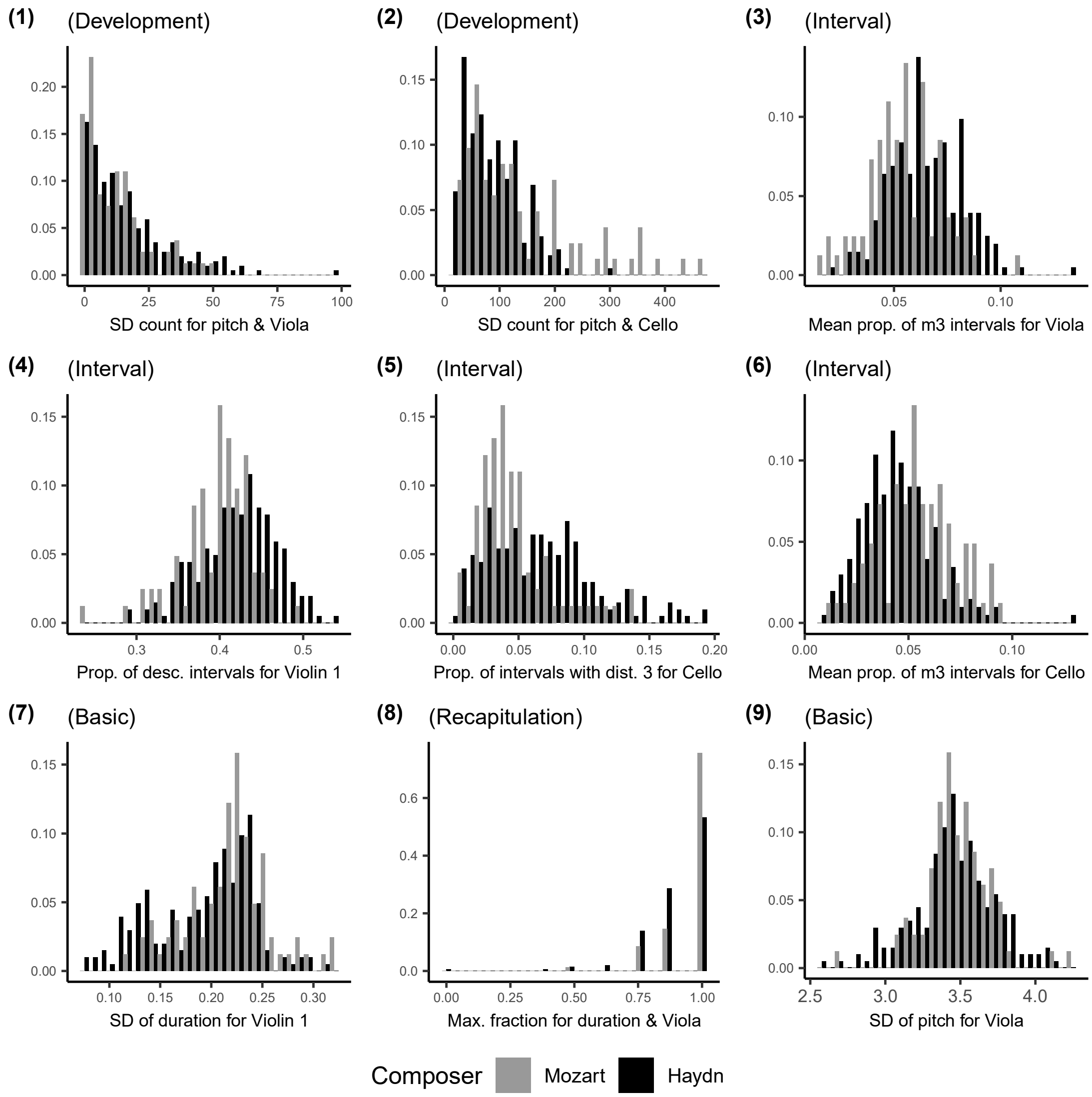

Figure 5: Side-by-side relative frequency plots by composer for the variables selected in the final model.

For the basic features, the standard deviation of duration for Violin 1 tends to be higher for Mozart (Figure 5 (panel 7)), as well as the standard deviation of pitch for Viola (panel 9). For intervals, several differences between these composers are identified. First, higher proportions of descending pairwise intervals in the first violin are associated with Haydn, rather than Mozart (panel 4), confirming differences between these composers' "surface tension". For minor third intervals, Haydn tends to have a higher mean proportion of pairwise intervals in Viola (panel 3), while Mozart has a higher mean proportion in Cello (panel 6). Additionally, Haydn is associated with a higher proportion of pairwise intervals with semitone distance 3 in the Cello (panel 5). It is interesting that out of all possible distances, semitone distance 3 was selected; perhaps this feature is another measure of minor third intervals. The inclusion of several minor third interval features in the model suggest distinctions in emotionalism between Haydn and Mozart string quartets.

