No cover image available

The Oxford Handbook of Music and Corpus Studies

Daniel Shanahan (ed.) et al.

https://doi.org/10.1093/oxfordhb/9780190945442.001.0001

Published: 2022 **Online ISBN:** 9780190945473 **Print ISBN:** 9780190945442

CHAPTER

Statistical Methods in Music Corpus Studies: Application, Use Cases, and Best Practice Examples

Daniel Müllensiefen, Klaus Frieler

https://doi.org/10.1093/oxfordhb/9780190945442.013.8 Pages C8.S1–C8.N2

Published: 18 August 2022

Abstract

In this chapter, the authors explain that there are two common goals in musical corpus analysis. The first is the description and comparison of musical corpora, the second is to establish relationships between musical structures and extra-musical data, which can refer to metadata of a particular musical piece (genre, style, and period labels, composer and performer attributions, etc.) or to listeners' perceptions and evaluations. The authors give a brief overview of basic and advanced statistical methods that have been employed in music corpus studies. The chapter covers descriptive statistics and visualizations, feature selection and aggregation using principal component analysis. In addition, random forests and linear regression methods for use in the context of corpus studies are briefly explained, as well as supervised and unsupervised classification techniques. Each topic and method is introduced with a conceptual explanation, suggestions for its application, and usage scenarios from the research literature.

Introduction

There are two common goals in musical corpus analysis. The first goal is the description and comparison of musical corpora, the second is to establish relationships between musical structures and extra-musical data. Extra-musical data can refer to metadata of a particular musical piece (genre, style, and period labels, composer and performer attributions, etc.) or to listeners' perceptions and evaluations (e.g., from perceptual experiments, preference indications, or by commercial success). The goal is to find statistical regularities that apply to not only one song, but instead also to a larger collection of musical pieces, a musical corpus as a whole, or even to music across different corpora. To achieve this goal, the necessary ingredients of research are one or more musical corpora as well as a mechanism for extracting meaningful features or descriptors of musical structure. We use the term "features" here in a very inclusive way to designate any system of attributes that can be derived from the musical representation system used to encode the music of the corpus itself. Related and mostly synonymous terms are "properties," "traits," "characteristics," and "descriptors." In addition, reliable extra-musical data are necessary so that they can be compared to and associated with musical features. The tool for relating musical features and extra-musical data are statistical methods. Statistical analysis enables the researcher to discover and describe regularities in musical data and is thus a central component for making sense of data in musical corpus studies. While it seems obvious that different research questions require different statistical methods, there is a danger that researchers are biased toward answering questions that lend themselves to familiar analysis methods, reflected in the popular wisdom that "when all you have is a hammer, every problem looks like a nail." Furthermore, as there is a general tradeoff between simplicity and precision of statistical models, these have to be chosen carefully in terms of research aims and feasibility. In order to raise awareness for the range of statistical analysis methods for musical corpus studies, we hope to provide a brief overview of a few common research scenarios together with suitable statistical approaches and specific techniques that can help to answer the corresponding research questions.

This chapter is neither intended to be exhaustive, nor can any of the statistical methods be covered in sufficient detail to serve as a practical how-to guide for running the analysis. Instead, the aim and scope of this chapter is to introduce typical research scenarios and use cases in musical corpus studies and then present best-practice options for statistical analysis. However, for each use case and corresponding analysis methods, references to in-depth treatments of the underlying statistical theory and practical application guides are provided. We assume that the reader has some grounding with quantitative methods including descriptive statistics and statistical tests, as well as linear regression and correlation, at the level of any basic undergraduate statistics textbook as presented, for example, in Field, Miles, and Field (2012).

The use cases described in the rest of the chapter span typical scenarios in music corpus studies. Early analysis stages often include exploration and visualization of musical features as well as feature aggregation and selection, and these often provide the foundation on which later analysis stages build. Supervised classification or regression are typical use cases for describing how musical features are associated with external data (e.g., how melodic or rhythmic features of pop tunes are related to their catchiness or their propensity to become an earworm). In contrast, cluster analysis exploits similarities and associations within the feature distributions of a musical corpus and aims to discover or introduce internal structures (clusters) among musical objects.

Visualization and Exploration

Once features are computed for the pieces of a musical corpus, one of the first steps of analysis should be the visualization of the data and their exploration by graphical devices and descriptive summary statistics. Tukey (1977) explains the benefits of exploratory data analysis and visualization fabulously in one of the first books exclusively devoted to the topic. With a perspective on music corpus studies, there are four primary reasons for visualizing and exploring musical corpus data before any further analyses are carried out.

1. From a graphical display of musical feature data, it is much easier and quicker to identify potential errors in the data, such as values outside of theoretical range or anomalies like a large proportion of identical values in a feature variable, than it would be by looking at the numbers alone. Data artifacts like these are not uncommon in musical corpus studies and can arise, for example, from coding mistakes when music is hand-coded or converted automatically into a machine-readable format (Huron 1988) or from faulty implementations of musical features. Visualizations often help to spot and fix errors quickly and are an essential tool for data tidying.

2. Visualizations of musical feature data provide information about the distribution of a musical feature variable and tell the analyst, for example, whether a feature is heavily skewed to one side, has a flat or a spiky distribution, or a close-to-normal (i.e., Gaussian or "bell curve") distribution. The shape of a feature distribution can be a parameter of interest by itself, and documenting and comparing the distributions of certain features (e.g., pitch range) across different genres or historical periods can be a musicological result in its own right. In addition, the shape of a musical feature distribution can suggest the need for data transformations that might be necessary for the use in subsequent statistical analysis models (linear models, e.g., assume Gaussian error terms as well as a linear relationship between predictor variables and the dependent variable; see section 4.1.6 in Tabachnick and Fidell, 2014, among others).

3. Visualizing the musical feature data of two or more variables on the same graph can also serve to explore interesting structures in data that provide evidence for the central research question and that were or were not hypothesized a priori. If prior hypotheses claim a substantial association between two variables, then this should be visible on a suitable graph. In this sense, data visualization is complementary to statistical hypothesis testing and can be a very convincing component of confirmatory data analysis. In fact, it has been argued that graphs are often better representations of scientific evidence than *p* values from significance testing (e.g., Smith et al. 2000).

