Visualizing music similarity: clustering and mapping 500 classical music composers

Patrick Georges¹ · Ngoc Nguyen²

Received: 29 September 2018 / Published online: 11 July 2019
© The Author(s) 2019

Abstract

This paper applies clustering techniques and multi-dimensional scaling (MDS) analysis to a 500×500 composers' similarity/distance matrix. The objective is to visualize or translate the similarity matrix into dendrograms and maps of classical (European art) music composers. We construct dendrograms and maps for the Baroque, Classical, and Romantic periods, and a map that represents seven centuries of European art music in one single graph. Finally, we also use linear and non-linear canonical correlation analyses to identify variables underlying the dimensions generated by the MDS methodology.

Keywords Mapping classical music composers · Similarity measures · Dendrograms · Hierarchical clustering · Multidimensional scaling · Canonical correlation · Music information retrieval

Introduction

This paper uses several techniques that permit to visualize and graph some aspects of a Classical composers' similarity matrix computed by Georges (2017). The number of composers analysed is 500 and the matrix is of dimension 500×500 , leading to 250,000 pairwise (bilateral) composers' similarity indices. Sheer dimension prevents easy reporting of the results in a standard article. Hence, 'visualizing' or 'translating' the similarity matrix into clusters and mapping of composers is an important communication tool. Before embarking on this mapping project, it may be useful to provide some background on the motivation for building a composers' similarity matrix.

✉ Patrick Georges
pgeorges@uottawa.ca
Ngoc Nguyen
ngoc.nguyen@wku.edu

¹ Graduate School of Public and International Affairs, University of Ottawa, Social Sciences Building, Room 6011, 120 University, Ottawa, ON K1N 6N5, Canada

² Department of Mathematics, Western Kentucky University, 1906 College Heights Blvd., Bowling Green, KY 42101, USA

Through digital music files and streaming, classical music consumers have access to an enormous range of composers and styles of classical music. However, the vast store of music makes the problem of what, exactly, to listen to, even more acute than during the pre-digital age. Many introductions to the classical music world are in the business of inculcation through lists of ‘mandatory’ composers and compositions to explore. Yet, as mentioned by Smith (2000), most people explore new subjects by starting with the familiar, and in the case of music, this may mean hearing a composer that one likes and searching for more music of the same type. Pandora, Spotify, Last.fm, YouTube and other music streaming platforms have algorithms proposing what an auditor may want to listen next. These algorithms and their improvement are largely tributary to the field of music information retrieval (MIR), which develops innovative content-, context- and user-based searching schemes, music recommendation systems, and novel interfaces to make the vast store of music available to all.¹

Content-based MIR aims at uncovering from an audio signal meaningful music qualities (rhythm, timbre, melody, harmony, loudness, etc.) that can be used for music similarity and retrieval tasks. The unit of study is typically the ‘composition’ and the analysis is to compare similarities/differences across audio signals from a series of compositions or from different segments of a same composition (self-similarity). This particular angle has been used among others, by Foote (1999), Pampalk (2006), and more recently by Weiss (2017) and Weiss et al. (2018) for classical music, and by Mauch et al. (2015) for popular music where they investigate the evolution of musical diversity and disparity and whether evolution has been gradual or punctuated. Context-based MIR is motivated by the fact that there are aspects not encoded in an audio signal or that cannot be extracted from it, but which are nevertheless important to human perception of music, for example, the cultural background of a composer.² In this case, the unit of analysis may also be the ‘composer’ or the ‘artist/performer’, not just the ‘composition’. Examples are Pachet et al. (2001) and Baccigalupo et al. (2008). In this vein, Smith and Georges (2014) propose a statistical analysis that captures similarity across pairs of composers by mean of pairwise comparison of presence-absence of personal musical influences (e.g., other composers/influencers). Data on personal musical influences are taken (and available) from ‘The Classical Music Navigator’ (Smith 2000; hereafter referred to as *CMN*) where each of the 500 composers of the database is associated with a list of composers who have had a documented positive influence on a subject composer.³ In essence, they infer similarities among composers by assuming that if two composers share many of the same personal musical influences, their music will likely have some similarities. On the other hand, if two composers have been influenced by

¹ One of the earlier survey article on MIR is Orio (2006). Schedl et al. (2014) provide a survey of more recent developments and applications.

² Besides content- and context-based MIR, the user also plays a key role in many MIR applications, including for recommendations, retrieval, and similarity analyses, which may also depend on user properties and user context. User context represents frequently changing factors such as activity, emotion and current social context of a user. User properties refer to constant or slowly changing characteristics of a user such as music taste and music education. See Schedl et al. (2014) for a survey of MIR explorations to model and analyse user behavior and to incorporate this knowledge into retrieval and recommendation systems.

³ See the ‘composers’ page of the *CMN*. For example, the set of composers who had a documented positive influence on Debussy is: {J.S. Bach, Wagner, Chopin, Tchaikovsky, Liszt, R. Schumann, Ravel, Fauré, Grieg, Rimsky-Korsakov, Mussorgsky, Franck, Gounod, Massenet, Satie, Borodin, Rameau, Albéniz, F. Couperin, Joplin, Delibes, Chausson, Lalo, Chabrier, Dukas, Alkan}.

very distinct sets of composers, then their music is likely to have little similarity.⁴ They use and compare the performance of a dozen similarity indices or measure of association commonly used in fields such as biogeography and biological systematics such as the first and second Kulczynski coefficients (1927), the Jaccard coefficient (1901), the Dice coefficient (1945), the Simpson coefficient (1943), the Smith coefficient (1983), the binary distance coefficient (Sneath 1968), and the binomial index of dispersion χ^2 statistic (Potthoff and Whittinghill 1966).

Smith and Georges (2015) examine how much, if any, additional explanation can be added to their first study when ‘ecological’ measures (i.e., other characteristics such as time period, geographical location, school association, instrumentation emphases, etc.) are also taken into account. Each composer is associated with a subset of 298 ecological categories (also extracted from and available at the *CMN*),⁵ and the authors infer a statistical association between pairs of composers by assuming, in essence, that if two composers share many ecological categories, then their musical ‘ecological niches’ are very similar, so that, in this sense, they may be considered similar.⁶ Although the two methods (personal musical influences and ecological characteristic) do not generate identical similarities rankings, their combination does provide, arguably, an improved system of ranking. Finally, Georges (2017) applies on the same database a new measure of similarity, equipped with a statistical significance test, the ‘centralised’ cosine measure. The centralised cosine is based on

⁴ For any pair of composers (i, j) for $i, j \in C$ (among the $n \times n$ possible pairs with $n=500$), we have a set I_i of all personal musical influences on composer i , and a set I_j of all personal musical influences on composer j . We are interested in capturing whether a composer $k \in C$ had a reported influence on both i and j , on i but not j , on j but not i , and on neither i nor j . Thus, for any pair (i, j) , $I_i \cap I_j = CI_{ij}$ is the set of composers k that have influenced both i and j ; $I_i - I_i \cap I_j = I_{i,-j}$ is the set of composers k that have influenced i but not j ; $I_j - I_i \cap I_j = I_{j,-i}$ is the set of composers k that have influenced j but not i and $DI_{ij} = I_{i,-j} \cup I_{j,-i}$ is the set of composers k that have influenced either i or j but not both. From this we can produce a count table for any pair (i, j) that sums the elements (the number of composers) in each of the four sets CI_{ij} , $I_{i,-j}$, $I_{j,-i}$, and $C - CI_{ij} - DI_{ij}$, and from which similarity indices for all pairs of composers (i, j) can be computed on the basis of well-known formulas. For example, the binomial index of dispersion used in Smith and Georges (2014) can be computed for any pair (i, j) as: $BID_{ij} = n(ad-bc)^2 / [(a+b)(c+d)(a+c)(b+d)]$, where a , b , c , d , and n are the count/number of composers in each of the five sets CI_{ij} , $I_{i,-j}$, $I_{j,-i}$, $C - CI_{ij} - DI_{ij}$ and C .