Many interpretations of the sonata-style features in the model align with Harutunian's assertions regarding Haydn and Mozart sonatas, while others are less conclusive. In Figure 5 (panel 1), Haydn's greater standard deviation counts at a high threshold (the 0.95 quantile) may confirm his "organic construction" in the development. In contrast, panel 2 displays higher standard deviation counts for Mozart than Haydn, suggesting the importance of threshold choice and segment length. The recapitulation maximum fraction of overlap for duration in panel 8 is more often 1 for Mozart than Haydn in the viola voice, which may validate Mozart's greater exposition-recapitulation similarity.

Features are present in all categories (except the exposition), indicating the importance of features ranging from basic to sophisticated. This also suggests most statistical differences between Haydn's and Mozart's string quartets can be explained by the basic, interval, recapitulation, and development features. Of these categories, some are more commonly represented than others. For example, there are four interval features, while only two basic features. The high count of interval features is expected, because of the fundamental role intervals serve in music. The sonata-style features include two from the development and one from the recapitulation. Both of the development features are "standard deviation counts", implying important differences between Haydn and Mozart in the extent of thematic material variation.

The counts of features from each voice can describe distinctions between the composers' handling of the voices in the string quartet. There are two features from Violin 1, none from Violin 2, three from Cello, and four from Viola. Contrary to our expectations, the "leading" violins account for only two features, while the lower accompanying voices (Cello and Viola) number seven features. These surprising results suggest that Mozart and Haydn handle their low accompanying voices differently, while their violin parts are more similar. The inclusion of nine monophonic features and no polyphonic features indicates that Mozart and Haydn may connect the string quartet voices together in a similar way but treat individual voices distinctly.

Features from pitch tracks outnumber features from duration tracks: there are seven from pitch, while only two from duration. One explanation is that the role of pitch is more prominent than rhythm in Classical Western music. Indeed, in a study with Western musical excerpts, Schellenberg, Krysciak, and Campbell (2000) found that pitch is more emotionally meaningful to listeners than rhythm.

6 Conclusion

We have conducted a quantitative analysis of Mozart versus Haydn string quartets, contributing to the vast musical scholarship of these composers. We proposed many novel summary features that are musically meaningful and related to the sonata form. Feature selection identified 9 important features, several of which pertained to the sonata structure. Features from the cello and viola voices were selected more often than ones from the violins, suggesting that Haydn and Mozart use their low accompanying voices distinctly. The 9 features were also commonly selected in leave-one-out (LOO) cross-validation, further signifying their importance and validating the robustness of our approach. Our Bayesian logistic regression models containing musical features achieved state-of-the-art classification accuracy: over 85% for LOO on almost all known movements. These strong results indicate that Haydn and Mozart string quartets can be discriminated with high accuracy without sacrificing musical interpretability.

Further directions for the Haydn-Mozart classification task involve prediction of ambiguously authored movements and a study of the "Haydn" movements. Some movements excluded from our study have spurious authorship, and our models could determine the probability those movements were authored by Mozart or Haydn. Another topic of interest to historians is the similarity of "Haydn" movements (the set of Op. 10 string quartets composed by Mozart) to movements composed by Haydn, which could be assessed quantitatively.

Beyond the Mozart-Haydn string quartet classification task, our sonata-style features can represent any music roughly following the A-B-A structure. Our interdisciplinary approach has prioritized both musical interpretability and quantitative validity. We recommend a similar framework be applied to other studies in MIR, so that these works can be fully appreciated both musically and mathematically. As the mathematician James Joseph Sylvester (1908) famously wrote, "May not music be described as mathematics of the sense, mathematics as music of the reason?".

References

De Leon, P. J. P., & Inesta, J. M. (2007). Pattern recognition approach for music style identification using shallow statistical descriptors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *37*(2), 248–257.

Downie, J. S. (2003). Music information retrieval. *Annual review of information science and technology*, *37*(1), 295–340.

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(5), 849–911.

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., & Zheng, T. (2016). arm: Data analysis using regression and multilevel/hierarchical models. 2015. *URL https://CRAN. R-project. org/package= arm. R package version*, 1–9.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157–1182.

Harutunian, J. M. (2005). *Haydn's and mozart's sonata styles: a comparison* (Vol. 113). Edwin Mellen Pr.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction*. Springer Heidelberg.

Herlands, W., Der, R., Greenberg, Y., & Levin, S. A. (2014). A machine learning approach to musically meaningful homogeneous style classification. In *Aaai* (pp. 276–282).

