

# HIERARCHICAL MULTIDIMENSIONAL SCALING FOR THE COMPARISON OF MUSICAL PERFORMANCE STYLES

BY ANNA K. YANCHENKO\* AND PETER D. HOFF†

*Department of Statistical Science, Duke University, \*[anna.yanchenko@duke.edu](mailto:anna.yanchenko@duke.edu); †[peter.hoff@duke.edu](mailto:peter.hoff@duke.edu)*

Quantification of stylistic differences between musical artists is of academic interest to the music community and is also useful for other applications, such as music information retrieval and recommendation systems. Information about stylistic differences can be obtained by comparing the performances of different artists across common musical pieces. In this article we develop a statistical methodology for identifying and quantifying systematic stylistic differences among artists that are consistent across audio recordings of a common set of pieces, in terms of several musical features. Our focus is on a comparison of 10 different orchestras, based on data from audio recordings of the nine Beethoven symphonies. As generative or fully parametric models of raw audio data can be highly complex and more complex than necessary for our goal of identifying differences between orchestras, we propose to reduce the data from a set of audio recordings down to pairwise distances between orchestras, based on different musical characteristics of the recordings, such as tempo, dynamics and timbre. For each of these characteristics, we obtain multiple pairwise distance matrices, one for each movement of each symphony. We develop a hierarchical multidimensional scaling (HMDS) model to identify and quantify systematic differences between orchestras in terms of these three musical characteristics and interpret the results in the context of known qualitative information about the orchestras. This methodology is able to recover several expected systematic similarities between orchestras as well as to identify some more novel results. For example, we find that modern recordings exhibit a high degree of similarity to each other, as compared to older recordings.

**1. Introduction.** The quantification of stylistic differences between musical artists is of interest in musicology and has uses for the general music-listening public, such as for music information retrieval and recommender systems. In this article we are particularly focused on the quantification of systematic differences among 10 different orchestras, based on data from audio recordings of the nine Beethoven symphonies. This is motivated by a desire to statistically quantify a variety of descriptions of heterogeneity among orchestras that has typically been done qualitatively, such as how musical performances might change over time or how orchestras from the United States systematically differ from European orchestras.

In general, information about stylistic differences among artists can be obtained by comparing their performances of a common collection of musical pieces. Quantitative analyses of different orchestral recordings has been explored in the music information retrieval community using tempo curve analysis (Peperkamp, Hildebrandt and Liem (2017)) and image analysis techniques with principal components analysis (Liem and Hanjalic (2015)). A more statistical approach to making such comparisons would be to fit a model to the audio data for each artist separately and then compare the estimated model parameters corresponding to each artist. However, from a data analysis perspective an audio recording of a piece of music

---

Received April 2020; revised August 2020.

*Key words and phrases.* Audio processing, Bayes, hierarchical modeling, functional data, multidimensional scaling.

is a complex, multivariate, highly structured time series with long-term dependencies that, in symphonic works, exist over multiple minutes.

Comparison is a critical question of interest for musicologists (Cook (2005)), and a quantitative comparison of systematic similarities across performances is of interest to the musicology and music information retrieval communities in several regards. First, such an approach allows for the quantitative comparison of a large corpus of musical pieces for a recommendation system (van den Oord, Dieleman and Schrauwen (2013)) or music corpus analysis task (Liem and Hanjalic (2015)) in the music information retrieval context. Second, such an approach is of interest for musicologists as well (Cook (2005)) since it allows for numerically exploring, and testing hypotheses of interest, such as whether more recent recordings exhibit less variation than older recordings (Liebman, Ornoy and Chor (2012)), a primary question of interest in this work as well. Finally, this type of analysis can serve as a starting point for insight into understanding differences in expressive playing from a psychological and cognition perspective (Desain and Honing (1994), Goebel and Widmer (2009), Liem and Hanjalic (2011a), Liem and Hanjalic (2011b), Penel and Drake (1998)). While the complexity of audio data allows for analysis of more nuanced features, such as musical style, this complexity also makes it challenging to develop accurate generative statistical models of musical audio data in its raw form. One popular approach is to use convolutional neural networks with dilated, causal convolutions (van den Oord et al. (2016)), and, while these can generate audio that mimics their input, the large number of estimated parameters can be difficult to interpret and, potentially, noncomparable across model fits. Additionally, WaveNet (van den Oord et al. (2016)) models very short audio segments of only a few seconds in length, making stylistic analysis over full orchestral works that are multiple minutes in length challenging.

As an alternative to such generative approaches, for the purpose of identifying differences between orchestras we propose to reduce the data from a set of audio recordings down to pairwise distances between orchestras, based on different musical characteristics of the recordings, such as tempo, volume dynamics and timbre. For each of these characteristics, we obtain multiple pairwise distance matrices, one for each movement of each symphony, resulting in 37 distance matrices for each of the three musical characteristics. Comparison of the orchestras may then proceed using statistical methods appropriate for analysis of distance data. Such methods might include distance-based analysis of variance (ANOVA) approaches used in the ecological community (Anderson (2001), McArdle and Anderson (2001), Minas and Montana (2014), Rizzo and Székely (2010)) or the related functional ANOVA (FANOVA) approach that was developed to analyze distance data in genomics with Gaussian processes (Vsevolozhskaya et al. (2014)). In this article we focus on adapting multidimensional scaling (MDS) to the specific task of combining information across multiple distance matrices to identify consistent differences between orchestras. MDS is a popular technique for analyzing distance data, originally developed in the psychology literature (Torgerson (1952)). Standard MDS generates an embedding of observed distance data into a Euclidean space so that the distances between objects in the embedding approximates the observed distances.

Our HMDS model can be viewed as a modification and extension of the Bayesian MDS model proposed by Oh and Raftery (2001). In Bayesian MDS (BMDS) the observed distance matrix is assumed to be equal to the distance matrix of a set of latent vectors in a Euclidean space, plus (truncated) Gaussian noise. Using a Markov chain Monte Carlo approximation algorithm, the posterior distribution of the latent vectors may be inferred from an observed matrix of distance data. Our proposed HMDS extends the BMDS model of Oh and Raftery (2001) in several ways to accommodate specific features of our data. Most importantly, BMDS was developed to analyze a single distance matrix and assumes that the “true” distances are Euclidean. In contrast, our data consist of 37 distance values for each pair of orchestras and each musical characteristic. Treating these 37 values as “replicates,”

our HMDS model is able to distinguish between differences that are “systematic,” that is, consistent across musical pieces, and differences that are idiosyncratic to particular pieces. Furthermore, our approach allows for the systematic component of the observed distance matrices to be non-Euclidean. This is useful, as the distance metrics we use to evaluate stylistic differences are not necessarily embeddable in Euclidean space. Another feature of our model is that we allow for differences in the potential for variation across replicates or pieces. This is critical for our application, as some musical pieces have much more potential for variation than others. Combining information across pieces without adjusting for this potential would tend to hide systematic effects. Finally, in contrast to the truncated normal model in BMDS, we model nonnegative distances using gamma distributions. This approach has the advantage of being able to accommodate skew in the distribution of observed distances and is perhaps a more natural choice for positive distance data. Additionally, our gamma model for observed distances permits the use of semiconjugate prior distributions which facilitates several posterior calculations.

Related to our approach, Park, DeSarbo and Liechty (2008) extended the BMDS model to multiple distance matrices in the specific context of capturing two types of heterogeneity in preference data. Their model combined two major types of latent utility models for preference data in a generalized, mixture model framework. However, in contrast to our work, Park, DeSarbo and Liechty (2008) specified normally distributed likelihood functions for their observed dominance score matrices that were specific to the two latent utility functions considered, as opposed to the general setting of skewed, positive distance data, as we consider here. Additionally, the model in Park, DeSarbo and Liechty (2008) specified the same variance across the multiple distance matrices and assumed Euclidean distances between embedding vectors, while HMDS allows for heterogeneous potential for variation for each distance matrix and relaxes the Euclidean assumption. Additional extensions to BMDS include Bayesian MDS with variable selection (Lin and Fong (2019)), which incorporated covariate information in the dimensionality reduction performed by classical MDS and allowed for heterogeneous variability by distance pair, and Bayesian MDS with simultaneous variable selection with dimension reparameterization (Fong et al. (2015)). In contrast to HMDS, both Lin and Fong (2019) and Fong et al. (2015) did not extend BMDS to multiple distance matrices and, again, assumed normal distributions for the observed distance data. In summary, our proposed HMDS model is distinguished from prior work, such as Fong et al. (2015), Lin and Fong (2019), Park, DeSarbo and Liechty (2008), by the extension to multiple distance matrices with heterogeneous variation, the modeling of nonnegative distances with gamma distributions and the relaxation of the Euclidean assumption of the systematic component of our observed distance matrices.

HMDS also extends previous approaches to the comparison of audio recordings (Liebman, Ornoy and Chor (2012), Liem and Hanjalic (2015), Peperkamp, Hildebrandt and Liem (2017), Sapp (2007), Sapp (2008)) from the perspective of the musical application of interest. HMDS seeks to find systematic similarities across recordings in a comparative and unsupervised way and, as such, does not rely on an annotated score or audio representation, which can be expensive to obtain, severely limiting the breadth of audio recordings to which such a method could be applied. Additionally, much previous comparison work has been limited to the comparison of a few pieces (Liem and Hanjalic (2015)) or a single instrument (Liebman, Ornoy and Chor (2012), Sapp (2007), Sapp (2008)). In contrast, HMDS allows for the analysis of systematic similarities of *many* pieces of *orchestral* recordings, and such large-scale analysis is critical to drawing musicological conclusions (Cook (2005)). Indeed, it would be prohibitive for a musicologist to listen to and compare pairwise 37 pieces recorded by 10 different orchestras, as we consider here. The comparison of recordings in this work is on a much larger scale, in terms of the number of pieces and the number of instruments (orchestral works), than the majority of previous work (Liebman, Ornoy and Chor

(2012), Liem and Hanjalic (2015), Peperkamp, Hildebrandt and Liem (2017), Sapp (2007), Sapp (2008)). Finally, HMDS is one of the few truly statistical, model-based approaches for orchestral recording comparison, as most existing approaches rely on purely algorithmic approaches (Liebman, Ornoy and Chor (2012), Liem and Hanjalic (2015), Peperkamp, Hildebrandt and Liem (2017)). In particular, our approach in HMDS allows us to explore several of the findings and questions raised in Liebman, Ornoy and Chor (2012), for example, if newer recordings are less idiosyncratic than older ones and the importance of geographic location in performance style for a much larger orchestral setting. While we do not specifically perform hypothesis tests in this work, the model-based nature of HMDS allows for natural extensions to hypothesis tests on such musical questions of interest which is a feature lacking in previous music information retrieval based approaches.

The audio processing procedure and musical metrics are discussed in Section 2. In Section 3 we first review BMDS, then develop our HMDS model and provide an algorithm for posterior approximation. In Section 4 we fit an HMDS model to the audio recordings of all nine Beethoven symphonies recorded by 10 different orchestras. We interpret the results and evaluate quantitatively the differences among the orchestras in the context of known, qualitative information. Our results recover several expected systematic similarities between orchestras as well as identify some more novel results. For example, we find that modern recordings exhibit a high degree of similarity to each other, as compared to older recordings. Conclusions and directions for future work are discussed in Section 5.