⁵ See the ‘Index of Forms and Styles of music’ in the *CMN*. For example, the ecological characteristics and style influences associated with Debussy are represented by the following elements: {ballets 1900 on, cello chamber music, chamber music 1825 to 1925, ‘Dance’ in composition title, etudes, flute unaccompanied, flute chamber music, harp chamber music, harp orchestral music, Impressionist style, nocturnes, operas 1900 on, orchestra incidental music, orchestra symphonic poems, orchestration, Paris composers 1800 on, piano unaccompanied 1775 to 1900, piano unaccompanied 1900 on, piano chamber music general, quartets for strings, song cycles and collections, songs 1800 to 1900, songs 1900 on, suites, trios for other combinations, viola unaccompanied or chamber music, violin chamber music 1850 on, Asian music, gamelan music, Renaissance Period music}. An 8-page list of all ecological categories is available in Smith and Georges (2015).

⁶ We use the word ‘ecological’ to highlight the context-based nature of these categories. Categories such as ‘harp chamber music’, ‘piano composition from 1800 onwards’, ‘impressionist style’ and ‘gamelan music’, in order to describe Debussy, or ‘second Viennese school’, ‘serial technique’ and ‘expressionist style’ for Schoenberg, are fundamentally different from content-based information extracted from a score or audio file (e.g., harmonic intervals, tonal complexity, chord transitions, etc.). While there might be, in general, a fine line between ‘content’ and ‘context’-based information, in this paper content-based information refers to audio features (as mentioned just above) that can be extracted from one specific piece of music, while context-based information represents more general information about the style of music (including instrumentation emphasis) of a composer.

Table 1 A 5×5 subset of a 500×500 matrix of composers' similarities, S_{comb}

	1. Bach, JS	2. Mozart, WA	3. Beethoven	4. Schubert	5. Brahms
1. Bach, JS	1.0000	0.1121	0.0962	0.0547	0.0847
2. Mozart, WA	0.1121	1.0000	0.3947	0.2684	0.2003
3. Beethoven	0.0962	0.3947	1.0000	0.4590	0.3252
4. Schubert	0.0547	0.2684	0.4590	1.0000	0.2600
5. Brahms	0.0847	0.2003	0.3252	0.2600	1.0000

Source: Georges (2017)

earlier literature in scientometrics and bibliographic couplings.⁷ It provides a more sensitive measure of similarity than the binomial index used by Smith and Georges (2014, 2015) as it also tracks composers who (consciously or not) attempted to 'differentiate' themselves from others, leading to negative values for the index. Using the centralised cosine measure, three 500×500 composers' matrices of similarity indices are computed, the first one (S_{infl}), on the basis of personal musical influences, the second one (S_{ecol}), on the basis of 298 ecological characteristics and the third one (S_{comb}), that combines information from personal musical influences and ecological data. A subset of this last matrix for the Top-5 major composers of the database is given in Table 1. Eventually, Georges (2017) compares the first two matrices and shows that the interactions between similarity information extracted from personal musical influences and from ecological characteristics can provide a basis for an evolutionary model of Western classical music. In particular the analysis aims to understand why composers are perceived to 'sound' different (on the basis of ecological characteristics) despite similar 'lineages' (personal influences network) or why they could 'sound' alike even when their 'lineages' are different. Using an analogy to biodiversity terminology, these two cases are respectively referred to as 'music speciation and evolution' and 'convergent evolution'. The paper finally proposes a statistical test in order to identify 'transitional' composers, 'innovators' and 'followers' in the development of Western classical music.

Besides the application proposed in Georges (2017), more can be done with a composers' matrix of distance or similarity. In the rest of the introduction we present the contribution of this paper with respect to what exists in the literature. First, this paper applies clustering techniques to the similarity matrix S_{comb} in order to compute dendrograms or family tree diagrams for composers. Previous applications range from fields as diverse as biology (the tree of life evolution) to linguistics (grouping of languages according to linguistic affinities). Classifying (or clustering) composers into pre-defined groups such as

⁷ The ordinary (non-centralised) cosine similarity measure (also known as the Salton's measure) is a statistic familiar to bibliometrics and scientometrics. The idea was mathematically formalized by Sen and Gan (1983) and later extended by Glänzel and Czerwon (1996) who also applied the methodology. As discussed in Smith et al. (2015), when all the vectors are Boolean vectors, the null distribution of the cosine similarity under the assumption of independence between two composers is unknown and has a nonzero mean; in order to derive a statistical test for the cosine measure, a centralised cosine measure was proposed (Giller 2012). The centralised cosine measure is the cosine measure computed on the centralised vectors, with respect to the mean (average) vectors, and this is the measure used in Georges (2017) as applied to classical music composers. The formula is: $CSCI_j = (ad - bc) / \sqrt{[(a+b)(c+d)(a+c)(b+d)]}$, where a, b, c, d are the count of composers in sets described in Footnote 4.

Renaissance, Baroque, Classical, Romantic, Modern, is of course a customary method in musicology. Recent computational methods based on statistical analyses of content-based information such as symbolic scores representations, MIDI files, or audio recordings (e.g., Rodriguez Zivic et al. 2013; White 2013; Weiss 2017; Weiss et al. 2018) typically show that composers tend to cluster in ways that conform to our intuitions about stylistic traditions. In this paper we will compare the clustering performance of our context-based approach versus the content-based approach of Weiss (2017). Beyond grouping composers on a fixed number of clusters, we also investigate, as Weiss (2017), hierarchical clustering, which better highlights evolution and trends in classical music. Weiss (2017) does it for 70 composers and we will again compare results in the next section. Besides the 70 composers analysed in Weiss, our method permits to compute dendrograms for up to 500 composers representing seven centuries of classical music. We will also cluster composers within specific period (Baroque, Classical, and Romantic).

The second objective of the paper is to apply multi-dimensional scaling (MDS) to the similarity matrix S_{comb} so as to generate maps of classical music composers. We occasionally see such maps on the internet that can be used to discover composers who are ‘close’ or ‘similar’ to a subject composer. Yet, these maps come with their own problems. For example, a search on Beethoven suggesting both Schubert, Stravinsky, the Spice Girls and Miles Davis, as ‘related’ composers is one example of a methodology that requires improvement.⁸ This in itself is perhaps not the key problem; the real issue is that these user-oriented maps rarely focus on the methodology and the data behind their construction. As a mapping methodology, MDS has been used in many fields from marketing research to cognitive psychology. Recently, Vehkalahti and Everitt (2019) suggest that MDS can be applied to classical music and they illustrate this with a map of ten classical composers. This said, our MDS application maps up to 500 composers and we also map them by main periods (Baroque, Classical, Romantic, etc.). In terms of data on composers, we refer readers to the *CMN* website, and to Smith and Georges (2014, 2015) and Georges (2017) for a complete description of the data collection, the computing of similarity matrices, and the similarity index that is used (centralised cosine measure). In the first section (Clustering techniques and dendrograms) and the second section (Multidimensional scaling), we will be using the 500×500 composer similarity matrix S_{comb} , built on the basis of combined personal musical influences and ecological data. Indeed, as discussed in Smith and Georges (2015) and already mentioned above, adding both sets of data provides an arguably improved composers’ similarity ranking and similarity matrix relative to the cases when we use only the personal musical influences or the ecological data. This said, we will often apply the methods analysed here to a partition of the full 500×500 matrix (i.e., a matrix of smaller dimension) in order to present graphs that remain readable if printed. Finally, the next section extends the MDS analysis to three dimensions and uses linear and nonlinear canonical correlation analyses as an attempt to discriminate factors explaining the MDS dimensions. In that section we will be using the similarity matrix S_{infl} , based on personal musical influences only, in order to extract MDS dimensions (the composers coordinates). Then canonical correlation will be applied to these computed MDS coordinates and a few selected composers’ ecological characteristics to assess the discriminatory power of ecological categories on the MDS coordinates.

⁸ See music-map.com. Accessed February 2, 2019.