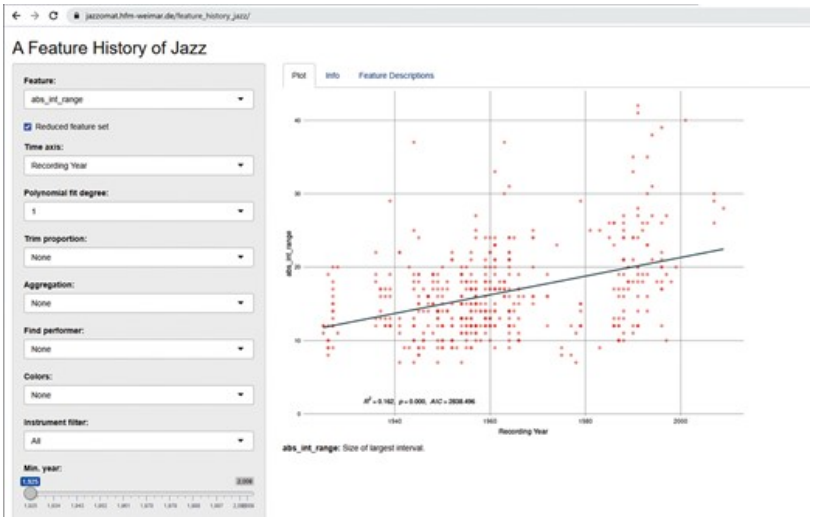
In contrast, when data visualizations are generated without a concrete hypothesis in mind but rather for the purpose of discovering novel structures in the data, then this analysis is considered exploratory. *Exploratory analysis* using graphical methods can be an extremely powerful way of gaining novel scientific insights and for generating new hypotheses. But exploring the data in this way is different from confirming a priori hypotheses for which one seeks empirical evidence, and it is absolutely necessary to label exploratory (graphical) analysis as such. In relatively new and emerging fields such as corpus-based musicology, however, strong theories and precisely testable hypotheses are relatively rare. Hence, data-driven exploration is an important first step and can sometimes generate the main result of a musical corpus study (e.g., the prevalence of the melodic arch contour in Western folksongs; Huron 1996).

In sum, it is good practice to start a music corpus study (or any quantitative study for that matter) with a descriptive and visual exploration of the dataset. Unfortunately, plots of raw feature distributions are rare in journal articles, and data are often reported in heavily aggregated form (e.g., reduced to means and standard deviations) which only characterize a few aspects of a distribution. However, recent trends in Open Science

(Munafò et al. 2017) and electronic publishing are facilitating the addition of rich data plots along with a journal article, if only in supplementary materials or accompanying web repositories.

One attractive alternative to supplementary plots of journal articles are *web apps*, such as the Shiny software framework, which can open up a dataset to the research community and the general public. Shiny is a package for designing interactive websites for the statistical programming environment R (Chang et al. 2017; R Core Team 2008). It is very well integrated in the RStudio development environment (RStudio Team 2016) and comparatively easy to use, enabling the combination of common data processing routines with the creation of interactive web interfaces. These web interfaces can be run on a local computer or launched on the internet. Shiny apps can be built incrementally by adding further analysis options after the web app has been launched. Web apps also facilitate the incorporation of download links for raw and processed data as well as analysis code and hence the implementation of many good Open Science practices. Web apps not only serve as a perfect supplement to a paper, but can also serve as a handy analysis resource during the writing process.

Figure 8.1



Screenshot of the Feature History Explorer Shiny web app (https://jazzomat.hfm-weimar.de/feature_history_jazz/) showing how the range of absolute intervals in jazz solos generally increases over time.

As an example, we describe the Shiny app "Feature History Explorer,"[1] which was developed in the context of the paper "A Feature History of Jazz Solo Improvisation" (Frieler 2018). The goal of this corpus study was to trace the development of features of jazz solo improvisations over time and identify well-known stylistic periods as well as to potentially uncover new periods of stylistic change or stability. To this end, a large set of numerical features describing melodic and rhythmic aspects of jazz solos was extracted from all 456 transcriptions in the Weimar Jazz Database (WJD; Pfleiderer 2017), along with relevant metadata, most notably the year of recording.

Figure 8.1 shows the central features of the Shiny app and its central "Plot" tab where a chosen feature (*y* axis) is plotted against a time variable (*x* axis). The data for this bivariate plot can be filtered and manipulated according to several different criteria, listed in the lower part of the navigation panel on the left. The resulting scatterplot includes a regression line and statistics for the corresponding linear (or polynomial) regression model. Plots and models are generated on the fly by the underlying R routines.

The filtering options in the navigation panel on the left-hand side allow the user to select individual instrument groups as well as the range of the recording year. In addition, users can preprocess the data and customize the plot in various ways. The web app allows the download of all features and meta-data for use

with different data processing tools and also includes descriptions of the most important features, including links to further documentation. It is also possible to include an HTML version of the companion paper (Frieler 2018) with direct links to plots and tables generated from the app. However, copyright restrictions often limit the degree of integration of journal paper and web app.

As a demonstration, Figure 8.1 shows a scatter plot of the feature "abs_int_range" against recording year. This is a simple feature that represents the range of the absolute interval size in a solo (i.e., the difference between the largest and the smallest semitone interval disregarding the sign [direction]). The *y*-axis of the graph ranges up to 42 (i.e., about 3½ octaves, which is the largest interval to be found in any solo of the WJD). This is a possible value (at least for some instruments), and hence there are no obvious data artifacts visible in this plot. But Figure 8.1 also shows that the WJD does not cover the whole time range equally, but rather that there is a high density of solos from the 1940s to the 1960s and again after 1985, while recordings from the 1970s are underrepresented. Hence, the graphical distribution of recordings across time provides evidence for the discussion of important questions about sampling strategies and the representativeness of music corpora (see Pfleiderer [2017] for a more in-depth discussion of the choices underlying the WJD).

Finally, the graph shows a clear increasing trend of the absolute interval range over time. This is in line with the hypothesis of a general trend toward higher virtuosity in jazz solos because the choice of large intervals has often been stylistically associated with greater virtuosity in recent decades (Frieler 2018). But filtering the data by instrument group reveals that instrument subgroups within the corpus also play a role: it is generally easier to play large intervals on the saxophone than on brass instruments, and the tenor saxophone is the largest instrument group in the WJD. In fact, this trend toward larger intervals is mainly driven by tenor saxophonists in this particular corpus.

The example demonstrates that data visualization can be a powerful technique for data exploration. Interactive visualization tools, such as the R/Shiny app presented here, can help to identify obvious artifacts in the data, gain insight into variable distributions and the underlying sampling process, and provide visual evidence for or against prior hypotheses and for the generation of novel hypothesis in an exploratory manner. In sum, data visualizations represent a view on the musical data that is a necessary complement to any subsequent statistical analyses.

Feature Selection and Aggregation

Features in computational musical research are commonly represented as numerical or categorical values and are derived from the digital music representation in which a corpus is encoded (e.g., MIDI, kern, MusicXML). Features such as the range of intervals in a melodic line are often constructed by a series of transformation and aggregation operations over a set of musical events. In the case of the range of intervals used in a melodic line (see description of feature "abs_int_range" in the previous section), this involves (a) assigning numerical values to the pitches of a melodic line, (b) calculating the distance between successive note events in terms of semitones, and (c) determining the maximum of the calculated distance values. In this case only three operations are necessary to define the feature. But many features are constructed from longer chains of transformations and aggregation operations. The combinatorial nature of the feature construction process gives rise to a potentially very large number of features that can be constructed to describe a given corpus. Feature computation toolboxes like FANTASTIC (Müllensiefen 2009), jSymbolic2 (McKay et al. 2017), the MIDIToolbox (Eerola and Toiviainen 2004), or the MeloSpyGUI (Pfleiderer et al. 2017) provide a vast number of features. Many of these features will be variants of the same idea and aim to measure the same music-analytic construct, but analytic theory is rarely precise enough to advise the researcher which feature variant to choose for subsequent analyses.