**2. Audio feature extraction.** Fully parametric models of raw audio data can be highly complex, especially for our goal of identifying differences between orchestras. We thus propose to reduce the audio recordings for each piece to pairwise distances between orchestras, based on different musical characteristics of the recordings. Specifically, for each recording we create three positive functions of time, representing tempo, volume dynamics and timbre over the duration of the recording. For each of these three audio characteristics and for each piece, a distance is computed between each pair of orchestras using the Hellinger metric distance between the corresponding functions. In the remainder of this section, we motivate this proposed audio processing methodology, starting with identification of musical features of interest, then the processing and aligning of the audio recordings and finally comparison of the the audio features to form distance matrices.

*2.1. Data and features of interest.* Our original data consist of audio recordings of all nine Beethoven symphonies recorded by 10 different orchestras (Table 1); the recordings span from the 1950s to 2016. Each movement is treated as a separate piece, resulting in a total of 37 pieces. For this work we consider three main musical features of interest: tempo, dynamics and timbre. Overall, we are interested in extracting features that represent artistic or expressive choices made by conductors and orchestras, rather than features that are specific to the recording process. For example, the overall volume of a recording is a function of the microphone placement during the recording process and is not an artistic choice made by the conductor. For all of the musical features considered, we attempt to isolate and remove artifacts of the recording process to focus on expressive musical features.

Tempo is the speed at which a piece is performed and often varies over the course of an orchestral piece of music. We are interested in relative tempo changes between orchestras and not in the overall speed of a recording. For example, if the score for a given piece calls for the tempo to accelerate at a specific point in the piece, one orchestra may accelerate over only one measure of music, while another orchestra may accelerate over three measures of music. Or, during the *accelerando* denoted in the musical score, one orchestra may double their tempo, while another orchestra may barely increase their speed at all. These types of relative tempo changes between orchestras are examples of tempo features of interest.

TABLE 1

*Orchestras, conductors and recordings years for the audio recordings of the nine Beethoven symphonies considered in this work*

Orchestra	Conductor	Recording Years
Academy of Ancient Music	Hogwood	1986–1989
Berlin Philharmonic	Rattle	2016
Berlin Philharmonic	von Karajan	1982–1984
Chicago Symphony Orchestra	Solti	1991
Leipzig Gewandhaus Orchestra	Masur	1989–1993
London Symphony Orchestra	Haitink	2006
NBC Symphony Orchestra	Toscanini	1939–1952
New York Philharmonic	Bernstein	1961–1967
Philadelphia Orchestra	Muti	1988–2000
Vienna Philharmonic	Rattle	2012

Dynamics refer to the relative changes in volume of an orchestra over the course of a piece, and we are interested in relative dynamic dissimilarities between orchestras rather than the overall volume. For example, suppose the score for a given piece calls for a *crescendo* (an increase in volume) at a specific point in the piece. One orchestra may play twice as loud at the end of the crescendo as they did at the beginning of the crescendo, while another orchestra may not noticeably increase their volume at all over the crescendo.

Musical timbre refers to the quality or color of an orchestra’s sound. For example, a violin and a trombone have different timbres and thus sound different from each other, even when playing the same note pitch. Individual orchestra members contribute to the overall differences in timbre between different orchestras; for this work we consider the global timbre of the entire orchestra as a feature of interest. While several methods for analyzing timbre exist (Grachten et al. (2013), Logan (2000), Sueur (2018)), we consider spectral flatness as a proxy for the timbre of the orchestra. Spectral flatness is a measure of the tonality of a sound, where a spectral flatness of 1 means that there are equal amounts of energy spread throughout the entire spectrum (white noise), while a spectral flatness close to 0 indicates that the energy in the audio signal is concentrated in only a few frequency bands, approaching a pure tone (Sueur et al. (2018)). It is important to note that timbre is an inherently perceptual property of sound and can be difficult to define precisely; thus, spectral flatness is really a measure of one aspect that influences timbre, rather than a direct measure of timbre itself (Müller (2015)).

Based on our measure of spectral flatness, a difference in timbre between two orchestras corresponds to a difference in the spread of energy across the spectrum. For example, at a specific point in a piece the flutes in one orchestra may play as loud as the lower pitched instruments, resulting in a spread of energy across frequency bands and a high spectral flatness. On the other hand, an orchestra where only lower instruments play at a loud volume for the same part in the piece would have spectral energy concentrated in fewer frequency bands and thus a lower spectral flatness. Spectral flatness can be sensitive to the recording technology used to produce the audio signal. For example, older recordings that were converted from analog signals may have less energy in the upper frequencies, due to the audio conversion, and thus lower spectral flatness values compared to modern recordings that do not clip the upper frequencies. However, we believe that spectral flatness is a good initial proxy for the relative timbre features of an orchestra.

*2.2. Audio processing and alignment.* In this subsection we describe the specific audio processing steps to transform the original audio recordings into data representations that can be used to calculate pairwise distances for each of the musical features or metrics of interest



described above. Our procedure consists of three steps—spectral preprocessing, alignment and calculation of features; each of which will be described in detail below.

*2.2.1. Spectral preprocessing.* Before calculating the musical metrics of interest, we need a musically meaningful representation of the audio signals that will allow for the calculation of these features. For example, we cannot determine the tempo of an orchestra for a given piece from the raw audio signal, so we will need another data representation that facilitates musical feature extraction. We use two different representations of the audio signal, the spectrogram and the chromagram. The spectrogram represents the power spectrum of the audio signal over the entire frequency range and is found via the short-time Fourier transform (STFT). The spectrogram contains the energy distribution of the audio signal over time and is used to calculate the spectral flatness metric. The spectrogram for orchestra  $i$  for piece  $p$  is a  $F \times T$  matrix,  $S_{ip}(f, t)$ , which represents the magnitude of the  $f$ th frequency band at time  $t$ . For all orchestras and all pieces, the frequency resolution of the spectrogram representation is 5 Hz and the temporal resolution is 0.1 seconds. We only use the magnitude information from the spectrogram and ignore the phase information.

The chromagram representation (Müller (2015)) can be calculated from the spectrogram and is used to calculate the tempo and dynamics metrics. The chromagram aggregates the amplitude of each frequency bin in the spectrogram across octaves to give one amplitude for each pitch in the 12 tone scale. This aggregation is robust to differences in instrument balance and intonation between orchestras. The chromagram is a  $12 \times T$  matrix, where each row corresponds to one note pitch or pitch class. For orchestra  $i$  and piece  $p$ , let  $\{f_q\}$  be the set of frequency bands that correspond to pitch class  $q$ . For example,  $q = 11$  corresponds to a  $B^b$  note pitch in any octave. Then, the chromagram,  $\psi_{ip}$ , can be calculated from the spectrogram as follows:

$$(1) \quad \psi_{ip}(q, t) = \sum_{k \in \{f_q\}} S_{ip}(k, t).$$

The 12-dimensional chromagram representation at time  $t$ ,  $\psi_{ip}(:, t)$ , can be thought of as approximating the notes that orchestra  $i$  plays at time  $t$ . Example chromagrams are shown in Figure 1.

In summary, we calculate the spectrogram  $S_{ip}$  and chromagram  $\psi_{ip}$  representations for all orchestras  $i$  and pieces  $p$ . All of the musical features of interest are derived from these two representations; the tempo and dynamics features are calculated from the chromagram representation, and the timbre features are calculated from the spectrogram representation.

*2.2.2. Alignment.* For all pieces  $p$ , orchestra  $i$  and orchestra  $j$  play the same notes in the same order. However, different orchestras may play these sequences of notes at different speeds. For example, in Figure 1(a), the Vienna Philharmonic and the New York (NY) Philharmonic play the opening of Beethoven's Fifth Symphony at different tempos; Vienna only holds the opening  $G$  pitch for about half a second and the following  $E^b$  pitch from 0.5 to 2.5 seconds, while NY plays the opening  $G$  for nearly an entire second and sustains the following  $E^b$  pitch for nearly three seconds. Our goal is to compare differences in tempo, dynamics and timbre when each orchestra is performing the *same part of each piece* to assess differences in artistic or expressive aspects of the performance. Before calculating the musical features of interest, we then need to temporally align the spectrogram and chromagram representations so that, on piece  $p$  at time  $t$  for orchestra  $i$  and at time  $t'$  for orchestra  $j$ , these two orchestras are performing the same part of the piece.

Since the chroma vectors approximate the notes that an orchestra plays at time  $t$ , we want to find the warping path of indices,  $w(t)$ , such that  $\psi_{ip}(:, t) = \psi_{jp}(:, w(t))$  for all  $t$ , subject to

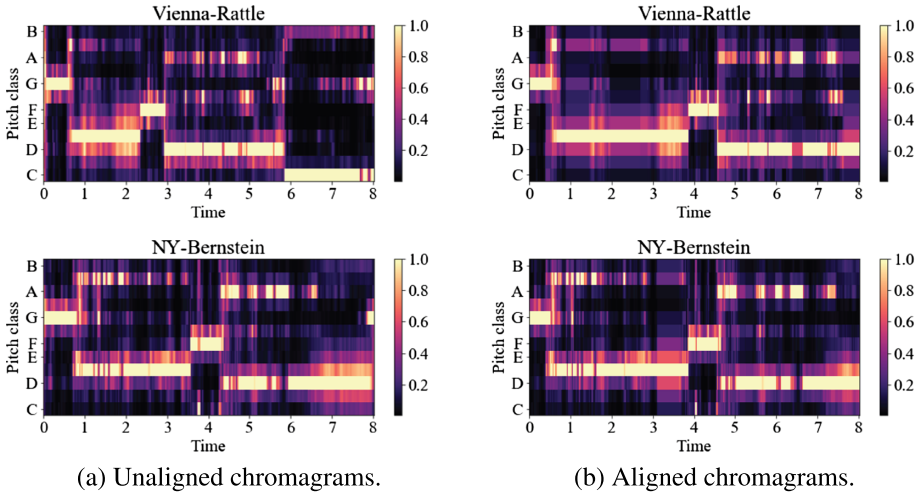


FIG. 1. (a) Unaligned and (b) aligned chromagrams for the opening of Beethoven No. 5 - *Mvt. 1* by the Vienna Philharmonic and the NY Philharmonic. The unaligned chromagrams show that the two orchestras play the same notes pitches but at different speeds and for different durations, so these two orchestras do not perform the same part of the piece at the same point in time. However, after alignment, the two orchestras do perform the same part of the piece at the same point in time, and musical features can be extracted.

the constraints that  $w(t)$  is monotonically increasing and that the orchestras start and end at the same point in the piece, that is,  $\psi_{ip}(:, 0) = \psi_{jp}(:, 0)$  and  $\psi_{ip}(:, T_{ip}) = \psi_{jp}(:, T_{jp})$ , where  $T_{ip}$  and  $T_{jp}$  are the lengths of piece  $p$  for orchestras  $i$  and  $j$ , respectively. This problem can be solved via dynamic time warping. Dynamic time warping finds a nonlinear warping path,  $w(t)$ , between the two chromagrams and is frequently used in music information retrieval (Ellis (2007), Kammerl et al. (2014), Kirchoff and Lerch (2011), Müller (2015), Peperkamp, Hildebrandt and Liem (2017), Thornburg, Leistikow and Berger (2007)). Dynamic time warping has also been used frequently within the larger context of MDS (Kruskal (1983)).