Clustering techniques and dendrograms

Musicologists have typically recounted the history of classical music by classifying composers into defined groups such as Renaissance, Baroque, Classical, Romantic, and Modern periods. See for example Taruskin (2010), and Taruskin and Gibbs (2013), henceforth T&G. Recently, clustering composers using statistical analyses of content-based information such as audio recordings in the case of Weiss (2017) and Weiss et al. (2018), have generated results that, on average, tend to conform to traditional musicology analyses. For example, Weiss (2017) selects commercial audio recordings of 1600 pieces (movements) for 70 composers and averages audio features (such as chord progressions, interval classes, and complexity features) of the recordings over each composer. This generates a composer feature matrix on which he performs principal component analysis (PCA) followed by K -means clustering on the first three principal components, with $K=5$ (i.e., the algorithm groups the 70 composers into exactly 5 clusters). In this section, we follow the same methodology on the same list of composers, applying PCA to our context-based similarity matrix S_{comb} (instead of an audio, or content-based, feature matrix) and then using K -means clustering on the first three principal components.⁹

Our results are given in Fig. 1 and can be compared with Figure 7.24 in Weiss (2017). As in Weiss, mostly, composers with a similar lifetime belong to a same cluster. The first cluster corresponds to the traditionally-defined Baroque period (e.g., Lully, Corelli, Vivaldi, J.S. Bach, Handel, D. Scarlatti). We then have a pre-classical cluster (e.g., J. Stamitz, J.C. Bach).¹⁰ This is followed by a classical cluster (e.g., J. Haydn, Mozart, Beethoven, Schubert), a Romantic cluster (e.g., Weber, Chopin, Bruckner, Mussorgsky, Mahler, R. Strauss), and finally a Modern cluster (e.g., Debussy, Ravel, Schoenberg, Bartók, Stravinsky, Messiaen, Boulez). There are very few surprises in Fig. 1 with respect to conventional grouping of composers, with perhaps the exception of classical composer Cimarosa clustered with earlier pre-classical composers.¹¹ It could also be argued that a few ‘transitional’ composers might be grouped into one instead of another cluster. For example C.P.E. Bach, a transitional composer between Baroque and Classical periods, often classified as Pre-classical

⁹ We actually do this for 65 of the 70 composers selected by Weiss because five of them (Giustini, Platti, Pleyel, Leopold Mozart, and Antheil) are not included in the list of 500 composers given in the *CMN* database.

¹⁰ The period of transition from Baroque to Classical is, according to T&G (2013), “a historiographical black hole that scholars tried to plug by searching for a ‘missing’ link between the Bach/Handel and Haydn/Mozart poles.” According to T&G (2013), “[no] period has been in greater need of fundamental research than that from the 1730s to the 1760s, what was long commonly known as Preclassic (and thus relegated by its very name to a status of relative insignificance) or labeled as Rococo, style galant, and *empfindsamer Stil* (sensitivity style).” T&G (2013) classify CPE Bach and JC Bach as, respectively, representatives of the ‘*empfindsamer Stil*’ and the ‘*style galant*’.

¹¹ Technically, one reason why Cimarosa (1749–1801) is pulled into this pre-classical group of composers is that some of his personal musical influences in our database are rather backward-looking (D. Scarlatti, Pergolesi, and Piccinni), and in terms of ecological characteristics, his name is associated with a relatively imprecise category of operas from 1600 to 1800. Interestingly, from a musicological point of view, Sheldon (1989) argues that “the essential problem of ‘Pre-Classic,’ is actually one of ‘Classic,’ an aesthetic normative concept based on the accomplishments of a few ‘great men’. In other words, the obvious drawback of ‘Viennese Classicism’ as a style period designation is that it excludes far more than it includes.[...] [C]ontemporaries (Germans as well as Italians) felt more appropriately represented musically by Antonio Sacchini, Giuseppe Sarti, Giovanni Paisiello, and Domenico Cimarosa.”

(see Footnote 10), is clustered in Fig. 1 with other Baroque composers, and Schubert, a transitional composer between Classical and Romantic periods, is clustered in the Classical period.¹²

Comparing our results with those of Weiss, his clustering does not lead to a Pre-classical cluster (even if such a period is typically mentioned by musicologists). Also he generates two parallel clusters covering the modern period (what he refers to the ‘avant-garde’ of that time (e.g., Schoenberg, Berg, Varèse, Boulez, and, more contentiously, Britten and Bartók) and a more moderate harmonic style (such as Prokofiev, Shostakovich, Hindemith, Milhaud). Observe that, as suggested by Griffiths (1978) in his ‘*Concise history of modern music: From Debussy to Boulez*’, our clustering of the modern period includes the impressionists, Debussy and Ravel, while they are both clustered with other Romantic composers in Weiss’ analysis. Also quite surprising in Weiss’ analysis, are the clustering of Vivaldi and, to a lesser extent, D. Scarlatti, with Classical composers, C.P.E Bach with Romantic composers, or Mussorgsky and Fauré in the (moderate harmonic) modern cluster.¹³ As Weiss (2017) argues, “[t]hese outliers point to the difficulties of clustering composers to a fixed number of top-level clusters”. This leads him to propose applying hierarchical clustering, as in bioinformatics where “phylogenetic trees are popular tools for clustering DNA sequences in order to highlight evolutionary development and trends.” In the following, we quickly describe hierarchical clustering and then apply it to the list of composers selected by Weiss. We finally apply hierarchical clustering to specific periods so as to highlight intra-period developments.

The agglomerative hierarchical clustering algorithm builds a cluster hierarchy displayed as a tree diagram called a dendrogram. In our case, the primary input for the algorithm is the 500×500 matrix of composers similarity S_{comb} . Typically, the algorithm applied to the composers’ similarity matrix begins with each composer in a separate cluster. In the very first step a two-composer cluster is formed between the two most similar composers. Then, at each successive step, the two clusters that are most similar are joined into a new cluster. Several methods are available to compute distance between clusters of composers

¹² There is a continuing question and problem as to how the music of Franz Schubert (1797–1828) is best to be understood, as representative of Classicism or of Romanticism. As argued by Gray (1978), “the Romantics created a Schubert in their own image with the aid of countless colorful legends which have little to do with fact. [...] The Romantics tended to dismiss Schubert’s early works, which are clearly Classical, with the important exception of his songs; they preferred to believe that the more personal late works simply burst forth in a paroxysm of Romantic inspiration.” Yet, Gray (1971) argues that Schubert’s works are largely reflective of the Viennese Classicism of Haydn, Mozart and Beethoven. Furthermore, he argues that “historians have usually drawn upon the form wherein Schubert is universally regarded as supreme—the Lied—as evidence that that he is, most of all, a romantic.” But this attitude may follow from a reverse historicism, conferring the accolade of ‘romantic’ upon Schubert’s songs because the art song is one of the most romantic forms in the hands of Schumann, Franz, Loewe, Brahms, and Wolf. Yet, according to Gray (1971), “that is exactly in his songs where Schubert most perfectly achieves classic stature.”

¹³ Domenico Scarlatti (1685–1757) is primarily a Baroque composer, clustered under Baroque in Fig. 1. Pre-classical (galant) music emerged during his long life, and there is some evidence that D. Scarlatti might have been the first great advocate of the Florentine piano (see Sutherland, 1995). Yet, it remains odd to see a clustering analysis such as the one in Weiss, which groups D. Scarlatti together with other Classical composers. It would be interesting to see how Weiss’ results are sensitive to his selection of recordings that use exclusively the piano, instead of the harpsichord. Carl Maria von Weber (1786–1826) is also clustered in Weiss together with other Classical composers, while he is clustered with Romantic composers in our Fig. 1. Weber’s grouping remains ambiguous as he was a transition composer from the Classical to the Romantic periods. Yet, according to Finscher (1983), his opera, *Freischütz* is often considered to be the first German Romantic opera (first performed on June 18, 1821), a genre which was to dominate the opera houses of the German-speaking countries until well beyond the middle of the [nineteenth] century, leading the way to the Romantic operas of Wagner.