There are two standard approaches for dealing with a large set of features which may include many correlated features: feature aggregation and feature selection.

Feature Aggregation

Baayen (2008, ch. 5) provides a concise treatment of the rationale for using principal component analysis (PCA) and factor analysis (FA) as feature aggregation techniques for (linguistic) corpus studies with examples in R. In general, feature aggregation aims to combine features that are intercorrelated into components (PCA) or factors (FA). The main idea behind feature aggregation is that correlated variables carry very similar information and, therefore, are partly redundant. These aggregation techniques exploit correlations and redundancies in features sets and reduce the number of variables from many raw features to a few aggregated components or factors (Baayen 2008).

PCA components and FA factors are specific linear combinations of features such that the intercorrelations among components or factors are minimal ("principal"), and thus the components or factors can be thought to represent different, independent aspects of the data. To understand why linear combinations can create minimal correlations, first note that combining features into weighted sums of features is equivalent to rotating the axes used to represent the features. Now consider two strongly correlated features (i.e., two features for which the data points lie near a single line in the plane). If one rotates the axes so that one axis coincides with this line, then the data points along this new axis will have a large amount of variation ("variance"). The variance along the other axes will be much lower, and, as such, if the other axes are dropped, the loss of information is minimal. In this sense the number of features can reduced while still keeping the most important information. This procedure can be generalized to arbitrarily many dimensions.

The components or factors resulting from this data reduction process can then be used as predictors in subsequent analysis models (e.g., regression models that can only estimate unbiased model coefficients if their predictor variables are not highly correlated). Often dozens or hundreds of features can be reduced to only a handful of components or factors, which usually have greater explanatory power than the individual feature variables. Generally, PCA and FA are conceptually and mathematically related and are often used for similar purposes. Revelle (2018) discusses similarities and differences between the two techniques, for example, the assumption, exclusive to FA, that a latent (i.e., unobserved) factor is causing the correlations between observed variables. PCA is employed much more frequently in the context of musical corpus studies, and so we will only refer to PCA as a feature aggregation technique in the remainder of this section. Most aspects of PCA modeling that we describe, however, also apply to FA models.

There are a number of challenges in the construction of PCA models. First, the researcher has to decide on the optimal number of components for a given dataset. There are at least eight different criteria for deciding the optimal number of components. However, these criteria rarely all agree when used with real datasets. Revelle (2018) discusses their respective merits and downsides. Several criteria make use of the eigenvalues of the correlation matrix, which are related to the variance in the correlational data that a component is able to explain. For example, the widely used *Kaiser criterion* suggests that we consider only PCA solutions where all components have an eigenvalue of greater than 1. Revelle (2018) recommends assessing the optimal number of components on a given dataset by several criteria and aiming to understand why they might disagree.

The second common challenge for PCA modeling is the interpretation and naming of the resulting components. This interpretation process is not part of the statistical procedure but an important aspect for establishing the validity of the PCA model and for its communication and dissemination. The meaning of each component should generally summarize the features that contribute most strongly to this component (i.e., features with large "component loadings"). In addition, the interpretation of the overall PCA solution,

comprising all components, should be plausible. From this perspective the interpretation of the solution can also serve as a criterion for deciding the optimal number of components.

Van Balen, Burgoyne, Bountouridis, Müllensiefen, and Veltkamp (2015) used PCA to reduce forty-four audio features to twelve components and fifty-eight symbolic features to also twelve components (i.e., achieving a data reduction factor of about 4) in a study on the features that would predict the "catchiness" of excerpts from a corpus of commercial pop songs. Catchiness was measured as recognition speed and accuracy with participants in a perceptual experiment. Here, the authors used parallel analysis to decide on the number of components necessary. Parallel analysis was first introduced by Horn (1965) and compares the eigenvalues of PCA or FA solutions for a given dataset to solutions using randomized versions of the same dataset not containing any meaningful correlational structure. Parallel analysis suggests accepting a model where all components have eigenvalues larger than the eigenvalues for the corresponding components extracted from the parallel but randomized datasets. Van Balen et al. (2015, table 1) also provided meaningful interpretations for the twelve components that describe the musical aspect summarized (e.g., harmony, melody), the type of measure (e.g., intensity, entropy) applied, and whether background information from the musical corpus was used in the construction of the features. Van Balen et al. (2015) then computed component scores for all music excerpts (i.e., values of the new aggregated features). The component score variables entered a subsequent linear regression model as predictor variables. Finally, backward elimination was used as a variable selection strategy (see later discussion) to eliminate all component score variables without significant power to explain catchiness. Their final model combines two audio feature components ("vocal prominence" and "melodic range conventionality") and two symbolic components ("melodic repetitivity" and "melodic/bass conventionality") to predict participants' performance on the recognition task. In summary, the purpose of reducing the large number of audio and symbolic music features to only twelve components in this study was to combine features that are highly correlated onto the same component and feed a relatively small number of uncorrelated predictor variables to the final regression model. The latter is often necessary to obtain a more robust output from any variable selection procedure and to obtain unbiased estimates for predictor variable coefficients.

Feature Selection

Unlike PCA or FA, feature selection procedures do not aggregate the information from several features but instead aim to select only the feature variables with the largest explanatory power within the context of a supervised regression or classification model. A very simple approach is to compute, for example, regression models for all possible subsets of feature variables and choose the subset of predictors that produces the best model fit on the given dataset. However, the exhaustive search for the best subset of feature variables can become computationally intractable for large numbers of features because the number of subsets grows factorially with the number of feature variables. In addition, the exhaustive search approach is also very likely to pick up on random variations in the data, giving rise to overfitting the model on the particular dataset, which impacts negatively on the model's generalizability to new datasets. Therefore, exhaustive search procedures are rarely used for item selection in practice unless the number of feature variables is very low (e.g., up to four).

Stepwise variable selection procedures represent a widespread and practical alternative to the exhaustive search through all possible subsets and are implemented in many software packages (e.g., R or SPSS). In stepwise selection procedures each variable is assessed for its explanatory power only once. The assessment is carried out in a serial fashion, starting either with the variable that is most closely (forward selection) or least closely (backward selection starting from a full model) associated with the outcome variable to be predicted. *Forward selection* starts from a model without any predictor variables, the so-called *null model*, which means "predicting" the dependent variable with its average value, and adds the strongest predictors in a step-by-step way. *Backward selection* starts from a full model including all available predictors and

removes the weakest predictors first. At each step the model with the assessed predictor is compared against the model without this predictor. Comparisons are typically done via statistical tests (e.g., likelihood ratio, F-test) that allow the researcher to decide whether the predictor makes a significant contribution to explaining variance in the outcome variable above and beyond the other predictor variables in the model. If its contribution to the model explanatory power is significant, then the predictor is added/retained in the model or otherwise discarded. The stepwise selection process ends when the addition/removal of variables from the model no longer increases its fit significantly. Backward selection procedures seem to be more common in practice, but a general recommendation is to run both a backward and forward selection on the same dataset and observe whether both strategies produce the same result (i.e., select the same subset of predictors). In their study on catchiness, Van Balen et al. (2015) used a stepwise selection procedure based on significance tests. To avoid overfitting, they set the significance level to a conservative value of $\alpha = .005$, which accounts for the fact that twelve significance tests are required for the stepwise assessment of the twelve PCA component as predictor variables.