Following Peperkamp, Hildebrandt and Liem (2017), we align each piece to a reference MIDI recording from [Kunsterfuge.com \(2018\)](#) (MIDI is a symbolic music representation and is simplified compared to the audio orchestral recordings). The result of the dynamic time warping is a warping path,  $w_{ip}(t)$ , for each orchestra  $i$  for each piece  $p$ , relative to the reference MIDI recording. Then, we have that  $\psi_{ip}(:, w_{ip}(t)) \approx \psi_{jp}(:, w_{jp}(t))$  and  $S_{ip}(:, w_{ip}(t)) \approx S_{jp}(:, w_{jp}(t))$  for all  $i, p$  and  $t$ . This equivalence is approximate, as specific chroma amplitudes can and do differ by orchestra; these differences correspond to variation in dynamics and timbre by orchestra, for example. After the alignment, however, all orchestras perform the same part of the piece at the same time. For simplicity, we also normalize the time by the length of each piece for each orchestra so that  $t \in [0, 1]$  for all orchestras  $i$  and pieces  $p$ . We can now use the aligned chromagrams and spectrograms to calculate our musical features of interest.

**2.2.3. Calculation of audio features.** After the alignment of our data representations, we are ready to calculate the specific musical metrics of interest. Starting with the aligned spectrograms and chromagrams, we calculate the tempo, dynamics and timbre features as normalized functions that are used to calculate pairwise distances between orchestras. Let  $\tilde{\psi}_{ip}$  denote the aligned chromagrams, that is,  $\tilde{\psi}_{ip}(:, t') = \psi_{ip}(:, w(t))$ , and similarly for the aligned spectrograms,  $\tilde{S}_{ip}$ , that will be used for the calculation of these musical metric functions. Example functions are shown in Figure 2.

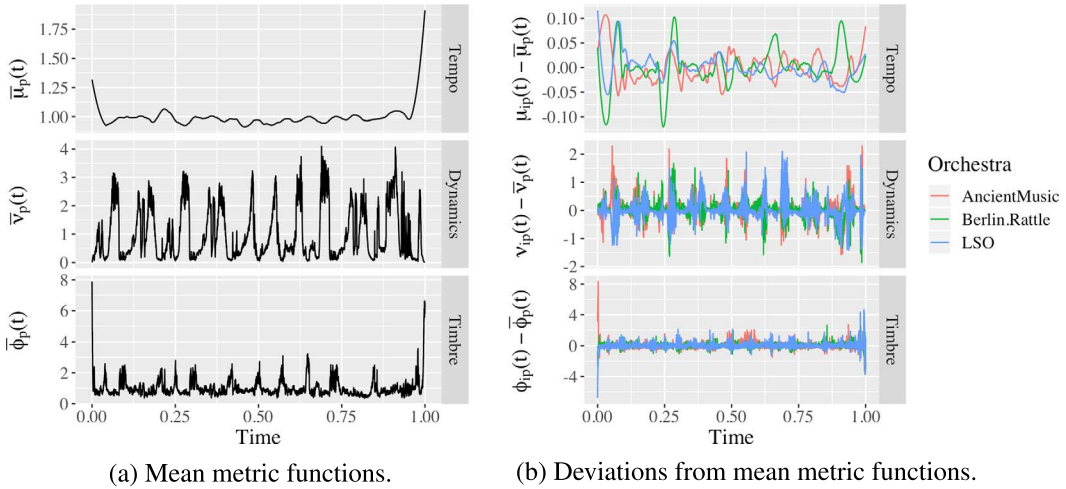


FIG. 2. (a) Mean metric functions and (b) deviations from mean metric functions by orchestra, for Beethoven Symphony No. 6, Movement 1. The time is normalized to be between 0 and 1, since all recordings are already aligned to the same reference recording.

The tempo function can be calculated using the warping path from the dynamic time warping alignment. That is, the tempo function for orchestra  $i$  on piece  $p$  is

$$\mu_{ip}(t) \propto w'_{ip}(t) = \frac{dw_{ip}(t)}{dt},$$

where  $\mu_{ip}(t)$  is the ratio of the tempo of orchestra  $i$  on piece  $p$  at time  $t$  relative to the reference recording for piece  $p$ . This means that at time  $t$  for piece  $p$ , if  $\mu_{ip}(t) = 4$  and  $\mu_{jp}(t) = 2$ , then orchestra  $i$  is playing two times as fast as orchestra  $j$ . Likewise,  $\mu_{ip}(t) < 1$  means that orchestra  $i$  is performing piece  $p$  slower than the reference recording at time  $t$ . The tempo curve is normalized such that

$$\mu_{ip}(t) = \frac{w'_{ip}(t)}{\sum_{s=0}^1 w'_{ip}(s)}.$$

The dynamics function can be calculated using the sum of the magnitudes of the aligned chromagram at each point in time, divided by the average volume for the entire piece. That is, the dynamics are calculated as

$$v_{ip}(t) \propto \frac{\sum_{q=1}^{12} \tilde{\psi}_{ip}(q, t)}{\frac{1}{12} \sum_{q=1}^{12} \sum_{s=0}^1 \tilde{\psi}_{ip}(q, s)}.$$

We normalize the dynamics curves by the average volume for that orchestra for piece  $p$ , as we are not interested in the overall volume of each orchestra. Again, we define  $v_{ip}(t)$  to be a normalized function so that  $\sum_{s=0}^1 v_{ip}(s) = 1$ . Then,  $v_{ip}(t) > v_{jp}(t)$  means that, relative to each orchestra's respective average dynamic for piece  $p$ , orchestra  $i$  is playing louder than orchestra  $j$  at time  $t$ .

Spectral flatness is the ratio of the geometric mean to the arithmetic mean and can be calculated as

$$\phi_{ip}(t) \propto F \times \frac{\sqrt[F]{\prod_{f=1}^F \tilde{S}_{ip}(f, t)}}{\sum_{f=1}^F \tilde{S}_{ip}(f, t)},$$

where  $\tilde{S}_{ip}(f, t)$  is the relative amplitude of the  $f^{\text{th}}$  frequency of the aligned spectrogram,  $F$  is the total number of frequencies for the aligned spectrogram (Sueur et al. (2018)) and we



have that  $\sum_{s=0}^1 \phi_{ip}(s) = 1$ . For the spectral flatness features,  $\phi_{ip}(t) > \phi_{jp}(t)$  means that the energy of orchestra  $j$  is concentrated in a smaller number of frequency bands than orchestra  $i$  for piece  $p$  at time  $t$ .

In summary, after transforming the audio signals to spectrogram and chromagram representations, temporally aligning the representations and calculating the musical metrics for tempo, dynamics and timbre, we have musical feature functions for tempo,  $\mu_{ip}(t)$ , dynamics,  $v_{ip}(t)$  and timbre,  $\phi_{ip}(t)$ , for all orchestras  $i = 1, \dots, 10$ , pieces  $p = 1, \dots, 37$  and time  $t \in [0, 1]$ . These functions are normalized to sum to 1 such that  $\sum_{s=0}^1 \mu_{ip}(s) = \sum_{s=0}^1 v_{ip}(s) = \sum_{s=0}^1 \phi_{ip}(s) = 1$ .

*2.3. Comparison of audio features.* Now, that we have normalized functions for each orchestra, for each piece and for each musical feature, we can calculate pairwise distance matrices using a density-based distance measure. We calculate the pairwise Hellinger distance between all orchestras for each piece to obtain  $M$  pairwise distance matrices for each musical metric. The Hellinger distance is a commonly used density-based distance measure, and for discrete distributions  $P = (p_1, \dots, p_K)$  and  $Q = (q_1, \dots, q_K)$  can be calculated as

$$(2) \quad H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^K (\sqrt{p_k} - \sqrt{q_k})^2}.$$

After calculating the pairwise Hellinger distance between orchestras, the distances within each metric are normalized to be between 0 and 1 over all pieces. The end result of the audio feature extraction procedure is  $M$  distance matrices for each of the three musical metrics, tempo,  $D^\mu$ , dynamics,  $D^v$  and timbre,  $D^\phi$ , where each distance matrix is of dimension  $N \times N$ , where  $N = 10$  is the number of orchestras and  $M = 37$  is the number of pieces. Then, for a specific musical metric,  $y_{ijp}$  is the Hellinger distance between orchestra  $i$  and orchestra  $j$  on piece  $p$  for that musical metric. We treat these  $y_{ijp}$  distances as our observed data for the modeling described in the next section. Additionally, each individual distance matrix is symmetric, so  $y_{ijp} = y_{jip}$  for all orchestras  $i$  and  $j$  and pieces  $p$ .

The overall audio processing methodology can thus be summarized as follows. We start with audio recordings of all nine Beethoven symphonies by  $N$  different orchestras. Each movement of a symphony is treated as a separate piece, resulting in  $M$  pieces. Using spectrogram and chromagram representations of the audio signals, we temporally align the orchestra representations by piece, using dynamic time warping, and calculate tempo, dynamics and timbre functions by piece. Then, we use the Hellinger density-based distance measure to calculate  $M$  pairwise replicate distance matrices (each of dimension  $N \times N$ ) for each of the three musical metrics. The data and accompanying code is available in the Supplementary Material (Yanchenko and Hoff (2020)) and online at <https://github.com/aky4wn/HMDS>.

**3. Hierarchical multidimensional scaling.** In this section we develop hierarchical multidimensional scaling, a statistical model for a sample of pairwise distance matrices among a common set of objects. The purpose of the model is to identify patterns in the distance matrices that are consistent throughout the sample as well as to quantify the variation of the distance matrices around these patterns. For example, in our application each matrix in our sample represents the pairwise distance between orchestras for a given piece for a specific musical metric, either tempo, dynamics or timbre. Our model-based approach is a modification and extension of the approach of Oh and Raftery (2001) who developed a probability model for a single dissimilarity matrix.

3.1. *Bayesian multidimensional scaling.* Oh and Raftery (2001) propose Bayesian Multidimensional Scaling (BMDS), a model-based version of classical MDS. For a *single* data matrix of pairwise distances between objects, the goals of BMDS are to find a low-dimensional representation of the objects of interest in Euclidean space and to measure the discrepancy between the Euclidean space and the observed distances. BMDS assumes that observed pairwise distance measurements are equal to a true distance measure plus observational noise, where the true distance measure is the Euclidean distance between latent embedding vectors for the pair. Let  $N$  be the number of objects or entities of interest and  $y_{ij}$  be the observed distance between object  $i$  and object  $j$ . Then, the BMDS model is defined as follows:

$$(3) \quad y_{ij} \sim \text{TruncNorm}(\|X_i - X_j\|_2, \sigma^2), \quad j > i, i, j = 1, \dots, N,$$

independently, where  $X_1, \dots, X_N$  are unobserved latent vectors in  $r$ -dimensional Euclidean space, one for each object  $i$ , and  $\sigma^2$  is an unknown scale parameter. TruncNorm is the normal density, truncated to be above 0. Note that for BMDS, the error term  $\sigma^2$  represents the deviation of the observed distances from Euclidean distances which could be attributed to either measurement error or misspecification of the (true) distances being Euclidean.