(as opposed to the distance between pairs of composers, which is the primary input), such as single linkage, complete linkage, simple average, centroid, median, group average (unweighted pair-group), Ward's minimum variance, etc.¹⁴ Here we use the group average method whereby the distance between two clusters is defined as the average distance between each of their members. Let the distance between composers/clusters i and j be represented as $d_{i,j}$ and suppose that these are the two most similar composers in the database. Then i and j will be merged into a new cluster k from which a new set of distance must be computed between k and any other composers/clusters m in the database using the formula: $d_{k,m} = (n_i/n_k)d_{i,m} + (n_j/n_k)d_{j,m}$ where n_x is the number of composers in cluster x . Hence, just after the very first step, when the two most similar composers are fused into a cluster k , the distance between this newly formed cluster and any other composer m is $d_{k,m} = (1/2)d_{i,m} + (1/2)d_{j,m}$, that is, the average of the distance between i and m and j and m .¹⁵ This permits to progressively build the cophenetic distance matrix which can then graphically be represented with a dendrogram.¹⁶

Figures 2, 3, 4 and 5 show dendrograms resulting from different subsets of the 500×500 similarity matrix S_{comb} , as starting points for the algorithm.¹⁷ Figure 2 focuses on the list of composers in Weiss (2017). Figures 3, 4 and 5 give dendrograms for composers of the Baroque, Classical, and Romantic periods. Dendrograms for other periods (Medieval and Renaissance, and twentieth century music (Impressionism, atonal, dodecaphonic, neo-Classical, neo-Romantic and 'avant-garde' composers)) and for all 500 composers are available from the authors upon request, to be studied from a deeper musicological angle. In all these figures, each joining of two clusters is represented by a node (the splitting of a vertical line into two vertical lines) and the vertical height of the split (as measured on the vertical axis) shows the distance between the two clusters (the larger the value the closer the clusters or individual composers). One typically reads these dendrograms by 'cutting' the tree at a specific height (see for example the horizontal dash line as the cut-level) from which one can examine and study the resulting main clusters. Figure 2 shows three main clusters (Baroque and Classical in blue, Romantic in green and Modern in red) and several lower-level sub-clusters. Without going into much detail, we can see that the results at sub-clusters levels are consistent with common knowledge, for example, Baroque composers are grouped by nationality (French: Lully, Rameau, F. Couperin; German: Telemann, J.S. Bach, Handel; and Italian: Corelli, Vivaldi, Albinoni). There is a group of transition composers from Baroque to Classical (C.P.E Bach, J.C. Bach, J. Stamitz), a core of Classical composers (Boccherini, Mozart, J. Haydn, M. Haydn, Cimarosa, Salieri) and late Classical composers including those transitioning into early Romantic music (Clementi, Beethoven, Schubert, Weber, Rossini). Mendelssohn, an early Romantic composer is also in this

¹⁴ Chapter 445 of the NCSS Statistical Software offers an accessible introduction to hierarchical clustering.

¹⁵ Note that, once fused, composers are never separated.

¹⁶ In the process of constructing a dendrogram, a cophenetic matrix is computed. It is a distance matrix wherein the original pairwise distance between composers are replaced by the computed distance between their clusters at the time these clusters are merged. The cophenetic correlation coefficient provides the correlation between both matrices. The cophenetic correlation coefficients for Figs. 2, 3, 4 and 5 are above 0.75, which is typically good.

¹⁷ The dendrograms have been generated with XLSTAT (Addinsoft 2019).

Fig. 2 European art music. Dendrogram for the 65 composers in Fig. 1

Fig. 3 European art music. Dendrogram for the Baroque period (unweighted-pair group average)

Fig. 4 European art music. Dendrogram for the Classical period (unweighted-pair group average)

Fig. 5 European art music. Dendrogram for the Romantic period (unweighted-pair group average)

group, paired with Beethoven.¹⁸ Pairs of composers in other periods seem reasonable, such as Debussy and Ravel, Schoenberg and Berg, Prokofiev and Shostakovich, Fauré and Grieg, Dvorák and Brahms, R. Strauss and Mahler, Chopin and Liszt, Robert and Clara Schumann. Recall that the list of composers studied here was taken directly from Weiss (2017) whose Figure 7.25 also gives a hierarchical clustering.¹⁹ Yet, as the author pinpoints, some of his results do not reflect meaningful stylistic similarities, such as the pairing of Sibelius and C.P.E. Bach, Fauré and Prokofiev, J. Stamitz and Corelli. Comparing our results with those in Weiss (2017) shows that using context-based data, instead of content-based (audio files) data, remains a useful approach for clustering and visualising similarities across composers.

It is interesting to note that grouping well-known composers into either a fixed number of clusters or into hierarchical clusters (as in Figures 1 and 2 and in Weiss 2017) roughly coincides with common musicological knowledge. However, more useful perhaps, would be to cluster a larger number of composers belonging (born into) a specific period, so as to shed light on ‘finer’ intra-period connections between composers (including composers less well-known than the very familiar names of Figs. 1, 2).²⁰ Figure 3 shows the dendrogram for the Baroque period with six main clusters that seem related to a regional/national/country dimension.²¹ Sanz is a single-element cluster representing Spain (a country which has long been isolated from the rest of Europe in terms of musical influences). We then see a cluster of English baroque composers (from Gibbons, in fact, late Renaissance, to Purcell), and an ‘Italian’ cluster (from the early Baroque opera composer Cavalli, up to the high/late baroque violin virtuoso Locatelli, passing by Corelli, A. Scarlatti, Vivaldi and Handel who, although ‘German’ and working in England, was very much influenced by ‘Italian’ music).^{22,23} German composers are divided into two clusters, the early to middle Baroque cluster, centered on the German keyboard (harpsichord and organ) tradition (from the early German Baroque of Schein, Schütz and Scheidt, culminating to J.S. Bach, and passing by some important keyboard composers such as Froberger, Buxtehude, Bruhns, Pachelbel and Böhm), and a high/late-Baroque cluster largely based on the city of Dresden (Zelenka, Heinichen, Weiss, and Quantz (a student of Zelenka)). Interestingly, the late Renaissance/early Baroque Italian keyboard composer, Frescobaldi, is part of the former cluster, which reminds us the influence that he had on the development of German

¹⁸ As argued by Todd (2008), “[Mendelssohn] is the staunch upholder of conservative aesthetic value; no radical reformer or innovator, he is the composer of finely polished chamber works that fall easily and unobtrusively in the Classical paradigms (duo sonata, piano trio, string quartet, and quintet) established by Mozart and Haydn in the eighteenth century and redefined by Beethoven in the early decades of the nineteenth.”

¹⁹ To be fully consistent with Weiss (2017), Fig. 2 is built using complete-linkage hierarchical clustering. But the unweighted pair-group algorithm provides very similar results to Fig. 2. Figures 3, 4 and 5 are built using the unweighted pair-group algorithm described in the text.

²⁰ Here, composers have been assigned to a specific period on the basis of their birthdate. Baroque (birth-dates): 1583–1698, Classical: 1699–1795, and Romantic: 1796–1860.

²¹ In this and the following figures, the name of the composers is preceded by his/her date of birth.

²² Recall that neither Italy nor Germany were politically unified at that time. Yet these geographical concepts are useful, here, from a clustering point of view.

²³ Technically, our database indicates ‘Italian music’ and ‘Italian opera’ as style influences for Handel. According to T&G (2013) “[Handel] spent his true formative years—from 1706 to 1710—in Florence and Rome, where he worked for noble and ecclesiastical patrons and met Alessandro Scarlatti, Corelli, and other luminaries of the day. He was known affectionately as *il Sassone*, ‘the Saxon,’ meaning really ‘the Saxon turned Italian,’ in the musical sense.”

keyboard music).^{24,25} Finally the last cluster is itself composed of two main sub-clusters. First we have the French baroque tradition (from Lully and Charpentier, up to late Baroque Leclair, passing by Marais, F. Couperin, Rameau and Boismortier); and then we have a Bohemian-Austrian sub-cluster of composers centered on Vienna and Salzburg represented by Schmeltzer, Biber (who was, allegedly, Schmeltzer's student), Fux, and the rather cosmopolitan composer Muffat). Violin, instead of harpsichord, is also more predominant in this group (at least for Schmeltzer and Biber) than in the German 'keyboard' cluster mentioned above. That the cosmopolitan German composer Telemann belongs to this last cluster is interesting as he is often portrayed as having incorporated and mixed several national styles (German, French, Italian, and Polish traditions) in his compositions.²⁶

The dendrogram for the Classical period is given in Fig. 4. First, observe two clusters with a single-element, Krebs and Billings. Krebs was a German composer and a student of J.S. Bach, and whose contrapuntal late baroque style was being supplanted by the newer *galant* style that eventually led to the Classical music period. Hence he does not belong to the Classical period and the cluster analysis neatly points out this fact. Billings was a Boston-based American choral music composer, which also isolates him from the European classical music tradition. The other three main clusters are the 'early' pre-Classical cluster (from Arne to C.P.E. Bach, passing by Martini and the opera composers Pergolesi and Galuppi), 'core' Classical (from Dittersdorf to Beethoven, passing by Grétry, Mozart, Cimarosa and Salieri) and 'late' Classical (from Clementi to Rossini and Mercadante, passing by Hummel, Field, Czerny, and Paganini). Whether Beethoven should be viewed as 'late' instead of 'core' Classical composer is an open question. The cluster analysis puts him close to J. Haydn, one of his teacher who epitomises core Classical music. Yet, Beethoven's mid-period compositions also point toward the emergence of Romantic music which would tend to position him into the late Classical cluster (transitioning into Romantic period).²⁷ Finally Fig. 5 shows the six clusters for the Romantic period.²⁸ First we have the operetta and waltz composers (e.g., Offenbach and J. Strauss Jr.). Then, the violin virtuosos (Sarasate, Vieuxtemps, Wieniawski), and a cluster of composers who predominantly wrote operas (from Bottesini to Wagner, passing by Bellini, Verdi, Massenet, Flotow and

²⁴ As recounted by Silbiger (1980), Frescobaldi's far-reaching influence on seventeenth- and eighteenth-century keyboard (organ) music is universally acknowledged today. The true inheritors of his art are to be found outside his native Italy. "[Frescobaldi's] brilliant pupil Johann Jacob Froberger carried the torch north and established a tradition there that was passed on by successive generations of German organists across at least a century."