An alternative to the assessment by significance tests is the use of information criteria, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). Here predictors are added to—or removed from—the model as long as they optimize the information criterion (beyond a given margin). Both AIC and BIC have a solid theoretical foundation (Raftery 1995) and balance the fit of the model to the data with a penalty for the complexity of the model (i.e., the number of predictor variables). The BIC also considers sample size and often delivers more robust results for large datasets.

Model-Based Feature Selection

A drawback of all stepwise selection procedures is that they are not guaranteed to identify the optimal subset of predictor variables because each predictor is only assessed at one point in the sequence of steps for its explanatory power. In addition, models derived from step-wise selection procedures may not generalize well to new datasets because the selection process is always based on the same dataset. *Model-based feature selection* is an alternative to running a potentially long series of stepwise model comparisons and enables the comparative assessment of the full set of predictor variables at the same time. This can be done in a linear regression model by shrinking the coefficients of weaker predictor variables to zero, which eliminates them from the regression model. Technically, this means finding a linear regression model with an additional constraint that the overall sum of absolute coefficient values should be minimal. This "shrinks" coefficients of variables with little explanatory power to zero, which effectively provides a feature selection. This technique is known as *Lasso shrinkage regression* and described in detail by Hastie, Tibshirani, and Friedman (2009). Hastie et al. in their textbook also describe *ridge regression*, which shrinks the coefficients of correlated predictor variables toward each other. Ridge regression can be a useful alternative to feature aggregation if there are high intercorrelations among predictor variables but the aim is to retain the original variables in the model rather than to aggregate them into factors or components.

In these circumstances one possible solution is to separate the feature selection step from the subsequent data modeling. An example of this strategy can be seen in Jakubowski, Finkel, Stewart, and Müllensiefen (2017) who use symbolic music features of pop tunes to model the propensity of tunes to become earworms for music listeners. Similar to van Balen et al. (2015), Jakubowski et al. (2017) make use of information from a large corpus of commercial Western pop songs to define so-called second-order features that reflect the commonness or conventionally of a feature value (first-order feature) with respect to the feature distribution in the corpus. Their initial set of variables is comprised of eighty-two features, and the aim of the data modeling is to classify tunes as either earworms or non-earworms based on a set of musical features. In a first step, Jakubowski et al. used a random forest (Breiman 2001) to predict the earworm status of each tune using all eighty-two features. *Random forests* (see more detailed explanation later) are able to cope with large sets of predictor variables as input and can produce variable importance scores for all

predictor variables, reflecting the contribution of each variable to the classification or regression accuracy of the model, including main and all interaction effects. Following the approach outlined by Strobl, Malley, and Tutz (2009), random forests can deliver unbiased estimates of variable importance even when predictor variables are substantially correlated. Jakubowski et al. (2017) selected three features with substantial variable importance score ("tempo" as well as two features measuring the conventionality of the melodic contour of the tune) as indicated by a random forest model. These three feature predictors where then used for explanatory modeling of the earworm status of the tunes. Here, Jakubowski et al. used different modeling techniques (a binary logistic regression and a classification tree model) that allow for straightforward understanding and interpretation of the contribution of each of the three feature variables in the model—something that is difficult to achieve with a complex random forest model. In this example the random forest only served the feature selection step. The variable importance scores from random forest models are particularly handy for this purpose, but other statistical techniques are better suited for the modeling and interpretation of the role that individual features might play for the perception of earworms.

Supervised Classification and Regression

Perhaps the most frequently used statistical techniques in music corpus studies are supervised classifications and regressions. The basic task consists in predicting an attribute (e.g., "musical genre," "composer," "peak chart position," "radio spins per month," "tempo," "performance indication") of musical objects by a set of predictor variables (e.g., features derived from pitch or rhythmic information, instrumentation). A supervised classification or regression model therefore aims to establish an association between the predictor variables (i.e., independent variables in statistical terms) and the target attribute (i.e., the dependent variable). If the target attribute is measured on a numerical scale (e.g., tempo or radio spins per month) then the corresponding model is a regression model. If the target attribute is categorical and its values are different labels without any numerical interpretation (e.g., genre or composer) then this constitutes a supervised classification model. Genre classification (e.g., Tzanetakis and Cook 2002) and author attribution (e.g., Jürgensen and Knopke 2004) are prominent problems tackled by computational musicologists working with large music corpora. The underlying assumption is that the target attribute is a useful shortcut description for whole set of features that commonly appear together. A supervised classification model can then be used to predict, for example, the author of a newly discovered musical piece. Similarly, a regression model can be employed to predict from suitable musical features the chart success of a song just released. In addition, supervised classification and regression models can also have an explanatory value by illuminating the process of how labels or scale value are assigned to musical pieces. For supervised classification and regression, the values of the target attribute are known at the modeling stage, and they are necessary to learn how the target attribute is structurally associated with the predictor variables (i.e., the learning or optimization process is "supervised"). In contrast, in unsupervised classification models (see later discussion) there is no distinction between target attributes and predictor variables. Here, the statistical model has to learn autonomously how variables are associated with each other and discover structure and groupings (i.e., "clusters") in the data, often based on measures of similarity or distance between observations.

There are many statistical techniques for supervised classification and regression that are suitable for musical corpus studies, and, although it is beyond the scope of this chapter to go into sufficient detail, there are excellent introductory examples. For example, both James, Witten, Hastie, and Tibshirani (2013) and Witten, Eibe, Hall, and Pal (2017) provide excellent introductory and advanced treatments of many supervised classification and regression techniques, and Weihs, Ligges, Mörchen, and Müllensiefen (2007) summarize their applications in music research. Therefore, in the following discussion we restrict ourselves

to the presentation of a single example specifically designed to illustrate a supervised classification model with the aim to predict the corpus that a melody belongs to purely from the melody's musical features.

For this example, 250 melodies were sampled from different music corpora. First, fifty melodies were taken from each of three subcorpora (German children's songs, folk songs from the Polish region of Warmia, and folk songs from Ireland) of the Essen Folk Song Collection (Schaffrath 1995). Additionally, fifty jazz melodies were sampled from the Weimar Jazz Database (Pfleiderer 2017), and fifty pop melodies were taken from the M^4S corpus of Western commercial pop songs (Müllensiefen et al. 2008). For each of these 250 songs a set of 138 melodic features was extracted with the help of the *melfeature* module of the MeloSpyGUI (Frieler 2017). Using these features, the supervised classification models were constructed with the aim of mapping each melody to its (sub-)corpus. As statistical techniques, we employed classification trees and their corresponding ensemble method, *random forests* (Breiman 2001).