Oh and Raftery (2001) describe a Markov chain Monte Carlo algorithm for approximating the posterior distribution of  $X_1, \dots, X_N$  and  $\sigma^2$ , conditional on the observed distance data  $\{y_{ij} : j > i\}$ . They specify independent priors for each of the unknown parameters. The latent  $X_i$  vector for each object is assumed to come from an independent,  $r$ -dimensional normal distribution with diagonal covariance matrix  $\Lambda$ , where the diagonal elements of  $\Lambda$  are inverse-gamma distributed. The error term,  $\sigma^2$ , is assigned an inverse-gamma prior distribution for conjugacy. The original BMDS model does not consider replications or multiple distance matrices, though later extensions do for a specific preference data application (Park, DeSarbo and Liechty (2008)), as discussed in Section 1.

3.2. *Hierarchical MDS for multiple distance matrices.* Given only a single distance matrix of observed distances, the BMDS model cannot distinguish between measurement error and the degree to which the systematic distances between objects differ from being Euclidean. For example, a large estimated  $\sigma^2$  value in equation (3) could indicate a large amount of measurement error in the observed pairwise distances *or* that the observed distances are not well represented by Euclidean distances between latent vectors (or some combination of these two factors). With *replicate distance matrices*, however, there is sufficient information to distinguish a non-Euclidean mean distance from across-sample measurement variation. We quantify the variation of the replicate distance matrices around a common mean distance matrix with the following hierarchical MDS (HMDS) model. Let  $y_{ijp}$  be the observed pairwise distance between entity  $i$  and entity  $j$  for observation  $p$ , where there are  $N$  total entities and  $M$  total replicate distance matrices. The HMDS model is given in equation (4),

$$(4) \quad y_{ijp} \sim \text{Gamma}\left(\psi, \frac{\psi}{\tau_p \delta_{ij}}\right), \quad j > i, i, j = 1, \dots, N, p = 1, \dots, M,$$

independently across pairs and replicates, where  $\psi, \tau_1, \dots, \tau_M$  and  $\{\delta_{ij} : j > i\}$  are parameters to be estimated. The gamma distribution is parameterized such that the mean of  $y_{ijp}$  is  $\tau_p \delta_{ij}$  and the variance of  $y_{ijp}$  is  $(\tau_p^2 \delta_{ij}^2) / \psi$ . Each  $\tau_p$  is a scale parameter for replicate distance matrix  $p$  that allows for each matrix to have a different “potential” for variation. The  $\{\delta_{ij} : j > i\}$  parameters represent the systematic dissimilarity between entities  $i$  and  $j$  across all  $M$  replicate matrices, while the  $\psi$  parameter serves as an overall scale factor. Note that  $y_{ijp} = y_{jip}$  for all entities  $i$  and  $j$  and replicate distance matrices  $p$ , so we only model explicitly the unique entries in each distance matrix.

Inclusion of the  $\tau_1, \dots, \tau_M$  parameters is important in our application, as we expect some pieces to have more inherent opportunities for variation than other pieces. For example, some pieces have numerous vague tempo markings that allow for a good deal of artistic interpretation and tempo variation between orchestras, as compared to other pieces that do not have many denoted changes in tempo. However, a piece’s “potential” for variation is a characteristic of the piece and is separate from the systematic differences between orchestras and must be handled accordingly. This potential for variation scales both the mean and the variance of the observed distances.

The proposed HMDS model differs in several ways from the BMDS model. First, the gamma distribution for the observed pairwise distances is a more natural choice for a positive random quantity than the truncated normal distribution of BMDS. The gamma distribution allows for skew in the observed pairwise dissimilarities and facilitates straightforward parameter estimation and inference, as will be described below. Second, the HMDS model in equation (4) does not restrict the  $\{\delta_{ij} : j > i\}$  parameters to correspond to Euclidean distances. This relaxes the assumptions of BMDS and is important in many applications, including our comparison of orchestral recordings, where the distance metrics used to compare objects are known to be non-Euclidean.

The HMDS parameters can be estimated with maximum likelihood estimation. The MLE estimates for the  $\delta_{ij}$  and  $\tau_p$  parameters satisfy the following system of equations:

$$\hat{\delta}_{ij}^{\text{MLE}} = \frac{1}{M} \sum_{p=1}^M \frac{y_{ijp}}{\hat{\tau}_p^{\text{MLE}}}, \quad \hat{\tau}_p^{\text{MLE}} = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j>i} \frac{y_{ijp}}{\hat{\delta}_{ij}^{\text{MLE}}}.$$

The MLE for  $\psi$  can be found by iteratively solving the following equation:

$$\frac{\Gamma'(\psi)}{\Gamma(\psi)} - \log \psi = \frac{2}{N(N-1)M} \left[ 1 + \sum_{i=1}^N \sum_{j>i} \sum_{p=1}^M \left( \log \left( \frac{y_{ijp}}{\tau_p \delta_{ij}} \right) - \frac{y_{ijp}}{\tau_p \delta_{ij}} \right) \right],$$

where  $\Gamma'(\psi)$  is the digamma function,  $\Gamma'(\psi) = \frac{d\Gamma(\psi)}{d\psi}$ . Importantly, note that the MLE estimates,  $\hat{\delta}_{ij}^{\text{MLE}}$ , might not be distances.

We choose to perform Bayesian inference for parameter estimation. Bayesian inference in the HMDS model naturally allows for parameter uncertainty estimates. Additionally, the space of  $N \times N$  distance matrices is quite high dimensional and complex, and Bayesian inference provides shrinkage toward a lower dimensional space. Finally, the choice of a hierarchical model for the  $\{\delta_{ij} : j > i\}$  parameters centered around a Euclidean space can aid in parameter interpretation. That is, we can specify a prior that puts the  $\{\delta_{ij} : j > i\}$  parameters near some actual distances that satisfy the triangle inequality. To that end, we model the  $\{\delta_{ij} : j > i\}$  parameters as

$$(5) \quad \delta_{ij} \sim \text{Inv-Gamma}(\gamma, (\gamma + 1)\|X_i - X_j\|_2), \quad j > i, i, j = 1, \dots, N,$$

independently across pairs. The inverse-gamma distribution is parameterized such that the prior mode for  $\delta_{ij}$  is  $\|X_i - X_j\|_2$ . The goal of the prior is to shrink the dissimilarity between orchestras  $i$  and  $j$  toward a distance metric that follows the triangle inequality. The  $\|X_i - X_j\|_2$  term fixes this metric space as an  $N - 1$  dimensional Euclidean space, though the  $\gamma$  parameter allows for potentially substantial variation from this Euclidean distance. Notably, this variation about a Euclidean distance, represented by the  $\gamma$  parameter in HMDS, is separate from the across-replicate sampling variability, represented by the parameter  $\psi$ . This separation of sources of variation is in contrast to the BMDS model.

This particular choice of an inverse-gamma prior further facilitates computation and interpretation. The effect of the prior on estimation of the  $\delta_{ij}$ ’s can be understood from the form

of their full conditional distributions. The conditional density of  $\delta_{ij}$ , given all other model parameters and the observed  $y_{ijp}$  pairwise distances, is

$$\delta_{ij}|\psi, \gamma, \{\tau_p\}_1^M, \{X_i\}_1^N, Y \sim \text{Inv-Gamma}\left(M\psi + \gamma, (\gamma + 1)\|X_i - X_j\|_2 + \psi \sum_{p=1}^M \frac{y_{ijp}}{\tau_p}\right),$$

where  $Y = \{y_{ijp} : i = 1, \dots, N, j > i, p = 1, \dots, M\}$ . The mode of this conditional distribution is

$$\frac{\gamma + 1}{M\psi + \gamma + 1}\|X_i - X_j\|_2 + \frac{M\psi}{M\psi + \gamma + 1}\hat{\delta}_{ij}^{\text{MLE}},$$

so the  $\delta_{ij}$  prior specified in equation (5) shrinks the MLE estimate toward a set of Euclidean distances.

The full conditional distribution of the  $\tau_p$  parameters, given all other model parameters and the observed  $y_{ijp}$  pairwise distances, is

$$\tau_p|\psi, \gamma, \{\delta_{ij} : j > i\}, \{X_i\}_1^N, Y \sim \text{Inv-Gamma}\left(\frac{N(N - 1)}{2}\psi + 1, \psi \sum_{i=1}^N \sum_{j>i} \frac{y_{ijp}}{\delta_{ij}}\right),$$

and the conditional mode is

$$\frac{\psi N(N - 1)}{\psi N(N - 1) + 4}\hat{\tau}_p^{\text{MLE}}.$$

The conditional mode of  $\tau_p$  depends on how close the observed pairwise distances (the  $y_{ijp}$ 's) are to the systematic dissimilarities (the  $\delta_{ij}$ 's) across all pairs,  $j > i$ . That is, if  $y_{ijp} = \delta_{ij} \forall j > i, i = 1, \dots, N$ , then  $\hat{\tau}_p^{\text{MLE}} = 1$  and the variation of the  $y_{ijp}$  distances is only scaled by  $\psi$ . In this case the observed distances are equal to the systematic dissimilarities, and there is no “potential” for variation for replicate matrix  $p$ . However, when  $y_{ijp} > \delta_{ij}$  across object pairs for a given replication matrix  $p$ , then  $\hat{\tau}_p^{\text{MLE}} > 1$  and that specific replication  $p$  has a high “potential” for variation.

For a full Bayesian analysis we also need to specify priors for the remaining unknown parameters,

$$\begin{aligned} (6) \quad & X_1, \dots, X_N \stackrel{\text{indep.}}{\sim} \mathcal{N}_{N-1}(0, \Lambda), \\ & \psi \sim \text{Gamma}(a_1, b_1), \\ & \gamma \sim \text{Gamma}(a_2, b_2), \\ & \tau_1, \dots, \tau_M \stackrel{\text{indep.}}{\sim} \text{Inverse-Gamma}(\alpha, \beta), \end{aligned}$$

where  $\Lambda$  is a diagonal matrix and  $a_1, b_1, a_2, b_2, \alpha$  and  $\beta$  are positive scalars. We set  $a_1, b_1, a_2, b_2$  to 0.01 and  $\alpha$  and  $\beta$  to 1 in our application. These choices of hyper priors specify weakly-informative prior distributions for  $\psi, \gamma$  and the  $\tau_p$ 's, following Gelman (2006). In particular, the choice of the shape parameter  $\alpha$  less than or equal to 1 in the prior for the  $\tau_p$  parameters means that we believe the main mass of the prior for each  $\tau_p$  is pushed toward 0, favoring small values of  $\tau_p$ , while still maintaining heavy tails, allowing for potentially very large  $\tau_p$  values. As we expect there to be different potentials for variation by piece  $p$ , this inverse-gamma prior specification allows for a wide spread of possible  $\tau_p$  values. We use an empirical Bayes approach to setting  $\Lambda$  by estimating the variance of embedding vectors from classical MDS performed for each piece. A full sensitivity analysis of the hyper-parameter choices for the gamma and inverse-gamma priors is included in Section 2 of the Supplementary Material (Yanchenko and Hoff (2020)); we find that the overall results are not very sensitive to the

choice of prior hyper-parameter values. Similarly to BMDS, the  $X_i$  vectors are assumed to explicitly follow a prior distribution here which is in contrast to the usual point estimates for the embedding vectors that result from classical MDS.