²⁵ Another important influencer of the North German organ tradition, especially on Scheidt, and whose influence was eventually handed on to Buxtehude and later yet on to J.S. Bach, is the Dutch composer and organist Sweelinck, who is excluded from this graph because he was born earlier, in 1562 (late Renaissance).

²⁶ As mentioned by Ritter (1994), "Telemann's works display a complete absorption and fusion of many individual national characteristics which establish him as an important link in the development of Baroque music and a precursor of the new *galant* style."

²⁷ See for example Cassedy (2010): "While some of [Beethoven's] earliest compositions in the 1790s had represented a departure from a musical past associated with Mozart and Haydn, by 3 years into the new century he was creating music that defiantly rejected traditions and redefined conventions. [...] [T]he Sonata in F minor, op. 57 (Appassionata), composed in 1804 and 1805, serves as the earliest and boldest example of Beethoven's break with the past and his entry into the territory of what [ETA] Hoffmann characterized as 'romantic.'" On the other hand Clementi, clustered in Fig. 4 with late Classical composers, is sometimes presented as the forerunner of Beethoven (see de Saint-Foix and Herter Norton 1931).

²⁸ To limit the size of the dendrogram in Fig. 5 (and the corresponding MDS map in Fig. 8), late Romantic composers born after 1861 have been excluded.

Gounod).²⁹ We have a single element cluster, Janáček, whose forward-looking innovative compositions led to modernism (see Steinberg 2006). And finally a very large group of ‘core’ Romantic composers (from Tarrega to Raff, passing by Mendelssohn, Liszt, Tchaikovsky and Mahler). This last cluster is made up of many sub-clusters. Without going into much detail, we observe that the sub-clusters have a typically strong regional component. First, Spain with Tarrega and Albéniz. Then comes Russia with ‘The Five’ of the new Russian School (Balakirev, Borodin, Rimsky-Korsakov, Mussorgsky, Cui), plus Tchaikovsky and a few other Russians. Then, a core of early to middle key German and German-influenced composers (from Robert Schumann to Brahms, passing by Schubert and Felix Mendelssohn, plus Liszt (Hungarian), Dvorák (Czech), Chopin (Polish/French), Franck (Belgian/French), Fauré (French) and Elgar (British)).³⁰ Then, a sub-cluster of French composers (from Chaminade to Chausson, and including the American pianist and composer of French/Creole origin Louis Moreau Gottschalk). Finally we have two sub-clusters which are not specifically based on a regional grouping. The first one, made of Sinding (Norwegian), MacDowell (American), Moszkowski and Paderewski (both Polish), represent a group of composers who are well-known for their piano compositions. The second, more heteroclitic group, is made up of a bunch of ‘national school’ composers whose music often became a means of asserting their national identity or else reflects their countries or regions of origin (Smetana and Fibich (Czech), Grieg and Svendsen (Norwegian), Gade (Danish), Stanford (Irish), Parry and Sullivan (English), Chadwick and Foote (Americans), Lalo (French), the German composers and pedagogues Carl Reinecke who taught to Grieg, Svendsen, and Stanford, and Rheinberger who taught to Chadwick. Arguably more difficult to cluster here, we also have Raff (Swiss), and Goldmark, Wolf, Bruckner and Mahler (Austrians). Interestingly, we also find two women in this sub-cluster, grouped with two German composers (C. Loewe and R. Franz) essentially known, as them, for their lieder and piano pieces: Clara Wieck-Schumann and Fanny Mendelssohn-Hensel, respectively the wife of Robert Schumann and the sister of Felix Mendelssohn (who we encountered previously in the ‘key’ German Romantic cluster). That they are not in the cluster of their husband or brother is perhaps the result of what T&G (2013) call ‘genius restrained’.³¹ Finally, one cluster in Fig. 5, which is made of Bruch, Hartmann, and Humperdinck, is the result of the algorithm used but does not seem to have much justification on the basis of their music. Humperdinck would probably fit better into the cluster of predominantly-opera composers, J.P.E. Hartmann (Danish) would fit better in the more heteroclitic national school cluster while Bruch (German) falls into the Romantic classicism exemplified by Brahms.

²⁹ Bottesini was also a famous double bass virtuoso and composer for his instrument.

³⁰ These non-German composers, although also having a ‘national’ school component (see later), have all been very much influenced by the symphony, concerto, chamber, lieder/songs and piano music tradition of their German colleagues in this cluster.

³¹ Technically, they are mainly associated in our ecological categories with lieder and piano pieces. As a comparison, R. Schumann and F. Mendelssohn are also associated with larger, more ambitious categories such as orchestral works (including concertos) and symphonies. As for ‘genius restrained’, T&G (2013) recount that “[Fanny Mendelssohn’s] artistic isolation and probably her creative blocks were the result of the discouragement she received, from her father and later from her brother, when it came to pursuing a career.” As for Clara Wieck-Schumann, a far more significant and public figure of the nineteenth century, T&G (2013) explain that “she stopped composing entirely after [her husband]’s death in 1856, when at age thirty-six she found herself a single mother of seven (one child had died).”

Multidimensional scaling: 2 dimensions

Multidimensional scaling (MDS) is a technique that generates a map displaying the relative positions of a number of objects based on a given set of pairwise distances between these objects. The following example may help to understand the essence of MDS. Given a geographical map of North America and a scale, one can compute the aerial distances between cities. If instead the initial data is a set of pairwise distances between North American cities, one can attempt to recover the geographical map of North America (within about a symmetry and/or rotation). MDS is a methodology that uses algorithms to implement this idea. Although MDS can generate a two-dimensional map that could perhaps, or hopefully, be interpreted as latitude and longitude in the geographical example, the technique per se can be used to generate more than two dimensions from a given distance matrix. A third dimension could perhaps be interpreted as the average altitude of cities with respect to sea level (i.e., topography).

We can apply the MDS methodology to the 500×500 matrix of pairwise distances across composers, S_{comb} .³² In this case, the MDS algorithm aims to position each of the 500 composers in an N -dimensional space (i.e., assigning coordinates) such that the primary distances between composers are preserved as well as possible, according to an optimisation procedure.³³ Choosing $N=2$ optimizes composers location in a two-dimensional scatterplot. In this case, given the distance d_{ij} between composers i and j , the algorithm generates the coordinates (x_i, y_i) and (x_j, y_j) . The MDS algorithm typically computes coordinates (x, y) so as to minimize a loss function called ‘stress’, which is a sum of squared errors between the actual distance across any two composers, d_{ij} , and the predicted distance d_{ij}^* computed by the algorithm:

$$\text{Stress} = \sqrt{\sum_{i < j} (d_{ij} - d_{ij}^*)^2 / \sum_{i < j} d_{ij}^2},$$

and where the predicted distances depend on the number of dimensions kept and the algorithm that is used. Stress values near zero are the best.³⁴

We first apply the MDS methodology to composers of specific periods (Baroque, Classical and Romantic) in Figs. 6, 7 and 8.³⁵ Note that the clusters that we have observed in dendrograms of Figs. 3, 4 and 5 are circled, as much as possible, on the MDS maps of the corresponding period, using the same color code. We can notice a strong correspondence between clusters in the dendrograms and the relative placement of composers on the MDS maps. The maps and dendrograms developed here also provide a music educational tool. For example, if one is interested in the guitar music of classical composer Fernando Sor,

³² Chapter 435 of the NCSS Statistical Software offers an accessible introduction to multidimensional scaling.