Generally speaking, classification (or decision) tree models work by recursively partitioning datasets into homogeneous subgroups where the target variable has mostly the same value (e.g., a corpus label as in this example). The data are split according to threshold values of the predictor variables. At each node of the tree, the predictor variable that maximally increases the homogeneity of the subgroups is selected for splitting. This process of partitioning the data into subgroups is repeated recursively until subgroups cannot be split any further either because a minimum group size criterion is reached or because there is no increase in homogeneity to be gained from a further split.

Tree models possess a number of characteristics that make them well-suited for the analysis of this dataset. Tree models use a built-in variable selection mechanism (see earlier discussion) and can easily cope with large sets of predictor variables. They do not assume a linear relationship between predictors and the target attribute (dependent variable). Finally, tree models are ideal for identifying higher-order interaction effects and also lend themselves very naturally to a graphical interpretation and understanding of the data (cf. Figure 8.3). However, despite these advantages, tree model can lack predictive power and generalizability of their results in comparison with other classification techniques. To overcome these limitations, Breiman (2001) suggested random forests as an extension of tree-based statistical models. In a random forest model, many trees are grown independently to predict the target attribute or classification label. For each tree only a bootstrap sample (i.e., a number of randomly selected data points) of the available data is used, and the number of explanatory variables is limited to random subset of all available predictor variables. From the many tree models grown within a random forest, the label that is predicted by the majority of tree models serves as the overall predicted classification label. The "importance" of a variable is then measured as the number of times it was selected for a node in all trees.

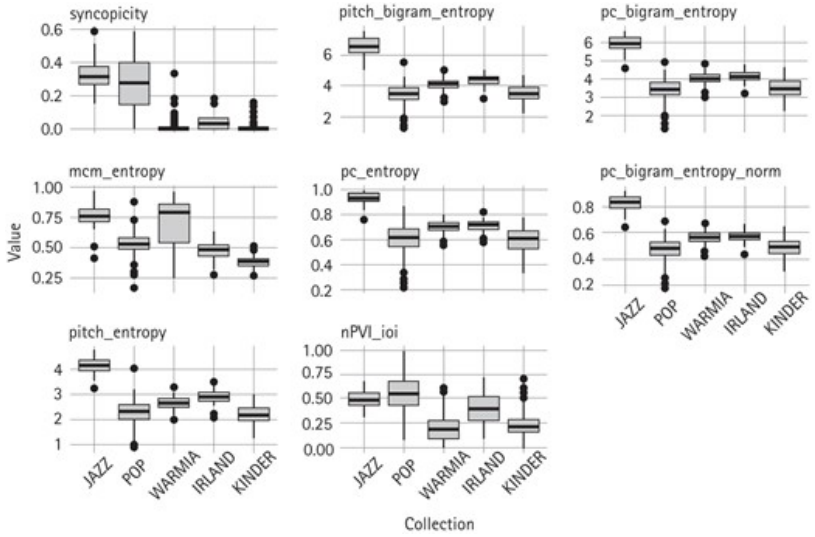
Random forests have been shown to have a superior prediction accuracy compared to individual tree models as well as compared to many other statistical predictive techniques (Fernandez-Delgado et al. 2014). Random forest models cannot be easily visualized but importance values for all predictor variables can be used to determine a subset of best predictors (see earlier discussion). For the current example, the eight feature predictors with the highest importance values according to the random forest are displayed in Figure 8.2.

The eight selected predictors are all related in some way to complexity. *Syncopicity* is the amount of syncopation in a melody; *pitch_bigram_entropy* is the entropy of consecutive pitch pairs, which can be interpreted as the predictability of the next pitch after knowing the preceding pitch value; *pc_bigram_entropy* is the entropy of consecutive pitch class pairs, which can be interpreted as the predictability of the next pitch class ("chroma") after knowing the preceding pitch class value; *mcm_entropy* is the entropy of distribution of metrical positions, which measures the uniformity in how metrical positions are occupied (a structured variable rhythm has lower *mcm_entropy* than a totally isochronous rhythm); *pc_entropy* is the entropy of pitch class distribution, which is higher for more

chromatic melodies, lower for diatonic, and even lower for pentatonic melodies; *pc_bigram_entropy_norm* is the normalized entropy of pitch class pairs, like *pc_bigram_entropy* but normalized to values between 0 and 1; *pitch_entropy* is the entropy of pitches, which measures the variability of pitch usage in a melody; and *nPVI_ioi* is the pairwise variability index of inter-onset intervals, which is a measure for the uniformity of rhythms (the lower the value, the more uniform the rhythm).

The boxplots show that the five different melody corpora can have very different values on some of these features. For instance, jazz and pop melodies have a much higher number of syncopations than the three subcorpora form the Essen collection. The jazz melodies clearly differ from the other four corpora in terms of *pitch_bigram_entropy* and other pitch-based features. Children's songs (Kinder), for example, have lower *mcm_entropy* values than melodies from all other corpora. Making use of the information of all features, the random forest model achieves an average classification accuracy of 88%, which is reasonably high considering the chance level of 20% on this five-class classification task.

Figure 8.2

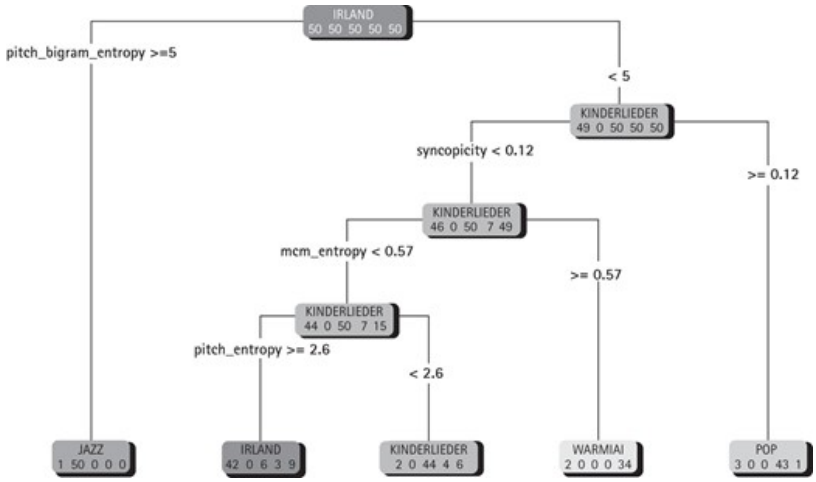


Boxplots of the most important variables according to a trained random forest on the example melody classification problem. *Syncopicity*: Amount of syncopation in a melody. *Pitch_bigram_entropy*: Entropy of consecutive pitch pairs. *Pc_bigram_entropy*: Entropy of consecutive pitch class pairs. *Mcm_entropy*: entropy of distribution of metrical positions. *Pc_entropy*: Entropy of pitch class distribution. *Pc_bigram_entropy_norm*: Normalized entropy of pitch class pairs. *Pitch_entropy*: entropy of pitches. *nPVI_ioi*: Pairwise variability index of inter-onset intervals.

