Compositional Data Analysis of Harmonic Structures in Popular Music

Burgoyne, J.A.; Wild, J.; Fujinaga, I.

DOI
10.1007/978-3-642-39357-0_4

Publication date
2013

Document Version
Final published version

Published in
Mathematics and Computation in Music

Compositional Data Analysis of Harmonic Structures in Popular Music

John Ashley Burgoyne[1], Jonathan Wild[2], and Ichiro Fujinaga[2]

[1] Institute for Language, Logic, and Computation
Universiteit van Amsterdam, Amsterdam, North Holland, The Netherlands
j.a.burgoyne@uva.nl
[2] Centre for Interdisciplinary Research in Music and Media Technology
McGill University, Montréal, Québec, Canada
{wild,ich}@music.mcgill.ca

Abstract. While analysing large corpora of music, many of the questions that arise involve the proportion of some musical entity relative to one or more similar entities, for example, the relative proportions of tonic, dominant, and subdominant chords. Traditional statistical techniques, however, are fraught with problems when answering such questions. *Compositional data analysis* is a more suitable approach, based on sounder mathematical (and musicological) ground. This paper introduces some basic techniques of compositional data analysis and uses them to identify and illustrate changes in harmonic usage in American popular music as it evolved from the 1950s through the 1990s, based on the McGill *Billboard* data set of chord transcriptions.

Keywords: compositional data analysis, popular music, harmony.

1 Introduction

Many questions in computational musicology involve the relative frequencies of different musical entities, for example, how often different jazz arrangers use tritone substitutions as opposed to more traditional dominant chords or how the typical distributions of pitches or pitch classes vary among different Renaissance modes. The analysis of data in this form is fraught with subtle challenges. Consider corpus-based research on harmony in popular music, which has been active recently due to the release of two large data sets [1,2]. If, as in de Clercq and Temperley's own analysis of their new data set [1], one examines the total number of instances of each chord root in the corpus without first normalising the counts of chords within each song to sum to a common total, one implicitly assumes that individual chords, rather than complete songs, have been sampled. This false assumption lends inappropriate extra weight to longer songs and insufficient weight to shorter songs; more importantly, it also precludes any meaningful method for computing or understanding the variance or covariance of the relative frequencies of chords across songs. Without some notion of variance, there is no way to predict whether the patterns of chords within the corpus overall are meaningful or merely the result of chance.

J. Yust, J. Wild, and J.A. Burgoyne (Eds.): MCM 2013, LNAI 7937, pp. 52–63, 2013.

Normalising the frequencies of chords to sum a common total seems like it would solve both problems, but in fact, it only solves the problem of weighting. It is possible and, in principle, meaningful to compute the variance of the relative proportions of each chord across songs, but in practise, it is difficult to know how to interpret such variances: being restricted to fall between zero and one, relative proportions cannot follow a normal distribution or anything like it. Worse, as has been known since the late 19th century, computing covariances or inter-correlations on data normalised in this way introduces spurious negative correlations that can cause seemingly similar analyses to draw opposing conclusions [3,4]. *Compositional data analysis* constitutes a collection of techniques that have been devised since the 1980s to handle these problems, originally for the use of mathematical geologists studying rock compositions [5]. It has since expanded to numerous other fields, but to the authors' knowledge, no researchers have yet used these techniques to analyse musical data.

In this paper, we introduce some basic techniques of compositional data analysis and, as a case study, use them to investigate two questions about harmony in American popular music from the later twentieth century: How did the distribution of harmonies in popular music (identified by their roots alone, for the sake simplicity) evolve throughout this period, and did the distribution of harmonies in a song have an impact on its popularity?

2 Method

2.1 Musical Material

We used the McGill *Billboard* data set for our musical material [2,6]. This collection contains complete transcriptions of the harmony and musical structure in a random sample of 1 379 singles drawn from *Billboard* magazine's Hot 100 chart between 1958 and 1991. Published weekly, the Hot 100 is a ranked list of the 100 singles that *Billboard* deems to be most popular in the United States at the time of publication, based on radio airplay and record sales. The *Billboard* data set is notable for the detail of its transcriptions and its careful sampling methodology, which is well-suited for longitudinal studies of popular music.