The joint posterior distribution of model parameters can be approximated with a Markov chain Monte Carlo algorithm. One such algorithm proceeds as follows:

1. For each  $i = 1, \dots, N, j > i$ , simulate

$$\delta_{ij} \sim \text{Inv-Gamma}\left(M\psi + \gamma, (\gamma + 1)\|X_i - X_j\|_2 + \psi \sum_{p=1}^M \frac{y_{jip}}{\tau_p}\right).$$

2. For each  $p = 1, \dots, M$ , simulate

$$\tau_p \sim \text{Inv-Gamma}\left(\alpha + \frac{N(N-1)}{2}\psi, \beta + \psi \sum_{i=1}^N \sum_{j>i} \frac{y_{ijp}}{\delta_{ij}}\right).$$

3. The remaining parameters,  $\{X_1, \dots, X_N, \psi, \gamma\}$ , can be updated via Metropolis-Hastings steps.

Full details of the Markov chain Monte Carlo algorithm are given in Section 1 of the Supplementary Material (Yanchenko and Hoff (2020)).

**3.3. Comparison to classical MDS approaches.** In the classical MDS literature, related approaches for three-way data include INDSCAL (Individual Differences in Scaling) and IDIOSCAL (Individual Differences in Orientation Scaling) (Carroll and Chang (1970)). Both of these approaches extend classical MDS to repeated observations or measurements between entities. However, unlike HMDS, a primary purpose of these methods is to identify a common low-dimensional subspace in which these distance matrices can be embedded and to describe across-matrix heterogeneity in this subspace. INDSCAL and IDIOSCAL both assume that there exists a group embedding space,  $Z$ , which represents commonalities among all entities. Each entity then has an individual space,  $X_i$ , that is a transformation of the group space  $Z$ , that is  $X_i = ZW_i$ , where  $W_i$  is a diagonal matrix with positive entries for INDSCAL and a general matrix for IDIOSCAL (Borg, Groenen and Mair (2018)). Both methods iteratively update  $Z$  and  $W_i$  to minimize the least-squares difference between the observed and fit distances. The weight matrix  $W_i$  can be interpreted as the transformations that stretch and compress the group space  $Z$  into each individual space  $X_i$  (Borg, Groenen and Mair (2018)). Similar to HMDS, this  $W_i$  matrix in INDSCAL and IDIOSCAL allows for variation by replication. INDSCAL and IDIOSCAL also share information across replicates by assuming a common group subspace,  $Z$ .

While these methods could be applied to our audio recording data, the objectives and assumptions of INDSCAL and IDIOSCAL are different from those of HMDS. In particular, our primary goal with HMDS is to estimate a distance matrix (the  $\delta_{ij}$ 's) that represents consistent differences between pairs of orchestras across pieces, rather than to find a common, reduced-dimension subspace in which to represent the across-matrix variations. Indeed, the  $\delta_{ij}$ 's of interest in HMDS need not be embeddable in a lower-dimensional Euclidean space. Although each  $X_i$  shares a common prior specification, the  $X_i$  vectors are not constrained to lie in a common "group space" like in INDSCAL and IDIOSCAL. The HMDS method is additionally applicable to distance matrices that are not Euclidean, which is especially important in our application, where the distances of interest are known to be non-Euclidean. Finally, a primary, important difference between HMDS and classical MDS approaches is that our approach is model-based, which allows for a description of uncertainty in the parameter estimates, as well as information sharing across the data matrices via the hierarchical model. We thus view HMDS as an extension of BMDS to replicate distance matrices that focuses on different modeling goals than the classical INDSCAL and IDIOSCAL approaches.



**4. Analysis of orchestral distance data.** In this section we fit the HMDS model to the replicate orchestral distance matrices to explore systematic differences between orchestras across pieces. We fit the HMDS model separately for each of the three musical distance metrics: tempo, dynamics and timbre. We check the MCMC approximation outlined in Section 3.2 and examine the goodness-of-fit of HMDS for our orchestral audio data. Finally, we analyze the learned parameters for each musical metric and find that the HMDS model is able to recover musically expected systematic differences between orchestras across pieces as well as suggest some unexpected similarities between orchestras.

**4.1. MCMC approximation and goodness-of-fit.** The posterior distribution of parameters in the HMDS model can be approximated using a Markov chain Monte Carlo (MCMC) algorithm, such as the one described in Section 3.2, or a more general MCMC algorithm (Stan Development Team (2019a)). We fit the HMDS model in Rstan with the default No-U-Turn Hamiltonian Monte Carlo sampler. Parameter values were randomly initialized, and the chain was run for 30,000 iterations, where the first 15,000 iterations were discarded as burn-in and no thinning was performed. The final 15,000 iterations were retained for posterior inference. We diagnose the posterior approximation with trace plots and effective sample size (ESS) diagnostics (Stan Development Team (2019b)). The ESS diagnostic uses the estimated autocorrelation of the Markov chain to assess the Monte Carlo approximation error. Specifically, an ESS of  $n_{\text{eff}}$  means that the Monte Carlo error is approximately equivalent to what it would be from  $n_{\text{eff}}$  independent simulations from the posterior distribution (Gelman et al. (2014)).

Trace plots for a subset of the  $\tau_p$  parameters for the tempo metric are shown in Figure 3 and trace plots for the remaining metrics are included in Section 2 of the Supplementary Material (Yanchenko and Hoff (2020)). The trace plots are used to identify any evidence of nonstationarity of the Markov chain which would indicate that the Markov chain would need to be run longer. All trace plots appear stationary for the 15,000 posterior simulations used for analysis. The median ESS values across all  $M$  pieces for the  $\tau_p$  parameters are 407 for the tempo metric, 189 for the dynamics metric and 274 for the timbre metric. While the ESS values for the  $\tau_p$  parameters do indicate that the posterior Markov chains are mixing slowly, the trace plots across parameters and musical metrics appear stationary.

We also analyze the goodness-of-fit of the HMDS model for our orchestral audio data using posterior predictive checks (Gelman et al. (2014)). We can evaluate the goodness of fit of the pair-specific sampling model by simulating posterior predictive values,  $\tilde{y}_{ijp} \sim \text{Gamma}(\psi, \psi/(\tau_p \delta_{ij}))$ , at each iteration of the Markov chain. These values can be compared to the observed  $y_{ijp}$ 's to evaluate the fit of the pair-specific sampling model by computing  $r_{ijp} = \log(\tilde{y}_{ijp}/y_{ijp})$  for each simulated  $\tilde{y}_{ijp}$ , where the ratio accounts for differences in

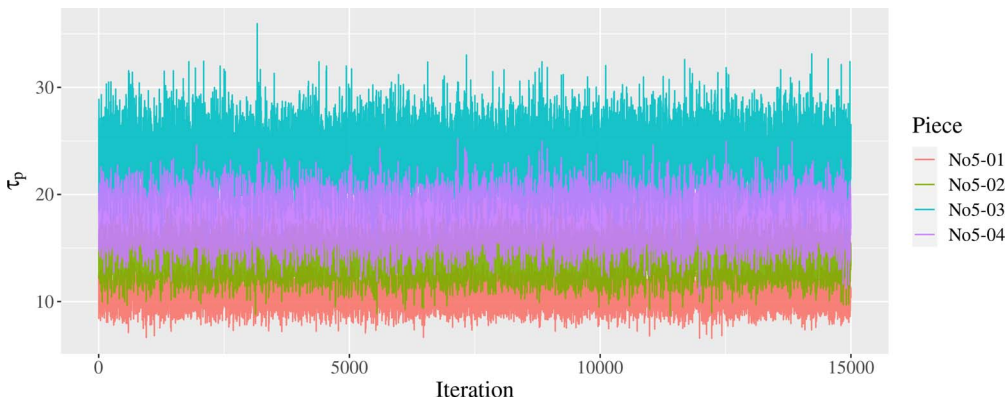


FIG. 3. Trace plot for Symphony No. 5 for  $\tau_p$  for the tempo metric.

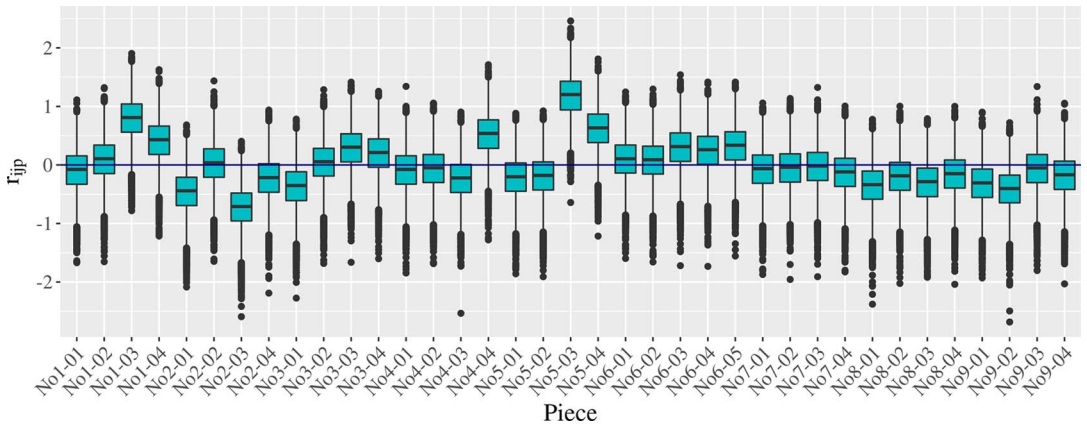


FIG. 4. Posterior predictive checks for the HMDS sampling model for one orchestra pair (Academy of Ancient Music and Vienna-Rattle) across all pieces for the tempo metric,  $r_{ijp} = \log(\tilde{y}_{ijp}/y_{ijp})$ .

scale due to  $\tau_p$ . The distribution of these ratios for one orchestra pair and all pieces for the tempo metric is displayed in Figure 4. Additionally, we can evaluate the coverage of these ratio distributions. For the tempo metric, 95.14% of the 95% highest-posterior density (HPD) intervals for each  $r_{ijp}$  contain 0. Likewise, 95.79% and 95.38% of the 95% HPD intervals for  $r_{ijp}$  contain 0 for dynamics and timbre, respectively. Overall, these results indicate that the sampling model for HMDS fits the data well.