Hillewaere, R., Manderick, B., & Conklin, D. (2010). String quartet classification with monophonic models. In *Ismir* (pp. 537–542).

Hontanilla, M., Pérez-Sancho, C., & Inesta, J. M. (2013). Modeling musical style with language models for composer recognition. In *Iberian conference on pattern recognition and image analysis* (pp. 740–748).

Hosmer, D. W., & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, *9*(10), 1043–1069.

Jesser, B. (1991). Interaktive melodieanalyse. *Peter Lang, Bern*, 258.

Johanson, B., & Poli, R. (1998). *Gp-music: An interactive genetic programming system for music generation with automated fitness raters*. University of Birmingham, Cognitive Science Research Centre.

Kaliakatsos-Papakostas, M. A., Epitropakis, M. G., & Vrahatis, M. N. (2011). Weighted markov chain model for musical composer identification. In *European conference on the applications of evolutionary computation* (pp. 334–343).

Klorman, E. (2016). *Mozart's music of friends: Social interplay in the chamber works.* Cambridge University Press.

Kosugi, N., Nishihara, Y., Sakata, T., Yamamuro, M., & Kushima, K. (2000). A practical query-by-humming system for a large music database. In *Proceedings of the eighth acm international conference on multimedia* (pp. 333–342).

Krumhansl, C. L. (2000). Rhythm and pitch in music cognition. *Psychological bulletin*, *126*(1), 159.

Larsen, J. P., & Feder, G. (1997). *The new grove haydn.* WW Norton & Company.

Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *Machine learning and applications, 2008. icmla'08. seventh international conference on* (pp. 688–693).

Levitin, D. J., & Rogers, S. E. (2005). Absolute pitch: perception, coding, and controversies. *Trends in cognitive sciences*, *9*(1), 26–33.

Mearns, L., Tidhar, D., & Dixon, S. (2010). Characterisation of composer style using high-level musical features. In *Proceedings of 3rd international workshop on machine learning and music* (pp. 37–40).

Mirka, D. (2009). *Metric manipulations in haydn and mozart: Chamber music for strings, 1787-1791.* Oxford University Press on Demand.

Nelder, J. A., & Baker, R. J. (2004). Generalized linear models. *Encyclopedia of statistical sciences*, *4*.

Pollastri, E., & Simoncelli, G. (2001). Classification of melodies by composer with hidden markov models. In *Web delivering of music, 2001. proceedings. first international conference on* (pp. 88–95).

R Core Team. (2017). *R: A language and environment for statistical computing. vienna, austria: R foundation for statistical computing; 2016.*

Schedl, M., Gómez, E., & Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, *8*(2-3), 127–261.

Schellenberg, E. G., Krysciak, A. M., & Campbell, R. J. (2000). Perceiving emotion in melody: Interactive effects of pitch and rhythm. *Music Perception: An Interdisciplinary Journal*, *18*(2), 155–171.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Selfridge-Field, E. (1997). *Beyond midi: the handbook of musical codes.* MIT press.

Taminau, J., Hillewaere, R., Meganck, S., Conklin, D., Nowé, A., & Manderick, B. (2010). Applying subgroup discovery for the analysis of string quartet movements. In *Proceedings of 3rd international workshop on machine learning and music* (pp. 29–32).

Temperley, D., & Tan, D. (2013). Emotional connotations of diatonic modes. *Music Perception: An Interdisciplinary Journal*, *30*(3), 237–257.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, *10*(5), 293–302.

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In *Advances in neural information processing systems* (pp. 2643–2651).

Van Kranenburg, P., & Backer, E. (2005). Musical style recognition—a quantitative approach. In *Handbook of pattern recognition and computer vision* (pp. 583–600). World Scientific.

Velarde, G., Cancino Chacón, C., Meredith, D., Weyde, T., & Grachten, M. (2018). Convolution-based classification of audio and symbolic representations of music. *Journal of New Music Research*, *47*(3), 191–205.

Velarde, G., Weyde, T., Chacón, C. E. C., Meredith, D., & Grachten, M. (2016). Composer recognition based on 2d-filtered piano-rolls. In *Ismir* (pp. 115–121).

Vos, P. G., & Troost, J. M. (1989). Ascending and descending melodic intervals: Statistical findings and their perceptual relevance. *Music Perception: An Interdisciplinary Journal*, *6*(4), 383–396.

Zaslaw, N. (1990). *The compleat mozart: a guide to the musical works of wolfgang amadeus mozart*. WW Norton & Company.

Zhang, J. L., Lin, M. T., Liu, J. S., & Chen, R. (2007). Lookahead and piloting strategies for variable selection. *Statistica Sinica*, 985–1003.

Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. *Big Data Research*, *5*, 2–8.