Out of these eight features, the subsequent classification tree model only made use of four features and achieved a classification accuracy of 85.2%. Hence the classification tree performs the classification task less accurately compared with the random forest, but, in turn, the tree model consists of only four very simple rules that can be directly translated to musicological knowledge. The leftmost branch of its graphical representation in Figure 8.3 shows that a single feature, *pitch_bigram_entropy*, is sufficient to identify all jazz solos (with only one Irish folk song erroneously misclassified into this group). *Pitch_bigram_entropy* measures the lack of predictability of the second pitch in two-pitch sequences, which is much larger (i.e., *pitch_bigram_entropy* ≥5) for jazz solos than for the melodies from the other corpora (*pitch_bigram_entropy* <5). Following the right branch from the top node, the next feature is *syncopicity*, which helps to classify nearly all pop songs (forty-three out of fifty, with four misclassifications). Syncopations are relatively rare in European folk songs, whereas they are much more frequent in commercial pop music. The most important feature differentiation for the remaining of folk songs is *mcm_entropy*, which measures the predictability of metrical positions. Here, high values are a feature of a large portion (34/50) of the Warmian (East Polish) folk songs. This corresponds to a large number of songs with an odd meter in the Warmia

corpus. The final branch uses the *pitch_entropy* feature, with German children songs being much more predictive in pitch content (lower entropy) than folk songs from Ireland.

Figure 8.3



Decision tree using the most important variables identified by a random forest model for the melody classification example.

This example demonstrates the principle of supervised classification using random forests and classification tree algorithms. One has to bear in mind that a classification tree does not make use of all existing differences between melodies from different corpora but uses only of the most discriminative ones. This focus on the most discriminative features produces simpler and more parsimonious models that can be visualized in a graphical structure. But the restriction to use only a reduced set of predictors also means a loss in classification accuracy. This tradeoff between model simplicity and predictive power applies to all supervised classification and regression techniques. Therefore, the choice of a particular statistical technique should be guided by the purpose of the statistical analysis and primary research question.

Clustering (Unsupervised Classification)

In contrast to supervised classification techniques, the purpose of unsupervised classification or *clustering*[2] is to group similar objects (e.g., music pieces) into the same cluster and assign corresponding labels that indicate cluster membership. Hence, the result of a clustering process is often the creation of a new categorical variable with cluster labels that can be interpreted as an index for the similarity between objects. Generally, objects from the same cluster are considered more similar than objects from different clusters. Therefore, clustering induces or retrieves structure in a dataset. Clustering can aim to discover "real" structures in the data (i.e., recreate variables that were part of the data generation process). For example, in musical corpus studies clustering can aim to cluster together songs from the same genre if the genre information itself was not available (e.g., Mörchen et al. 2006). When cluster analysis is used for objective structure discovery, the clustering model can, at least in theory, be compared to "ground truth" data that would indicate the true cluster membership. Alternatively, "constructive" clustering (Hennig 2015) can be used to induce a new grouping structure (e.g., by clustering commercial Western pop songs into the two discrete groups "hits" vs. "non-hits"; see Frieler et al. 2015) while acknowledging that the underlying construct *commercial success* is not discrete but a continuous variable that comprises numerical variables such as peak chart position and duration in the charts. In this case the clustering solution cannot be evaluated against any ground truth data, but is determined by the choices of the researcher regarding the clustering algorithm and the desired number of clusters.

Clustering has a long tradition in musical corpus studies and probably started with the desire to group tunes in large folk song collections into tune families. Béla Bartók's classification of Serbo-Croatian folk songs in the Parry collection is an early example of the systematic ordering of melodies in a corpus by structural features (Bartók and Lord 1951; regarding methodology, see Bartók 1976). Later Wolfram Steinbeck (1982) introduced formal cluster analysis methods to folk song classification using features of musical structure that were computed from a symbolic music representation. Currently, cluster analysis is still used in cross-cultural research to classify musical pieces into a (small) number of groups that reflect their similarity (e.g., the clustering of ethnographically recorded songs into dance, lullaby, healing, and love songs; Mehr et al. 2019).

There are many different approaches to cluster analysis, many of which are discussed in the textbooks by Kaufman and Rousseeuw (1990) and Everitt, Landau, Leese, and Stahl (2011). But, on a conceptual level, there are generally three steps in a cluster analysis that also apply to clustering in musical corpus studies: the computation of a dissimilarity matrix, the application of a clustering algorithm, and the interpretation and description of the cluster solution.

Dissimilarity matrix. Once features are extracted for a collection of music pieces, the dissimilarity between each possible pair of pieces can be computed using a wide range of distance (or similarity) measures. The choice of distance measure strongly depends on the type of features that are used to describe the pieces in the collection. Some distance measures (e.g., Euclidean distance) can only be applied to numerical features, while others (e.g., Gower's coefficient; Gower 1971) also work with mixed-type data (i.e., a set of features that includes categorical, ordinal, and metrical features). In any case, the dissimilarity value for a pair of pieces should summarize the dissimilarity across the features extracted from the pieces. However, dissimilarity measures are not necessarily always computed directly from feature variables that describe specific aspects of a music piece. Instead, similarity between musical objects might be obtained directly by comparing two musical sequences, as, for example, by techniques that compare symbol strings (Müllensiefen and Frieler 2004) or by compression-based measures (e.g., Pearce and Müllensiefen 2017).

Clustering algorithm. The aim of most clustering algorithms is to find a cluster solution where each cluster is fairly homogeneous (i.e., the distance between objects within the cluster is small) and all clusters are well separated from each other (i.e., the distance between objects of different clusters is large). But homogeneity and separation can both be achieved in various ways. Additionally, within-cluster homogeneity and between-separation can be different goals, and, depending on the priorities of the researcher, different weightings may lead to different clusterings. There is an abundance of different clustering methods available that optimize homogeneity and separation in very different ways. Additionally, cluster methods differ in how the number of clusters is chosen, how the relationship between clusters is defined (e.g., hierarchical vs. nonhierarchical partitioning), how outlier cases are dealt with, and what geometrical cluster shapes are permissible. Finally, there are different methods for assessing the stability and fit of cluster solutions with regards to the data. This long list of options and approaches makes it clear that there is no optimal clustering method for musical corpus analysis per se but that the choice for a particular algorithm, implementation, and evaluation method needs to be made very much with an understanding of the data and the research goals in mind.

Interpretation. The outputs of most clustering algorithms are simply alphanumerical labels (e.g., "A," "B," "C," etc.) that are assigned to the members of each cluster. These labels do not have any a priori meaning, and whether a specific cluster of music pieces is labeled "A" or "B" is often completely arbitrary. Hence, once a clustering solution has been obtained and evaluated, the individual clusters need to be described and interpreted, and labels can potentially be replaced with meaningful cluster names. Here, the distribution of features within each cluster can be very useful for identifying summary characteristics of each cluster and to distinguish different clusters from each other. There is no principled technical or statistical solution to the interpretation of a cluster solution, but it is the responsibility of the researcher to find a convincing

argument for interpreting clusters on the basis of the empirical evidence at hand (e.g., see labeling of clusters of harmonic sequences in popular music assigned by Shaffer et al. 2019).