Likewise, we can evaluate the goodness-of-fit of the hierarchical portion of HMDS with posterior predictive checks. At each iteration of the Markov chain, we can calculate  $\|X_i - X_j\|_2$ , simulate  $\tilde{\delta}_{ij} \sim \text{Inv-Gamma}(\gamma, (\gamma + 1)\|X_i - X_j\|_2)$  and then simulate posterior predictive values,  $\tilde{y}_{ijp} \sim \text{Gamma}(\psi, \psi/(\tau_p\tilde{\delta}_{ij}))$ . We compute  $r_{ijp} = \log(\tilde{y}_{ijp}/y_{ijp})$  for each simulated  $\tilde{y}_{ijp}$  and the distribution of these ratios for all orchestra pairs, averaged across pieces (i.e.,  $\bar{r}_{ij}$ ) is displayed in Figure 5. Again, we can access the coverage of the distribution

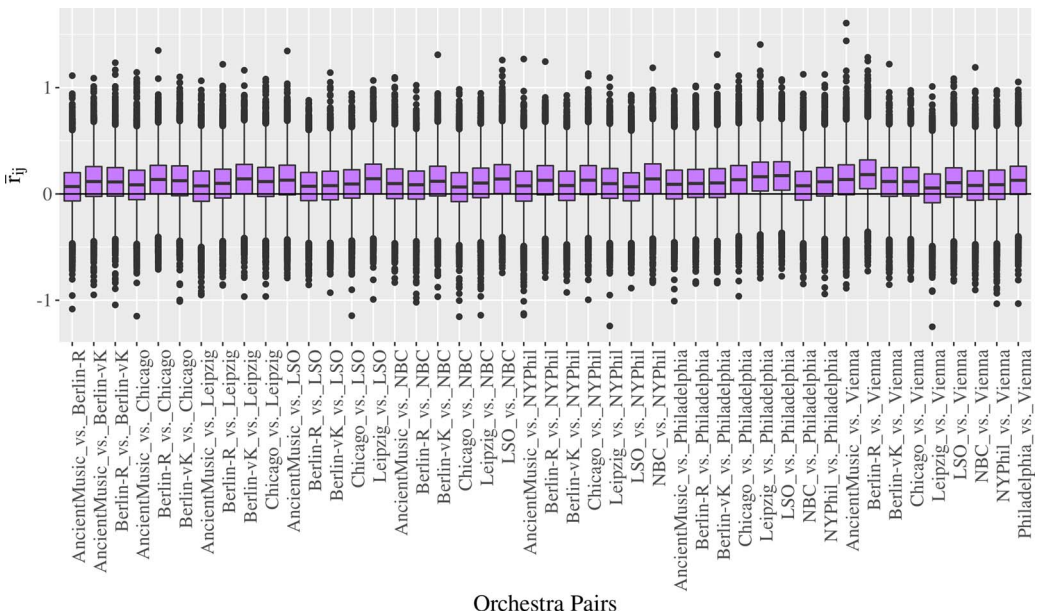


FIG. 5. Posterior predictive checks for the hierarchical sampling model for the tempo metric, averaged across all pieces.

of  $\bar{r}_{ij}$ ; 97.96%, 100% and 99.64% of the 95% HPD intervals for  $\bar{r}_{ij}$  for the tempo, dynamics and timbre metrics, respectively, contain 0.

**4.2. Potential for variation by piece.** The  $\tau_p$  parameters were introduced into the HMDS model to account for heterogeneous variation in the potential for across-orchestra differences by piece, where this potential for variation is determined by musical characteristics of the specific piece. The posterior distributions for each piece for  $\tau_p$  are given in Figure 6 for the tempo metric and in Section 2 of the Supplementary Material (Yanchenko and Hoff (2020)) for dynamics and timbre. Across all three metrics the  $\tau_p$  parameters were able to recover the different potentials for across-orchestra variation by piece, and the results correspond to musical expectation based on the score of each piece.

For example, for the tempo metric, the posterior mean of  $\tau_p$  for Symphony No. 6, Movement 1 (No6-01) was among the lowest for all pieces, while Symphony No. 9, Movement 2 (No9-02) was the highest posterior mean (Figure 6). This suggests that there is more potential for across-orchestra variation in tempo for No9-02 than for No6-01. Indeed, this corresponds to features of the musical score for these two pieces (International Music Score Library Project (2019)): piece No9-02 has approximately 24 marked tempo changes, while piece No6-01 has no marked tempo changes and only a fermata on the last note of the piece. (We include fermatas and grand pauses in our approximate count of marked tempo changes. A fermata indicates that a note should be held for a length of time determined by the conductor, while a grand pause indicates a break of complete silence in the piece for an amount of time again determined by the conductor.) Each tempo change denoted in the score is relative,

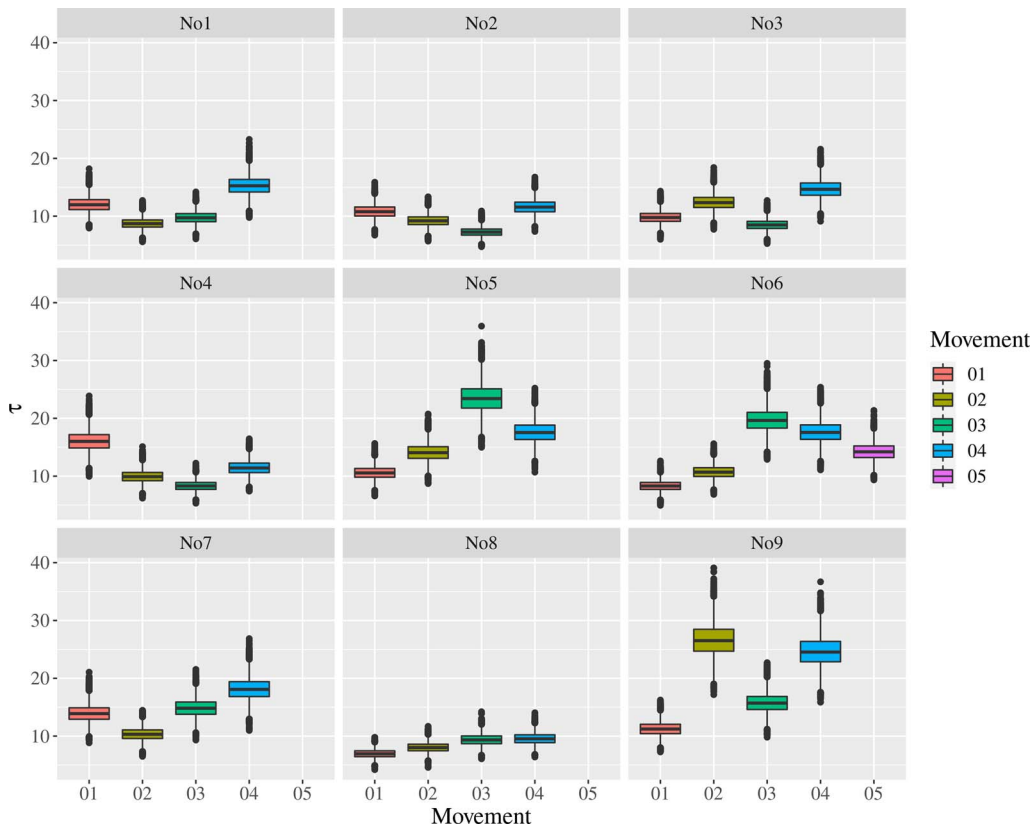


FIG. 6. Posterior distributions for the  $\tau_p$  parameters for the tempo metric by Beethoven symphony. The  $\tau_p$  parameters were able to recover the variation in the potential for across-orchestra differences by piece, for example, No9-02 as compared to No6-01.

thus allowing for a high level of variation in interpretation between different orchestras on pieces with many marked tempo changes. While tempo changes that are not written in the score can and do occur, the score is the “ground truth” and a proxy for the expected potential of variation. The posterior distributions of the  $\tau_p$  parameters for the dynamics and timbre metrics are summarized in Section 2 of the Supplementary Material (Yanchenko and Hoff (2020)) and also show a correspondence to musical markings in the score of each piece.

4.3. *Systematic differences between orchestras.* The primary motivation for this work was to explore systematic variation between orchestras across pieces for various musical metrics with the secondary goal of relating systematic differences to known characteristics of the orchestra and recording, such as the year of the recording. The latent distance parameters  $\{\delta_{ij} : j > i\}$  in the HMDS model capture this systematic variation and are the main parameters of interest for our application. From a musical perspective we do expect systematic differences between orchestras. For example, the Academy of Ancient Music performs on period instruments which sound different (in terms of dynamics and timbre) from the modern instruments used by the other orchestras considered here. Additionally, the musical metrics that we consider can be strongly influenced by the conductor, so we expect the two orchestras under Sir Simon Rattle (Berlin and Vienna) to be similar. As we will show, the HMDS model is able to recover these expected musical results and to additionally suggest some orchestral similarities that are more surprising.

For each metric we present a summary of the  $\delta_{ij}$ ’s as a heatmap and a dendrogram formed via hierarchical agglomerative clustering. For each heatmap, darker colors correspond to smaller values of  $\delta_{ij}$  which indicate that orchestra  $i$  and orchestra  $j$  are more similar to each other. For each dendrogram, orchestras that are more similar to each other are joined together at a lower height on the dendrogram. We focus on the posterior mean of each  $\delta_{ij}$ , as the posterior distribution of each  $\delta_{ij}$  is not skewed and fairly symmetric. Additionally, at each iteration of the Markov chain, the  $\delta_{ij}$  values are very close to the values  $\|X_i - X_j\|_2$  and thus satisfy the triangle inequality.

4.3.1. *Tempo.* A heatmap and dendrogram representing the estimated  $\delta_{ij}$ ’s for the tempo metric are provided in Figure 7(a) and Figure 7(b), respectively. As expected, the recordings by Sir Simon Rattle with the Vienna Philharmonic and the Berlin Philharmonic are quite

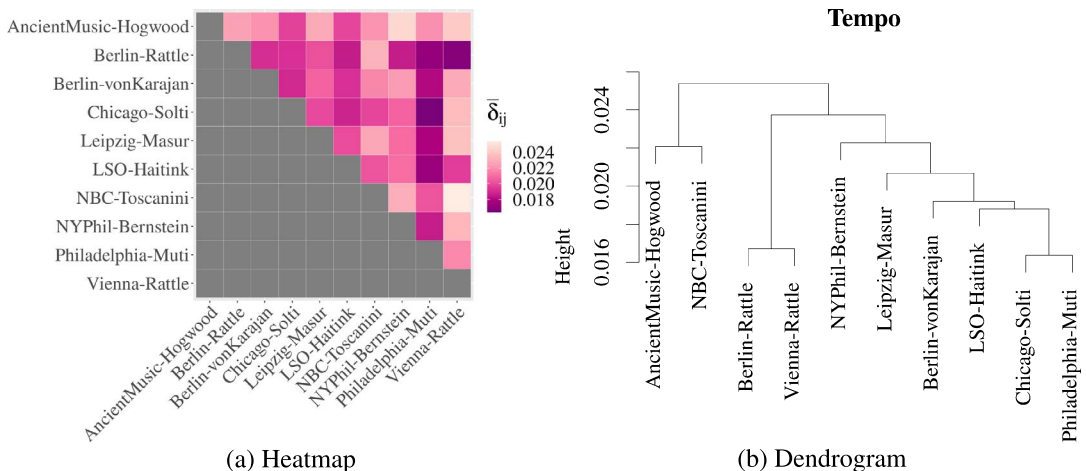


FIG. 7. (a) Heatmap of the posterior means of  $\delta_{ij}$  and (b) dendrogram found via hierarchical agglomerative clustering on the posterior mean of  $\delta_{ij}$  for the tempo metric. The two recordings by Rattle with Vienna and Berlin are very similar.

similar across pieces. These recordings were made within five years of each other under the same conductor, and the conductor has a large degree of control over the tempo of each piece. Additionally, the two Rattle recordings are also very similar to the LSO-Haitink recordings in terms of tempo. The data from these three orchestras are the most recent, in terms of the year in which the recordings were made. However, there does not appear to be any clear evidence of consistent similarity between orchestras by continent. Somewhat surprisingly, the Academy of Ancient Music and the NBC Symphony Orchestra are similar to each other and quite different from the other orchestras. This result can be interpreted in terms of the style of both orchestras and the scholarship of their conductors. The Academy of Ancient Music uses scholarship to imitate the way that these pieces would have been performed in Beethoven's time. The NBC Symphony Orchestra recordings, on the other hand, were conducted by Arturo Toscanini (born in 1867), who was a contemporary of musicians that were alive in Beethoven's time (Beethoven died in 1827) and who studied a style of conducting similar to that of the era when these pieces were composed.