³³ Our initial composers’ proximity matrix does not represent pairwise distances across composers, d_{ij} , but pairwise similarities, s_{ij} . Typically similarity indices are converted into distance indices using the formula: $d_{ij} = \sqrt{s_{i,i} + s_{j,j} - 2s_{ij}}$.

³⁴ Following Kruskal (1964), a value of 0 is a perfect goodness-of-fit, 0.05 is good, 0.1 is fair and 0.2 is poor. More recent articles caution against using this advice since acceptable values of stress depends on the quality of the distance matrix and the number of objects in that matrix.

³⁵ The MDS maps have been generated with XLSTAT (Addinsoft 2019).

Fig. 6 European art music. The Baroque period. An application of the MDS methodology

Fig. 7 European art music. The Classical period. An application of the MDS methodology

Fig. 8 European art music. The Romantic period. An application of the MDS methodology

then Fig. 7 (west side) suggests that the music of Giuliani and Carulli might also appeal to them. And indeed these two composers also mainly composed for the classical guitar.

Finally, Fig. 9 presents MDS results based on a 200×200 partition of the 500×500 similarity matrix S_{comb} .³⁶ Hence the map represents 200 ‘key’ composers among a set of 500.³⁶ The date of birth of composers is never used in our methodology to assess the similarity of composers.³⁷ Yet, the map unfolds the history of classical music, starting in the South and moving ‘clockwise’ as centuries pass by. We start our journey in the fourteenth century in France with Guillaume de Machaut and the Medieval ‘Ars Nova’ tradition. We then encounter the musical practice of the Renaissance (1400–1600) with Dufay, in what is now Belgium, a practice that eventually spreads to the rest of Europe through Josquin, Lasso, Palestrina, Gesualdo, Gabrieli, Victoria, and Byrd. Baroque music (1600–1750) emerges in ‘Italy’ (Monteverdi, Frescobaldi), and then in ‘Germany’ (Schütz), evolving in phases to reach mid-Baroque (Lully, M.-A. Charpentier, Buxtehude, Pachelbel, Corelli, and A. Scarlatti), during which classical music progressively emancipates from previously predominant vocal works, and high/late Baroque (F. Couperin, Telemann, Vivaldi, Albinoni, Rameau, J.S. Bach, Handel, D. Scarlatti, and Tartini). The Classical period (1750–1820) arises with a period of transition that is represented by early or pre-Classical composers (C.P.E. Bach, J.C. Bach, and Gluck) followed by ‘core’ Classical composers (Boccherini, J. Haydn, M. Haydn, W.A. Mozart, Clementi, and Beethoven), and ‘late’ Classical composers (Hummel, Schubert, Spohr, and Rossini). We have moved clockwise to more than 90° and are now on the North-West side of the map. Romanticism (1810–1920) starts and develops into two branches. Towards the outside of the map we encounter composers who are mainly remembered for their opera (Weber and the first German Romantic opera, Bellini and Donizetti and the Bel Canto (showcase of the singers’ talents), Meyerbeer and the ‘French Grand Opera’, followed by Glinka who signals the emergence of Russia in the classical music world. Then comes Verdi and his progressive focus on the dramatic aspects of opera, Wagner and his ‘Music Dramas’, the French with Gounod, Bizet, Massenet, Delibes, and the Verismo (realism) tradition epitomised by Leoncavallo. This has moved us another 90° and we are now on the North side of the map. The Verismo tradition eventually dies out with Puccini and Mascagni (in the North-East quadrant). Meanwhile, composers for who opera is not the main or only contribution, represent the other branch of Romanticism. It can be traced back from Beethoven and Schubert (that we encountered before, in the North West side of the map). Moving inside the circle (instead of toward/along the border), we meet Mendelssohn, R. and C. Schumann, Brahms, Bruckner, Liszt, Franck, Saint-Saëns, Dvorák, Tchaikovsky who merges with a few other Russians (Balakirev, Borodin, Rubinstein, Mussorgsky, and Rimsky-Korsakov).

We are now in the late Romantic (North-East) quadrant with composers born well into the second half of the nineteenth century and who also composed during the twentieth century. The history of European art music becomes more complex to recount. Composers adopted distinct paths, as if there were ‘two centuries in one’. As Pauls (2014) puts it: “An outstanding feature of the twentieth-century has been the divergence of European ‘art’

³⁶ The 200 composers are selected on the basis of their ranking in terms of ‘importance’, a ranking proposed by Smith (2000). A (very crowded) sister map of 500 composers is available from the authors.

³⁷ Of course this does not mean that there is no time dimension in our data set. Clearly, data on the personal musical influences of a subject composer will also include some contemporary composers, and the ecological data have general references to periods. For example for Debussy (from Footnote 5): ballets 1900 on, chamber music 1825 to 1925, etc.

Fig. 9 European art music. Seven centuries in one graph. An application of the MDS methodology

music into two general areas which do not overlap to the same extent that they do in previous centuries. That is, the performing repertoire is at odds, sometimes dramatically so, with a competing canon of works considered to be of greater importance from an evolutionary historical point of view”.³⁸ Thus, we see the late Romantic composers (Sibelius, Rachmaninoff), and to the left on the graph, we encounter the late German/Austrian Romantic tradition (R. Strauss, Mahler, and Reger). Then, reacting to the extreme form of German/Austrian late Romanticism, the impressionism movement emerges in France (Debussy, Ravel), while Bartók’s interest for ethnomusicology leads him to synthesize Modernism and folk music. The First World War shatters the ideals of Romanticism and Stravinsky responds with the Neoclassicism movement, an anti-Romantic aesthetic, while a more rational form of expression (expressionism/atonal and dodecaphonic/12-tone/serial music) emerges in Germany (Schoenberg, Berg, and Webern). Later, in France, Messiaen explores rhythm, the Indonesian gamelan, birdsong, and absorbs many non-Western influences. These developments push us in the South East quadrant of the twentieth century music. Towards the inside, we have ‘neo-classical’ and ‘neo-romantic’ currents with relatively accessible (moderate harmonic style) composers of the twentieth century (Hindemith, Honegger, Shostakovich, Martinů, Schnittke, Britten, Walton, Hovhaness, Arnold, etc.) who still compose broadly tonal music in the ‘Classical’ form or the ‘Romantic’ tradition but with elements of atonality, serialism, chromaticism, and in the case of Schnittke, poly-stylistic techniques, and Lutoslawski who has used most of the compositional techniques of the twentieth century (but never considered himself as part of the ‘avant-garde’). But towards the outside of the circle we meet the ‘second century’ within the twentieth century, the ‘avant-garde’ composers and their experimental works: Varèse (organised sound, early electronic music), Carter (metrical modulation), Ligeti (micropolyphony), Cage (electronic and aleatoric/chance music), Boulez (total serialism), Stockhausen (electronic, aleatoric, and serial), Xenakis (computer, electronic, stochastic processes and game theory applied to music), Berio (non-conventional vocal techniques), Feldman (chance, slowly evolving music), Reich (minimalism), Takemitsu (combining oriental and western music) and others, including Penderecki who in the 1970 s re-embraced neoromantic-sounding elements.

To conclude, the MDS map (and its sister map of 500 composers not included here) gives a basic overview, in a single graph, of the evolution of classical music across seven centuries.

Multidimensional scaling (more than 2 dimensions), and canonical correlations

In the previous section, we have presented MDS results on 2-dimension graphs. In Fig. 6 (the Baroque period), the first dimension (the x-axis) seems to be ‘somewhat’ linked to time (composers to the right of the graph are typically born before those on the left). The second dimension (the y-axis) seems to be somewhat associated with the country/region of

³⁸ As the history of music unfolds, the dispersion of (contemporary) composers on the map increases, suggesting larger style differences among composers, a trend which has accelerated in the twentieth century as seen on the East side of the map. Rutherford-Johnson (2017) in his *Music after the Fall* provides an highlighting discussion of how much diverse and fragmented contemporary composition has become, since 1989, a period not captured in Fig. 9.

origin of composers (Italians above, then Germans and Austrians in the center, and French and English composers below). Yet, this ‘guess’-estimation is approximate and appears more difficult to do in maps for the Classical and Romantic periods in Figs. 7 and 8.