A good example of the use of cluster analysis in musical corpus studies is the paper on the cultural and geographical spread of song types in traditional Taiwanese music by Savage and Brown (2014). Their primary research question concerns the prevalence of distinct song families among twelve different indigenous peoples in Taiwan. The authors used a corpus of 259 traditional Taiwanese songs collected from across the entire island. Each song was characterized by twenty-six categorical and ordinal features that reflected different aspects of musical structure and are modeled after Alan Lomax's Cantrometrics classification system (Lomax 1976; Savage et al. 2012). From the matrix of song features (259 songs and 26 features), Savage and Brown computed the dissimilarity matrix of all possible song pairs (33,411 different dissimilarity values) using a distance measure (Rzeszutek et al. 2012) very similar to Gower's coefficient of similarity (Gower 1971). The dissimilarity matrix was then taken as input to the *k*-means clustering algorithm, which partitioned the 259 songs into *k* different clusters. In *k*-means clustering the clusters are disjoint (i.e., each song belongs to only a single cluster) and do not possess any hierarchical structure (i.e., no cluster is part of another cluster). The researcher needs to decide in advance into how many different clusters the collection of songs should be partitioned. Because there are no strong suggestions from prior research or ethnomusicological theory for the number of clusters, Savage and Brown ran the *k*-means algorithm several times with values for *k* ranging from 1 to 25. For each cluster solution, they compute the within-group sum of squares as a criterion of cluster (in)homogeneity and then use the magnitude of decrease on this criterion for deciding the optimal number of clusters. On their data, the decrease of cluster inhomogeneity was less strong after five clusters, and therefore they chose the five-cluster solution. In a final step, they interpreted the five clusters in terms of the most important aspects of musical structure (e.g., metrical regularity, rhythmic variability, most prevalent scale and contour types, typical phrase length). Savage and Brown do not claim that the five clusters correspond to any objective reality, but they arrived at their clustering solution via observing a model–data fit index. Hence, their cluster analysis induces structure in a constructive way.

The song clusters are subsequently used in the main analysis of the study where the authors showed that there are large differences in the prevalence of these song clusters in different geographical regions and among the twelve indigenous Taiwanese peoples. Consequently, they derived cluster names from the geographical region where each song cluster is most prevalent.

The use of cluster analysis in their study helps to distinguish between geographical region, cultural group, and type of song, where the type of song is defined purely by musical features. Only this constructive definition of song type by cluster analysis enables the authors to draw a differentiated picture of the musical overlap and the musical similarities (as well as differences) across different indigenous cultural groups. It is a good example of how cluster analysis can induce in a musical corpus a useful summary structure that is not visible from the distributions of a large feature set.

Conclusion

This chapter is intended as an inspiration for exploring the wide range of statistical methods that might be suitable for music corpus analysis. We encourage making the most of the musical corpus and the corresponding feature data, which usually take long to compile and are often tricky to extract. Musical corpus studies are usually very data-rich and offer many different perspectives of scientific inquiry that can be investigated by applying statistical methods.

The descriptions of individual methods presented in this chapter do not cover all aspects necessary for their application to a real dataset. But we hope that each section conveyed the central message that the choice for

or against individual methods and analysis options needs to be made with a clear research question in mind and a profound knowledge of the data available in the corpus. Across the individual sections, we have provided possible answers to questions that frequently appear in musical corpus studies: "Does my data show the expected trends?" and "Is there any interesting structure in the data that I have not been aware of?" "What are the most important features in my dataset?" and "How can I combine several features that seem to measure the same thing?" "Can I predict interesting attributes of the musical pieces in my corpus from musical features?" "How can I group together pieces that are musically similar?" Obviously, there are many more questions that can arise as part of a musical corpus study, but exploration, prediction, and the construction of a classification structure are tasks that are relevant for investigation and comparison of musical corpora, and answering these questions with the appropriate statistical methodology is a core part of research in this field.

References

Baayen, R. H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.
Google Scholar Google Preview WorldCat COPAC

Bartók, B. (1976). Why and how do we collect folk music? In F. Faber (Ed.), *Bartók, Béla Essays* (pp. 9–24). London: Benjamin Suchoff.
Google Scholar Google Preview WorldCat COPAC

Bartók, B., and Lord, A. B. (1951). *Serbo-Croatian folk songs: Texts and transcriptions of seventy-five folk songs from the Milman Parry Collection and a morphology of Serbo-Croatian folk melodies*. New York: Columbia University Press.
Google Scholar Google Preview WorldCat COPAC

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.
Google Scholar WorldCat

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2017). *shiny: Web Application Framework for R*. Retrieved from https://CRAN.R-project.org/package=shiny.
Google Scholar Google Preview WorldCat COPAC

Eerola, T., and Toiviainen, P. (2004). MIR in Matlab. The MIDI Toolbox. In *Proceedings of 5th International Conference on Music Information Retrieval (ISMIR 2004)* (pp. 22–27). Barcelona: Universitat Pompeu Fabra.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster analysis: Everitt/cluster analysis*. Chichester, UK: Wiley. https://doi.org/10.1002/9780470977811.
Google Scholar Google Preview WorldCat COPAC

Fernandez-Delgado, M., Cernadas, E., and Barro, S. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181.
Google Scholar WorldCat

Field, A., Miles, J., and Field, Z. (2012). *Discovering statistics using R*. London: Sage.
Google Scholar Google Preview WorldCat COPAC

Frieler, K. (2017). Computational melody analysis. In M. Pfleiderer, K. Frieler, J. Abeßer, W.-G. Zaddach, and B. Burkhart (Eds.), *Inside the Jazzomat: New perspectives for jazz research* (pp. 41–84). Mainz: Schott-Campus. Retrieved from http://schott-campus.com/jazzomat/.
Google Scholar Google Preview WorldCat COPAC

Frieler, K. (2018). A feature history of jazz solo improvisation. In W. Knauer (Ed.), *Jazz @ 100. An alternative to a story of heroes* (vol. 15). Hofheim am Taunus: Wolke Verlag.
Google Scholar Google Preview WorldCat COPAC

Frieler, K., Jakubowski, K., and Müllensiefen, D. (2015). Is it the song and not the singer? Hit song prediction using structural features of melodies. In W. Auhagen, C. Bullerjahn, and R. von Georgi (Eds.), *Jahrbuch Musikpsychologie* (vol. 25). Göttingen: Hogrefe-Verlag.
Google Scholar Google Preview WorldCat COPAC

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857. https://doi.org/10.2307/2528823.
Google Scholar WorldCat

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*

(2nd ed.). New York: Springer.
Google Scholar Google Preview WorldCat COPAC

Hennig, C. (2015). Clustering strategy and method selection. In C. Hennig, M. Meila, F. Murtagh, and R. Rocci (Eds.), *Handbook of cluster analysis*, (pp. 703–732). New York, NY: Chapman and Hall/CRC. Retrieved from http://arxiv.org/abs/1503.02059.
Google Scholar Google Preview WorldCat COPAC

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447.
Google Scholar WorldCat

Huron, David. (1988). Error categories, detection, and reduction in a musical database. *Computers and the Humanities*, *22*(4), 253–264.
Google Scholar WorldCat

Huron, David. (1996). The melodic arch in Western folksongs. *Computing in Musicology*, *10*, 3–23.
Google Scholar WorldCat

Jakubowski, K., Finkel, S., Stewart, L., and Müllensiefen, D. (2017). Dissecting an earworm: Melodic features and song popularity predict involuntary musical imagery. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(2), 122–135. https://doi.org/10.1037/aca0000090.
Google Scholar WorldCat

James, G., Witten, D., Hastie, T., and Tibshirani, R. (Eds.). (2013). *An introduction to statistical learning: With applications in R*. New York: Springer.
Google Scholar Google Preview WorldCat COPAC

Jürgensen, F., and Knopke, I. (2004). A comparison of automated methods for the analysis of style in fifteenth-century song intabulations. *CIM04 - Conference on Interdisciplinary Musicology*. Graz, Austria.