4.3.2. *Dynamics*. The estimates of the  $\delta_{ij}$ 's for the dynamics metric also confirm prior musical expectations. In both the heatmap (Figure 8(a)) and the dendrogram (Figure 8(b)) for dynamics, the Academy of Ancient Music appears to be an outlier. This is expected, as Ancient Music is the only orchestra to record on period instruments which cannot play as loudly as modern instruments. Thus, the contrast in the range of dynamics (from loudest to softest volume) is smaller on period instruments than modern orchestral instruments. NBC also performed on older instruments (from the 1950s) which were not as loud as the modern orchestral instruments used by the remaining orchestras. Unlike the tempo metric, the dynamics metric shows evidence for some clustering by continent (Figure 8(b)). Vienna-Rattle, Berlin-von Karajan and LSO-Haitink are very similar, while all of the American orchestras are on the same main branch of the dendrogram. Finally, as shown in Figure 8(a), there is again evidence for higher similarity between more modern recordings, as Vienna-Rattle, Berlin-Rattle, LSO-Haitink and Philadelphia-Muti are very similar.

4.3.3. *Timbre/spectral flatness*. Finally, the analysis of the posterior means of  $\delta_{ij}$  for timbre again align with our musical expectation. Ancient Music is again an outlier for the timbre metric, due to the use of period instruments which have a fundamentally different

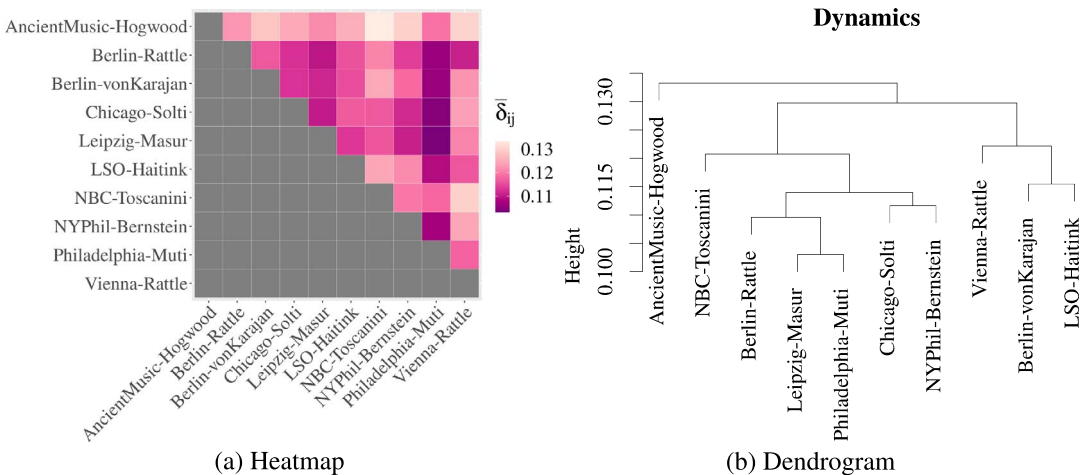


FIG. 8. (a) Heatmap of the posterior means of  $\delta_{ij}$  and (b) dendrogram found via hierarchical agglomerative clustering on the posterior mean of  $\delta_{ij}$  for the dynamics metric. The Academy of Ancient Music appears to be an outlier.



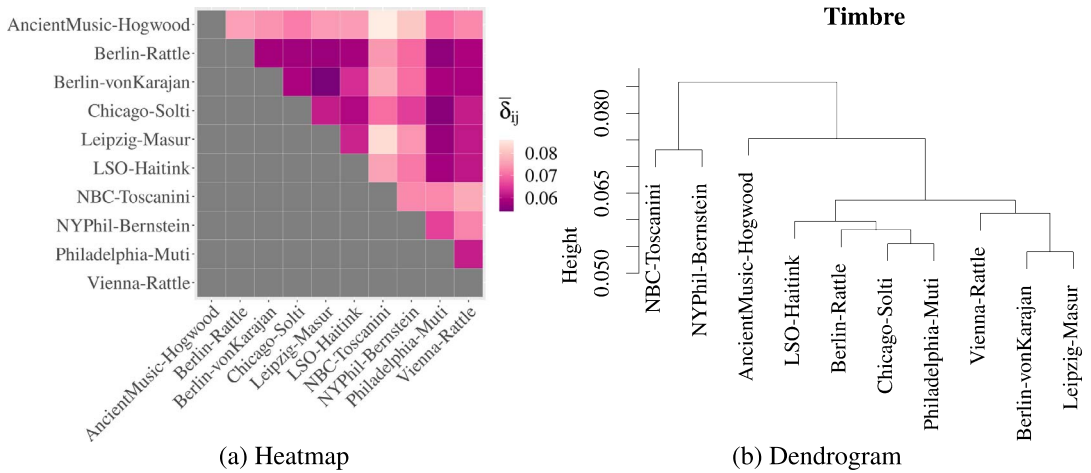


FIG. 9. (a) Heatmap of the posterior means of  $\delta_{ij}$  and (b) dendrogram found via hierarchical agglomerative clustering on the posterior mean of  $\delta_{ij}$  for the timbre metric. Newer recordings are more similar to each other than older recordings.

timbre than modern orchestral instruments (Figure 9(a)). Additionally, the NBC-Toscanini recordings and NY-Bernstein recordings are similar to each other and different from most of the other orchestras (Figure 9(b)), likely due to specifics of the recording technology at the time these recordings were made. Again, there is evidence for newer recordings being more similar to each other, as the LSO-Haitink, Berlin-Rattle, Chicago-Solti and Philadelphia-Muti are quite similar to each other in Figure 9(b). Finally, there is no consistent evidence for similarity by continent. While the NY-Bernstein recordings are similar to Chicago and Philadelphia in Figure 9(a), Philadelphia is also very similar to several European orchestras.

4.4. *Low-dimensional embedding.* While the focus of many classical MDS methods is often on dimensionality reduction of the learned embeddings, the primary purpose of HMDS is to share information and to identify differences (via the  $\delta_{ij}$  parameters) that are consistent across multiple distance matrices, rather than to provide a low-dimensional visualization of the distance data. For our application these estimated  $\delta_{ij}$ 's are not well represented by points in a Euclidean space with only two dimensions. For example, a principal components analysis on the  $X_i$  vectors indicates that all of the principal components are roughly equal in importance (Section 2 of the Supplementary Material (Yanchenko and Hoff (2020))).

We instead use a nonlinear dimensionality reduction approach to visualize the  $X_i$  vectors. The Procrustes transformation is first used to align each sampled  $X_i$  to the posterior mean for each metric individually. The aligned  $X_i$  vectors lie in nine-dimensional space, and a t-SNE (van der Maaten and Hinton (2008)) embedding is then used to visualize the  $X_i$  vectors in two-dimensional space. The t-SNE embedding is able to retain local and global structure in the data through the dimensionality reduction by keeping points that are close in the higher dimensional space close in the lower dimensional space (van der Maaten and Hinton (2008)). The posterior distributions for the embedded  $X_i$  vectors separate well by orchestra and are arranged similarly to the findings with the hierarchical clustering for  $\delta_{ij}$  above (Figure 10). For example, for the tempo metric the  $X_i$  vectors for Berlin-Rattle and Vienna-Rattle are quite close to each other, indicating that the recordings by these two orchestras are similar in terms of tempo. The  $X_i$  vectors for NBC and the Academy of Ancient Music are also close to each other, again indicating similarity in terms of tempo for these two orchestras.

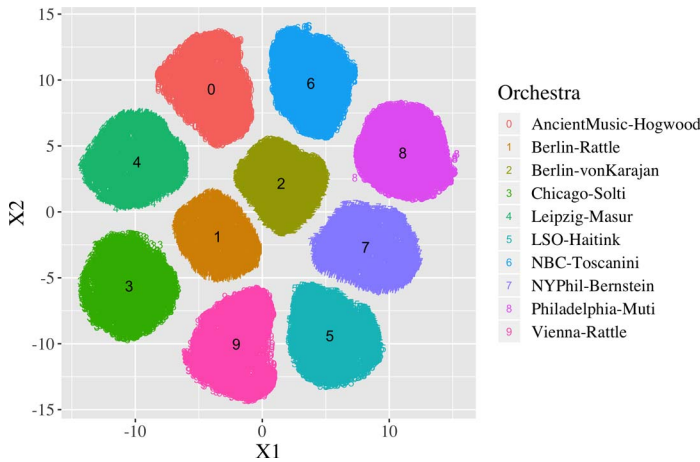


FIG. 10. A  $t$ -SNE embedding for the posterior distribution of the  $X_i$ 's for the tempo metric. Each point represents an iteration of the Markov chain for  $X_i$  embedded in a two-dimensional space. The similarity between orchestras corresponds to the results for the hierarchical clustering of the posterior means for  $\delta_{ij}$ . Additionally, the  $X_i$  vectors separate well by orchestra.

**5. Conclusions and future work.** In order to quantify systematic differences between orchestras, we developed a hierarchical, model-based approach to multidimensional scaling. This method generalized the BMDS model of Oh and Raftery (2001) in several ways, including the extension to modeling heterogeneous replicate distance matrices. We applied HMDS to the comparison of different orchestras across the Beethoven symphonies. The proposed HMDS model was successful in uncovering systematic differences between orchestras across pieces, and the  $\tau_p$  parameters were able to capture variation in the potential for across-orchestra differences. The overall analysis of the posterior means for  $\delta_{ij}$ , the systematic dissimilarity between orchestras  $i$  and  $j$ , for the three musical metrics yielded some expected and surprising results. As expected, the recordings by Vienna-Rattle and Berlin-Rattle were found to be very similar across all three musical metrics, but especially tempo, upon which the conductor has a large influence. Additionally, the Academy of Ancient Music was confirmed as an outlier in terms of dynamics and timbre, due to their use of period instruments. Surprisingly, we found that NBC and the Academy of Ancient Music were most similar to each other in terms of tempo, and this might be attributed to their adherence to artistic styles from Beethoven's time.

However, a few other interesting results also emerged. Across all three of the musical metrics, Philadelphia-Muti and Leipzig-Masur were quite similar. This was not an obvious result, as Riccardo Muti and Kurt Masur had different conducting backgrounds and experiences. Additionally, the Philadelphia Orchestra and the Leipzig Gewandhaus Orchestra are both among the most well known American and European orchestras, respectively, with very different histories. Philadelphia is known for the "Philadelphia Sound," developed under Leopold Stokowski in the early 1900s, while Leipzig has a long and storied history of conductors, including Felix Mendelssohn.