One difficulty is that many ecological categories (actually 298 variables including a composer’s country, period, city, style of music (opera, orchestral, etc.), schools of association, etc.) and all personal musical influences are ‘squeezed’ into just two MDS dimensions. If we increased the number of dimensions, perhaps we might be able to identify or ‘match’ some ecological variables with MDS dimensions. One task of MDS is indeed to determine the number of appropriate ‘dimensions’ in the MDS model. The usual technique is to run the MDS algorithm for a series of dimensions (e.g., from 2 to 6 dimensions), and adopt the model with the smallest number of dimensions that achieves a reasonably small value for the stress loss function. Table 2 and Fig. 10 show Kruskal’s stress values we obtained for dimensions 1 to 6 for the 200 ‘Top’ composers analysed in Sect. 2 (Fig. 9). A high stress value tends to create some distortion on the MDS maps, misplacing some composers. In Fig. 9 for example, Giordano (1867–1948) should be close to Puccini and Mascagni (Verismo) instead of being localised on the West side of the map (on the x-axis). Menotti (1911–2007) whose scores are often reminiscent of Puccini is also oddly located in the West side of the map. Marcel Dupré (1886–1971), a French organist and composer, should perhaps be right in the middle of the North East quadrant half way between Widor and Messiaen (instead of being in the South side on the y-axis). Moving from 2 to 3 dimensions permits to lower the stress value significantly. Therefore we should at least explore MDS results with three dimensions. Choosing $N=3$ optimizes the 200 composers location in a three-dimensional space. Given the original distance d_{ij} between composers i and j , the algorithm generates the coordinates (x_i, y_i, z_i) and (x_j, y_j, z_j) .

Instead of drawing three-dimensional graphs that are difficult to visualise our objective is to attempt to identify variables explaining these dimensions. In the example of a geographical map, the three dimensions can be interpreted as latitude, longitude, and perhaps altitude (if this is a ‘level’ or topographic map). Can we say something about the MDS-derived dimensions of a (three-dimension) music map? To address this issue, as mentioned in the Introduction we now use the similarity matrix that is built on personal musical influences only, S_{infl} (instead of S_{comb}). In this setting, composers who have been influenced by a common series of composers tend to be more similar than composers who have been influenced by very distinct composers and they will tend to be closer on the MDS map. Assuming three dimensions, the MDS methodology generates coordinates (x_i, y_i, z_i) for any composer i among the ‘Top’ 200, and we wish to say something about what these three dimensions represent. To do this we also characterize each composer by a few ecological characteristics (period, country of birth, main city where the composer lived) and the composer’s rank from 1 to 200 (as assigned by Smith 2000) which reflects the relative importance of a composer. We then use linear and non-linear canonical correlations to explore the relationship between ecological variables and MDS dimensions.

Canonical correlation analysis is a method for exploring the relationships between multivariate sets of variables. We do not necessarily think of one set of variables as independent and the other as dependent. It is the multivariate extension of correlation analysis. In our example, each composer i is described by a series of coordinates (x_i, y_i, z_i) representing the three dimensions (M_1, M_2, M_3) generated by the MDS methodology, and by a series of ecological variables (countries, cities, period, and the composers’ rank). It might be possible to create pairwise scatter plots with variables in the first set (e.g., the MDS dimensions of the composers), and variables in the second set (e.g., the ecological characteristics of the composers). But if the dimension of the first set is p and that of the second set is

q , there will be pq such scatter plots, and it may be difficult, if not impossible, to look at all of these graphs together and interpret the results. Similarly, one could compute all correlations between variables from the first set and variables in the second set, however interpretation is difficult when pq is large. Canonical correlation analysis allows us to summarize the relationships into a lesser number of statistics while preserving the main facets of the relationships. In a way, the motivation for canonical correlation is very similar to principal component analysis. It is another dimension reduction technique.

There are two types of canonical correlations: linear and non-linear. The mathematics behind both approaches (and their relationships) and the results obtained are described in details in a working paper, Georges and Nguyen (2019), for the interested reader. Here we just provide some basic results. In order to perform linear canonical data analysis all ecological variables had to be converted to dummy variables. For example, the variable Period was converted into binary variables (Baroque, Classical, Romantic, and twentieth century) that take value of 1 or 0 whether a composer belongs or not to such a specific period. The results described in Georges and Nguyen (2019) show that the twentieth century has the most discrimination power with the second (M_2) dimension generated by the MDS methodology. The romantic period has the most discrimination power with the first dimension (M_1) and classical period has the most discriminatory power with the third dimension (M_3). The limit of this analysis, however, is that inference that can be deduced based on linear canonical analysis might not show us the discrimination power of the ecological variables, say, Period, with regards to the MDS dimension/axis, but instead it shows the discrimination power of the dummy variable ‘20-century’ (whether or not a composer belongs to the twentieth century era of music) and the MDS dimension/axis. To some researchers it would be best if the results could be obtained for categorical variables such as Period, Country, or City as a whole. This indeed motivates the use of nonlinear canonical analysis.

A first difference with linear canonical correlation is that the non-linear method can be used to explore the relationship among more than two sets of variables. In Georges and Nguyen (2019) we have considered three sets of variables. Set 1 includes the three MDS dimensions, Set 2 includes City and Country, and Set 3 includes Period and Rank. A second difference between linear and non-linear canonical correlation is that linear canonical correlation works on original (numerical) variables while non-linear canonical correlation works on rescaled variables. Thus, non-linear canonical correlation analysis does not assume an interval level of measurement (in which distances between attributes/categories/scores are meaningful) or assume that the relationships are linear. Variables can be scaled as either nominal (taking a name or label), ordinal, or numerical. In our example, City, Country, and Period are treated as nominal. Ordinal variables are variables with natural, ordered categories.³⁹ By choosing the measurement levels (nominal, ordinal or numerical) of the variables, one tells the non-linear canonical correlation algorithm to choose an optimal transformation of a variable among a set of admissible transformations for that measurement level.⁴⁰

³⁹ For example, we can rate quality of customer services as poor, reasonable, good, or excellent. In this case, there is a natural order of different categories for quality of customer services.

⁴⁰ If nominal level is chosen, observations from one category remain in the same category under rescaling. If ordinal level is chosen, the order of the original (coded) values should be maintained (i.e., if an observation belongs to a ‘lower’ category before transformation, it should not be in ‘higher’ categories after transformation). Note that the measurement level chosen is not necessarily ‘apparent’, based on the meaning of the original variables. Instead, it can be chosen to improve the fit of the model (IBM SPSS Categories 22, 2013). For this reason, in our example, Rank is also treated as nominal.

Table 2 Best stress obtained for each dimension

Dimensions	1	2	3	4	5	6
Kruskal's stress	0.672	0.385	0.277	0.219	0.181	0.156
Iterations	31	205	356	415	369	437
Convergence	0.000	0.000	0.000	0.000	0.000	0.000

Fig. 10 MDS stress values for dimensions 1 to 6

Results in Georges and Nguyen (2019) show that the second MDS dimension is mainly discriminated by Period. Also, the first and third MDS dimensions are mainly discriminated by rank. Besides Ranking and Period, City and Country also contribute some discrimination power but the fact that these variables are correlating to more than one MDS dimension makes it difficult to identify ‘grouping’ of composers in the MDS space. In other words, the placements of composers on a MDS map are determined by a combination of different variables. This is completely understandable for a complicated and large dataset as the one we consider here. When considering the localisation/proximity of composers on a MDS map on the basis of personal musical influences, it is natural to expect that many factors must have had an impact on why they felt this influence, including the period of music, the country or city, the type of music, the types of instruments, the importance (ranking) of composers, etc. These various factors contribute to so much variability/noise in the data that it is somewhat impossible to find a simple one-to-one pairing of MDS variables and ecological variables. One possibility left for future research is to increase the MDS dimensions beyond the three dimensions used here, to add further ecological variables, and to increase the number of composers in the canonical analysis from 200 to 500. Perhaps this might improve the identification of the MDS dimensions.

Conclusion

This paper applies clustering techniques and MDS analysis to a medium size (500×500) composers' similarity matrix built on the basis of personal music influences and ecological data. We provide dendrograms and maps for the Baroque, Classical, and Romantic periods and a map (of 200 'Top' composers) that represents seven centuries of European art music in one single graph. Dendrograms and maps for other periods (Medieval and Renaissance, Impressionism and twentieth century classical music (serial, neo-Classical, neo-Romantic composers, 'avant-garde', etc.), and for all 500 composers are available from the authors, but their analysis is left for a subsequent musicological study.