For the analyses in this paper, we extracted the roots of the chords on each beat and totalled the number of beats spent on each root pitch class. Following John Snyder [7], who demonstrated that, because so many key changes are too brief to be recorded, statistical analysis with roots recentered at formal key changes can lead to misleading conclusions, we then replaced the pitch classes with their corresponding scale degree in the prevailing overall key of the song in which they appeared. We did consider enharmonic equivalents to have the same pitch class, however, despite Snyder's recommendations against it, because the pitch spellings used in chord symbols for popular music typically favour convenience over theoretical correctness. We normalised the counts for each root scale degree to yield what we call the *root composition* for each song: the relative proportion of time spent on chords rooted at each scale degree.

2.2 Compositional Data Analysis

The key insight behind compositional data analysis is that compositional data can be best understood as a collection of *log odds ratios*, the logarithm of the ratios of the individual component values against one another [8]. In this sense, compositional data analysis is quite similar to logistic regression, particularly the multinomial logit model [9,10]. Unlike the individual components of a composition, log odds ratios range from $-\infty$ to ∞ and can be modeled with any of probability distribution with support across the real line. Compositional data analysis typically employs a multivariate normal distribution to yield, in effect, a multinomial probit model. This model allows for reliable statistical answers to questions like 'How many subdominant chords are there, on average, relative to tonic chords, and is that ratio correlated with the number of dominant chords?' or 'Is there a difference between the frequency of use of major mediant and submediant (III and VI) and that of the minor mediant and submediant (♭III and ♭VI)?'. It also allows researchers to identify notable deviations from these norms and to quantify how these patterns evolved over time.

When analyzing compositional data, it is particularly important consider which ratios of components are the most interesting or relevant. Because compositional data represent parts of a complete whole (they typically, for example, are constrained to sum to 1 or 100 in order to represent proportions or percentages), a composition with N components (12 in our case, one for each pitch class), has only $N-1$ degrees of freedom (11 in our case). Thus, the $N(N-1)/2$ possible log ratios to consider must be reduced in some principled way to only $N-1$. There are infinitely many ways to do so, but analogous to the choice of contrasts for factors in an analysis of variance (ANOVA), some ways have more convenient mathematical behaviors or are easier to interpret. The dominant approach today starts with a binary tree over the components of a composition [11]. Each non-leaf node of this tree corresponds to a *balance*: a weighted log ratio of the geometric mean of all leaves underneath one branch against the geometric mean of all leaves underneath the other branch, or more formally, a function $b : \triangle^N \to \mathbb{R}$ such that

$$b(\boldsymbol{p}) = \left(\frac{|L||R|}{|L|+|R|}\right)^{1/2} \log \frac{\left(\prod_{i\in R} p_i\right)^{1/|R|}}{\left(\prod_{i\in L} p_i\right)^{1/|L|}} \quad , \tag{1}$$

where L is the set of all leaves in the left branch of the balance and R is the set of all leaves in the right branch. The weight is used to normalise the balance such that, under a special geometry that is used for compositional data analysis, they correspond to a unit-norm vector. Balances generated from binary trees are orthogonal in this geometry, and thus a collective set of such balances are said to be an *isometric log ratio* (ILR) transformation of the underlying composition [12,13]. ILR-transformed compositions have $N-1$ components that capture information about all N components of the original composition, and because of the underlying orthogonality, they behave well mathematically [14]; moreover, because researchers are free to design the binary tree as they see fit, it is possible to create balances that have a useful interpretation.

Figure 1 illustrates the binary tree we used to generate the balances for our experiments in this study. The shaded boxes (leaf nodes) represent chord roots and the white circles (non-leaf nodes) represent balances. Left branches represent the denominator of the log ratio and right branches represent the numerator. Balance 1, for example, is the log ratio of the number of beats spent on subdominant chords over the number of beats spent on tonic chords. Balance 2 is the log ratio of the geometric mean of the number of beats spent on either tonic or subdominant chords over the number of beats spent on dominant chords. The other balances are analogous. Because of the logarithm, positive values of the balance imply that there are relatively more beats spent on chord roots under the right branch of a balance; negative values imply that there are relatively more beats spent on the left branch.