Finally, the evidence for more similarity across tempo, dynamics and timbre in the more recent recordings was not necessarily expected a priori (Liebman, Ornoy and Chor (2012)). While some of this is due to changes in recording technology and the fact that Rattle conducted both the Vienna and Berlin Philharmonics, this result suggests that there may be less variation among newer recordings by different orchestras, as compared to older recordings.

It is important to note that in terms of the musical application of interest, tempo was the only metric considered that was independent of the recording technology used for each orchestra's recordings and, from that perspective, was the most indicative of artistic differences

between orchestras. Dynamics and timbre both depend on the specifics of recording technology and are also related to each other, as louder instruments will have different timbres than softer overall instruments. Additionally, it should be noted that the different orchestras considered performed on different brands and makes of instruments which also contributed to the differences in dynamics and timbre. Thus, future work could include additional and more nuanced musical metrics to further support our findings here. Alternative musical metrics would also be important to explore in the case of other composers, particularly composers with less potential for variation along the musical metrics considered here.

Future extensions to the proposed HMDS model could consider a hierarchical extension to share information across metrics and learn an overall distance for each orchestra pair. Other extensions could include exploring constraints on the parameter space for the  $X_i$  vectors to allow for identifiability of these parameters, such as in Bakker and Poole (2013). Additionally, while HMDS is designed for the quantitative comparison of orchestral audio recordings, the model is flexible enough to be of use in several different application domains with similar data structures. Of critical importance here is that comparing entity  $i$  on replication  $p$  to entity  $j$  on replication  $p'$  does not make sense, meaning that approaches similar to distance-based ANOVA (Anderson (2001), McArdle and Anderson (2001)) cannot be used. Thus, HMDS is especially applicable for medical settings, for example, where there are many different types of measurements made (replications  $p$ ) but where each type of measurement cannot be compared directly to another, and, instead, the question of interest is the existence of systematic similarities between patients or individuals across measurement types. Furthermore, the  $\tau_p$  parameters allow for different potentials of variation for each measurement type, a fact that may be especially relevant in many medical applications as well.

**Acknowledgments.** The authors would like to thank Mike Kris for helpful discussions about the interpretation of results in a musical context. The authors would also like to thank the four anonymous referees and the Editor for their constructive and helpful comments that improved the quality of this paper.

## SUPPLEMENTARY MATERIAL

**Supplement to “Hierarchical multidimensional scaling for the comparison of musical performance styles”** (DOI: [10.1214/20-AOAS1391SUPPA](https://doi.org/10.1214/20-AOAS1391SUPPA); .pdf). We provide details of one possible Markov chain Monte Carlo algorithm for inference in the HMDS model, and additional pre-processing details, model diagnostics and results.

**Source code and data for “Hierarchical multidimensional scaling for the comparison of musical performance styles”** (DOI: [10.1214/20-AOAS1391SUPPB](https://doi.org/10.1214/20-AOAS1391SUPPB); .zip). R code for the proposed HMDS model and data for the comparison of Beethoven symphonies across 10 orchestras and multiple musical metrics.

## REFERENCES

- ANDERSON, M. J. (2001). A new method for non-parametric multivariate analysis. *Austral Ecology* **26** 32–46.
- BAKKER, R. and POOLE, K. T. (2013). Bayesian metric multidimensional scaling. *Polit. Anal.* **21** 125–140.
- BORG, I., GROENEN, P. J. F. and MAIR, P. (2018). *Applied Multidimensional Scaling and Unfolding*. 2nd ed. *SpringerBriefs in Statistics*. Springer, Cham. MR3793129 <https://doi.org/10.1007/978-3-319-73471-2>
- CARROLL, J. D. and CHANG, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart–Young” decomposition. *Psychometrika* **35** 283–319.
- COOK, N. (2005). Towards the complete musicologist? Invited Talk at the International Symposium on Music Information Retrieval.
- DESAIN, P. and HONING, H. (1994). Does expressive timing in music performance scale proportionally with tempo? *Psychol. Res.* **56** 285–292.

- ELLIS, D. (2007). Chroma feature analysis and synthesis. <https://labrosa.ee.columbia.edu/matlab/chroma-ansyn/>.
- FONG, D. K. H., DESARBO, W. S., CHEN, Z. and XU, Z. (2015). A Bayesian vector multidimensional scaling procedure incorporating dimension reparameterization with variable selection. *Psychometrika* **80** 1043–1065. MR3423037 <https://doi.org/10.1007/s11336-015-9449-x>
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 <https://doi.org/10.1214/06-BA117A>
- GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677
- GOEBL, W. and WIDMER, G. (2009). On the use of computational methods for expressive music performance. In *Modern Methods for Musicology: Prospects, Proposals and Realities* 93–113. Ashgate, London.
- GRACHTEN, M., GASSER, M., ARZT, A. and WIDMER, G. (2013). Automatic alignment of music performances with structural differences. In *Proceedings of the 14th International Society for Music Information Retrieval Conference* 607–612.
- INTERNATIONAL MUSIC SCORE LIBRARY PROJECT (2019). IMSLP: Petrucci Music Library. [https://imslp.org/wiki/Category:Beethoven,\\_Ludwig\\_van](https://imslp.org/wiki/Category:Beethoven,_Ludwig_van).
- KAMMERL, J., BIRKBECK, N., INGUVA, S., KELLY, D., CRAWFORD, A. J., DENMAN, H., KOKARAM, A. and PANTOFARU, C. (2014). Temporal synchronization of multiple audio signals. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 4603–4607.
- KIRCHHOFF, H. and LERCH, A. (2011). Evaluation of features for audio-to-audio alignment. *J. New Music Res.* **40** 27–41.
- KRUSKAL, J. B. (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Rev.* **25** 201–237. MR0700350 <https://doi.org/10.1137/1025045>
- KUNSTDERFUGE. COM (2018). Franz Liszt transcriptions. <http://www.kunsterfuge.com/liszt.htm#Transcriptions>.
- LIEBMAN, E., ORNOY, E. and CHOR, B. (2012). A phylogenetic approach to music performance analysis. *J. New Music Res.* **41** 195–222.
- LIEM, C. C. S. and HANJALIC, A. (2011a). Expressive timing from cross-performance and audio-based alignment patterns: An extended case study. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* 519–524.
- LIEM, C. C. S. and HANJALIC, A. (2011b). Expressivity in musical timing in relation to musical structure and interpretation: A cross-performance, audio-based approach. In *Proceedings of the 42nd International AES Conference on Semantic Audio* 255–264.
- LIEM, C. C. S. and HANJALIC, A. (2015). Comparative analysis of orchestral performance recordings: An image-based approach. In *Proceedings of the 16th International Society for Music Information Retrieval Conference* 302–308.
- LIN, L. and FONG, D. K. H. (2019). Bayesian multidimensional scaling procedure with variable selection. *Comput. Statist. Data Anal.* **129** 1–13. MR3855195 <https://doi.org/10.1016/j.csda.2018.07.007>
- LOGAN, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval*.
- MCARDLE, B. H. and ANDERSON, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* **8** 290–297.
- MINAS, C. and MONTANA, G. (2014). Distance-based analysis of variance: Approximate inference. *Stat. Anal. Data Min.* **7** 450–470. MR3303162 <https://doi.org/10.1002/sam.11227>
- MÜLLER, M. (2015). *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, Cham. MR3382223 <https://doi.org/10.1007/978-3-319-21945-5>
- OH, M.-S. and RAFTERY, A. E. (2001). Bayesian multidimensional scaling and choice of dimension. *J. Amer. Statist. Assoc.* **96** 1031–1044. MR1947251 <https://doi.org/10.1198/016214501753208690>
- PARK, J., DESARBO, W. S. and LIECHTY, J. (2008). A hierarchical Bayesian multidimensional scaling methodology for accommodating both structural and preference heterogeneity. *Psychometrika* **73** 451–472. MR2447325 <https://doi.org/10.1007/s11336-008-9064-1>
- PENEL, A. and DRAKE, C. (1998). Sources of timing variations in music performance: A psychological segmentation model. *Psychol. Res.* **61** 12–32.
- PEPERKAMP, J., HILDEBRANDT, K. and LIEM, C. C. S. (2017). A formalization of relative local tempo variations in collections of performances. In *Proceedings of the 18th International Society for Music Information Retrieval Conference* 158–164.
- RIZZO, M. L. and SZÉKELY, G. J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Stat.* **4** 1034–1055. MR2758432 <https://doi.org/10.1214/09-AOAS245>
- SAPP, C. S. (2007). Comparative analysis of multiple musical performances. In *Proceedings of the 8th International Conference on Music Information Retrieval* 497–500.

- SAPP, C. S. (2008). Hybrid numeric/rank similarity metrics for musical performance analysis. In *Proceedings of the 9th International Conference on Music Information Retrieval* 501–506.
- STAN DEVELOPMENT TEAM (2019a). RStan: The R interface to Stan. R package version 2.19.2.
- STAN DEVELOPMENT TEAM (2019b). Stan Reference Manual, 2.21 ed.
- SUEUR, J. (2018). *Sound Analysis and Synthesis with R*. Springer, Berlin.
- SUEUR, J., AUBIN, T., SIMONIS, C., LELLOUCH, L., BROWN, E. C. et al. (2018). Package ‘seewave’: Sound Analysis and Synthesis. <https://cran.r-project.org/web/packages/seewave/seewave.pdf>.
- THORNBURG, H. D., LEISTIKOW, R. J. and BERGER, J. (2007). Melody extraction and musical onset detection from framewise STFT peak data. *IEEE Trans. Audio Speech Lang. Process.* **15** 1257–1272.
- TORGERSON, W. S. (1952). Multidimensional scaling. I. Theory and method. *Psychometrika* **17** 401–419. MR0054219 <https://doi.org/10.1007/BF02288916>
- VAN DEN OORD, A., DIELEMAN, S. and SCHRAUWEN, B. (2013). Deep content-based music recommendation. In *Advances in Neural Information Processing Systems* 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 2643–2651. Curran Associates, Red Hook.
- VAN DEN OORD, A., DIELEMAN, S., ZEN, H., SIMONYAN, K., VINYALS, O., GRAVES, A., KALCHBRENNER, N., SENIOR, A. W. and KAVUKCUOGLU, K. (2016). WaveNet: A generative model for raw audio. CoRR abs/1609.03499.
- VAN DER MAATEN, L. and HINTON, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** 2579–2605.
- VSEVOLOZHSKAYA, O. A., ZAYKIN, D. V., GREENWOOD, M. C., WEI, C. and LU, Q. (2014). Functional analysis of variance for association studies. *PLoS ONE* **9** 1–13. <https://doi.org/10.1371/journal.pone.0105074>
- YANCHENKO, A. K. and HOFF, P. D. (2020). Supplement to “Hierarchical Multidimensional Scaling for the Comparison of Musical Performance Styles.” <https://doi.org/10.1214/20-AOAS1391SUPPA>, <https://doi.org/10.1214/20-AOAS1391SUPPB>