Comparing our results with those in Weiss (2017) shows that using context-based data, instead of content (audio files) data, remains a useful approach for clustering and visualising similarities across composers. We have shown that grouping well-known composers into either a fixed number of clusters or into hierarchical clusters roughly coincides with common musicological knowledge. In other words, statistical methods can cluster and map composers in a way a musicologist could perhaps organize and recount the history of music on the basis of his/her specific knowledge. In this sense, the maps and dendrograms developed here might also enrich the musicologists' toolkit for basic music education. In this perspective, these graphs (especially those for the twentieth century not included here as they are too large) might benefit from the setting of a web page with a search option. One can envisage 'directed' paths or 'interactive' searches that would lead the user to explore the evolution of Western classical music across centuries or during a specific period, with audio and video files hyperlinked to composers' names. A drop-down menu could also identify/confirm the 5 or 10 most similar composers to a subject composer based on Georges (2017). This is left for a future project.

Acknowledgements The authors thank Charles H. Smith for providing access to the data collected in the Classical Music Navigator for this project. We also benefited from discussions and comments by C.H. Smith on the initial stage of the project.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Addinsoft. (2019). *XLSTAT statistical and data analysis solution*. Long Island, NY, USA. <https://www.xlstat.com>. Accessed 31 Aug 2018.
- Baccigalupo, C., Plaza, E., & Donaldson, J. (2008). Uncovering affinity of artists to multiple genres from social behaviour data. In *Proceedings of the 9th international conference on music information retrieval*, Philadelphia, PA, USA.
- Cassedy, S. (2010). Beethoven the Romantic: How ETA Hoffmann got it right. *Journal of the History of Ideas.*, 71(1), 1–37.
- de Saint-Foix, G., & Herter Norton, M. D. (1931). Clementi, forerunner of Beethoven. *The Musical Quarterly.*, 17(1), 84–92.
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26, 297–302.
- Finscher, L. (1983). Weber's *Freischütz*: conception and misconceptions. *Proceedings of the Royal Musical Association*, 110(1), 79–90.

- Foote, J. (1999). Visualising music and audio using self-similarity. In *MULTIMEDIA '99 Proceedings of the seventh ACM international conference on Multimedia (Part 1)* (pp. 77–80). <https://doi.acm.org/10.1145/319463.319472>.
- Georges, P. (2017). Western classical music development: A statistical analysis of composers similarity, differentiation and evolution. *Scientometrics*, 112(1), 21–53.
- Georges, P., & Nguyen, N. (2019). *Visualizing music similarity: Clustering and mapping 500 classical music composers*. Mimeo: University of Ottawa.
- Giller, G. L. (2012). The statistical properties of random bitstreams and the sampling distribution of cosine similarity. <http://dx.doi.org/10.2139/ssrn.2167044>. Accessed 17 January 2015.
- Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional, and institutional level. *Scientometrics*, 37(2), 195–221.
- Gray, W. (1971). The classical nature of Schubert's lieder. *The Musical Quarterly*, 57(1), 62–72.
- Gray, W. (1978). Schubert the Instrumental composer. *The Musical Quarterly*, 64(4), 483–494.
- Griffiths, P. (1978). *A concise history of modern music: From Debussy to Boulez*. London: Thames & Hudson.
- IBM Corp. (2013). *IBM SPSS categories 22, version 22.0*. Armonk, NY: IBM Corp.
- Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. de la Société Vaudoise de la Science Naturelle*, 37, 547–579.
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1–27.
- Kulczynski, S. (1927). Zespoly roslin w Pieninach. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. ser. B: Sciences Naturelles*, 1(suppl. 2), 57–203.
- Mauch, M., MacCallum, R., Levy, M., & Leroi, A. (2015). The evolution of popular music: USA 1960–2010. *Royal Society Open Science*, 2, 150081.
- NCSS Statistical Software. Documentation available at <https://www.ncss.com/>. Accessed 31 Aug 2018.
- Orio, N. (2006). Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, 1(1), 1–90.
- Pachet, F., Westerman, G., & Laigre, D. (2001). Music data mining for electronic music distribution. In *Proceedings of the 1st international conference on web delivering of music*, Florence, Italy.
- Pampalk, E. (2006). Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns. In *Proceedings of the international symposium on music information retrieval*.
- Pauls, H. (2014). *Two centuries in one. Musical Romanticism and the twentieth century*. Ph.D. thesis, Rostock University of Music and Theatre.
- Potthoff, R. F., & Whittinghill, M. (1966). Testing for homogeneity. I. The binomial and multinomial distributions. *Biometrika*, 53, 167–182. <https://doi.org/10.1093/biomet/53.1-2.167>.
- Ritter, M. (1994). Telemann: master of the 'German style'. *Early Music*, 22(4), 696–698.
- Rodriguez Zivic, P. H., Shifres, F., & Cecchi, G. A. (2013). Perceptual basis of evolving Western musical styles. *Proceedings of the National Academy of Sciences*, 110(24), 10034–10038.
- Rutherford-Johnson, T. (2017). *Music after the fall. Modern composition and culture since 1989*. University of California Press, Oakland, CA.
- Schedl, M., Gomez, E., & Urbano, J. (2014). Music information retrieval: recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2–3), 127–261.
- Sen, S. K., & Gan, S. K. (1983). A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library and Information Studies*, 30(2), 78–823.
- Sheldon, D. A. (1989). The concept *galant* in the 18th century. *Journal of Musicological Research*, 9(2–3), 89–108.
- Silbiger, A. (1980). The roman Frescobaldi tradition, c. 1640–1670. *Journal of the American Musicological Society*, 33(1), 42–87.
- Simpson, G. G. (1943). Mammals and the nature of continents. *American Journal of Science*, 241, 1–31. <https://doi.org/10.2475/ajs.241.1.1>.
- Smith, C. H. (1983). A system of world mammal faunal regions. I. Logical and statistical derivation of the regions. *Journal of Biogeography*, 10, 455–466.
- Smith, C. H. (2000). The classical music navigator. <http://people.wku.edu/charles.smith/music/>. Accessed 30 August 2018.
- Smith, C. H., & Georges, P. (2014). Composer similarities through 'the Classical Music Navigator': Similarity inference from composer influences. *Empirical Studies of the Arts*, 32(2), 205–229.
- Smith, C. H., & Georges, P. (2015). Similarity indices for 500 classical music composers: Inferences from personal musical influences and 'ecological' measures. *Empirical Studies of the Arts*, 33(1), 61–94.

- Smith, C. H., Georges, P., & Nguyen, N. (2015). Statistical tests for ‘related records’ search results. *Scientometrics*, 105(3), 1665–1677.
- Sneath, P. H. A. (1968). Vigour and pattern in taxonomy. *Journal of General Microbiology*, 54, 1–11. <https://doi.org/10.1099/00221287-54-1-1>.
- Steinberg, M. P. (2006). The politics and aesthetics of operatic modernism. *The Journal of Interdisciplinary History*, 36(4), 629–648.
- Sutherland, D. (1995). Domenico Scarlatti and the Florentine piano. *Early Music*, 23(2), 243–246 and 249–256.
- Taruskin, R. (2010). *The Oxford dictionary of western music*. Five-volume edition, Oxford, New York: Oxford University Press.
- Taruskin, R., & Gibbs, C. (2013). *The Oxford history of western music*. New York, Oxford: Oxford University Press.
- Todd, L. R. (2008). *Mendelssohn essays. Chapter 7: The Chamber Music of Mendelssohn*. New York: Routledge.
- Vehkalahti, K., & Everitt, B. S. (2019). *Multivariate analysis for the behavioral sciences*. Boca Raton, FL: CRC Press, Francis & Taylor Group.
- Weiss, C. (2017). *Computational methods for tonality-based style analysis of classical music audio recordings* (Doctoral dissertation). Ilmenau University of Technology. Ilmenau, Germany.
- Weiss, C., Mauch, M., Dixon, S., & Müller, M. (2018). Investigating style evolution of Western classical music: A computational approach. *Musicae Scientiae*. <https://doi.org/10.1177/1029864918757595>.
- White, C. W. (2013). *Some statistical properties of tonality, 1650–1900* (Doctoral dissertation). Yale University, New Haven, CT.