Kaufman, L., and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.
Google Scholar Google Preview WorldCat COPAC

Lomax, A. (1976). *Cantometrics: An approach to the anthropology of music*. Berkeley: University of California Extension Media Center.
Google Scholar Google Preview WorldCat COPAC

McKay, C., Cumming, J., and Fujinaga, I. (2017). Characterizing composers using jSymbolic2 features. *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*. Suzhou, China.

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., and Glowacki, L. (2019). Universality and diversity in human song. *Science*, 366(6468), 1–17. DOI: 10.1126/science.aax0868
Google Scholar WorldCat

Mörchen, F., Mierswa, I., and Ultsch, A. (2006). Understandable models of music collections based on exhaustive feature generation with temporal statistics. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)* (p. 882). Philadelphia, PA: ACM Press. https://doi.org/10.1145/1150402.1150523.

Müllensiefen, D. (2009). *FANTASTIC: Feature ANalysis Technology Accessing STatistics (In a Corpus)* (Technical Report). London: Goldsmiths.
Google Scholar Google Preview WorldCat COPAC

Müllensiefen, D., and Frieler, K. (2004). Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. *Computing in Musicology*, *13*, 147–176.

Google Scholar WorldCat

Müllensiefen, D., Wiggins, G., and Lewis, D. (2008). High-level feature descriptors and corpus-based musicology: Techniques for modelling music cognition. In A. Schneider (Ed.), *Hamburger Jahrbuch für Musikwissenschaft: Vol. 25. Systematic and comparative musicology* (pp. 133–156). Frankfurt: P. Lang.
Google Scholar Google Preview WorldCat COPAC

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021.
Google Scholar WorldCat

Pearce, M., and Müllensiefen, D. (2017). Compression-based modelling of musical similarity perception. *Journal of New Music Research*, *46*(2), 135–155. https://doi.org/10.1080/09298215.2017.1305419.
Google Scholar WorldCat

Pfleiderer, M. (2017). The Weimar jazz database. In M. Pfleiderer, K. Frieler, J. Abeßer, W.-G. Zaddach, and B. Burkhart (Eds.), *Inside the Jazzomat: New perspectives for jazz research* (pp. 19–41). Mainz: Schott-Campus. Retrieved from http://schott-campus.com/jazzomat/.
Google Scholar Google Preview WorldCat COPAC

Pfleiderer, M., Frieler, K., Abeßer, J., Zaddach, W.-G., and Burkhart, B. (Eds.). (2017). *Inside the Jazzomat: New perspectives for jazz research*. Mainz: Schott-Campus.
Google Scholar Google Preview WorldCat COPAC

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111. https://doi.org/10.2307/271063.
Google Scholar WorldCat

R Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from http://www.R-project.org.
Google Scholar Google Preview WorldCat COPAC

Revelle, W. (2018). *An introduction to psychometric theory with applications in R*. Retrieved from http://personality-project.org/r/boo.k.
Google Scholar Google Preview WorldCat COPAC

RStudio Team. (2016). *RStudio: Integrated development environment for R*. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/.
Google Scholar Google Preview WorldCat COPAC

Rzeszutek, T., Savage, P. E., and Brown, S. (2012). The structure of cross-cultural musical diversity. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1733), 1606–1612. https://doi.org/10.1098/rspb.2011.1750.
Google Scholar WorldCat

Savage, P. E., and Brown, S. (2014). Mapping music: Cluster analysis of song-type frequencies within and between cultures. *Ethnomusicology*, *58*(1), 133. https://doi.org/10.5406/ethnomusicology.58.1.0133.
Google Scholar WorldCat

Savage, P. E., Merrit, E., Rzeszutek, T., and Brown, S. (2012). CantoCore: A new cross-cultural song classification scheme. *Analytical Approaches to World Music*, *21*(2), 87–137.
Google Scholar WorldCat

Schaffrath, H. (1995). The Essen folksong collection. In D. Huron (Ed.), *Database containing 6,255 folksong transcriptions in the Kern format and a 34-page research guide*. Menlo Park, CA: CCARH.
Google Scholar Google Preview WorldCat COPAC

Shaffer, K., Vasiente, E., Jacquez, B., Davis, A., Escalante, D., Hicks, C., McCann, J., Noufi, C., Salminen, P. (2019). A cluster analysis of harmony in the McGill Billboard dataset. *Empirical Musicology Review*, *14*, 146–162. https://doi.org/10.18061/emr.v14i3-4.5576.
Google Scholar WorldCat

Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., and Archibald, A. B. (2000). Scientific graphs and the hierarchy of the sciences: A Latourian survey of inscription practices. *Social Studies of Science*, *30*(1), 73–94. https://doi.org/10.1177/030631200030001003.
Google Scholar WorldCat

Steinbeck, W. (1982). *Struktur und Ähnlichkeit: Methoden automatisierter Melodieanalyse* (vol. XXV). Kassel: Bärenreiter.
Google Scholar Google Preview WorldCat COPAC

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973.
Google Scholar WorldCat

Tabachnick, B. G., and Fidell, L. S. (2014). *Using multivariate statistics* (Pearson new international edition, 6th ed.). Harlow: Pearson.
Google Scholar Google Preview WorldCat COPAC

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
Google Scholar Google Preview WorldCat COPAC

Tzanetakis, G., and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* *10*(5), 293–302. https://doi.org/10.1109/TSA.2002.800560.
Google Scholar WorldCat

Van Balen, J., Burgoyne, J. A., Bountouridis, D., Müllensiefen, D., and Veltkamp, R. C. (2015). Corpus analysis tools for computational Hook Discove. *Proceedings of the, 16th International Society for Music Information Retrieval Conference*. Malaga, Spain.

Weihs, C., Ligges, U., Mörchen, F., and Müllensiefen, D. (2007). Classification in music research. *Advances in Data Analysis and Classification*, *1*(3), 255–291. https://doi.org/10.1007/s11634-007-0016-x.
Google Scholar WorldCat

Witten, I. H., Eibe, F., Hall, M., and Pal, C. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Amsterdam: Elsevier.
Google Scholar Google Preview WorldCat COPAC

Notes

1 https://jazzomat.hfm-weimar.de/feature_history_jazz/.

2 We use the terms “unsupervised classification” and “clustering” synonymously in this section.