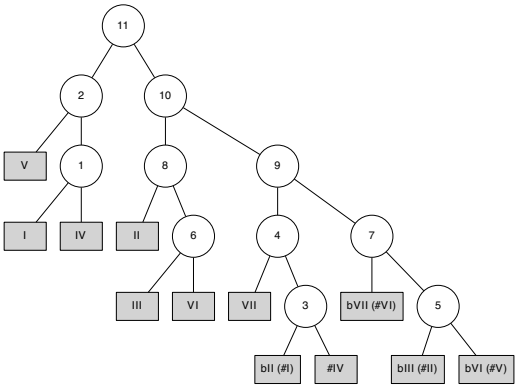


Fig. 1. Binary partition of chord roots used to generate balances (contrasts) for isometric log-ratio (ILR) transformations. Shaded boxes represent chord roots and unshaded circles represent balances. Negative values of a balance reflect a relative greater proportion of chords with roots anywhere in the left-hand tree of that balance; positive values reflect a great proportion in the right-hand tree. Although the structure arose from a hierarchical clustering technique designed to reduce the effect of zero counts, it yields relatively interpretable balances. Balance 11, for example, contrasts the basic chords (I, IV, and V) against all other chords; Balance 9 contrasts chords that arise in the minor mode or as borrowings from the minor mode (♭III, ♭VI, and ♭VII) against exotic chords (natural VII and chords rooted on altered scale degrees).

3 Results

3.1 Root Compositions

Zero Counts and Balances. Only 24 songs contain chords rooted on all twelve scale degrees, and thus, most root compositions contain at least one component with a count of zero beats. Compositional data analysis relies on the existence of ratios and logarithms, and thus these zero components require special treatment. This topic is an open field of research in compositional data analysis [15], and

the issues involved are more subtle than they might seem. We chose an easy-to-implement strategy: adding one beat to every count in every composition in our data set, or in statistical language, applying the *Bayes–Laplace prior*. The disadvantage of this approach is that it alters the ratios between all components slightly, although the effect is small for ratios with large counts, such as the ratios among the most common chords.

As one would expect, some roots are much rarer than others (e.g., 84% of songs in the data set have no instances ♭II). Moreover, these zero components are correlated: Songs without ♭VI chords, for example, are likely not to have ♭III chords. In order to reduce the distortions arising from the replacement of zeros, we used a technique from the R package 'compositions' that takes advantage of these correlations to derive balances that avoid ratios between zero and non-zero components as much as possible, resulting in the structure depicted in Fig. 1.[1] Much as this technique helped, one should nonetheless interpret balances between rare roots (e.g., Balances 3 and 4) with care. Any songs that contain none of the roots involved in a balance will necessarily have a balance value of zero, biasing the results toward a more even distribution of roots. This bias reduces the power of statistical tests involving these balances, and thus there is a risk that we have overlooked effects related to them.

Effects of Decade. We undertook a multivariate analysis of variance (MANOVA) on the root compositions to test for the effect of the decade in which a single was first distributed (1950s, 1960s, 1970s, 1980s, or 1990s) on its root composition. We found that the decade had a highly significant effect on root compositions: $\mathrm{V} = 0.14$, $F(44, 5388) = 4.53$, $p < .001$. In order to understand the effect of decade on specific balances, we then examined univariate analyses of variance (ANOVA) for each balance independently. The decade was significant at $p < .05$, controlling for false discoveries with the Benjamini–Hochberg procedure [17], for Balance 2 [$F(1, 4) = 4.42$, $MSE = 7.26$, $p = .001$], Balance 5 [$F(1, 4) = 6.48$, $MSE = 8.20$, $p < .001$], Balance 7 [$F(1, 4) = 5.60$, $MSE = 9.84$, $p < .001$], Balance 8 [$F(1, 4) = 6.85$, $MSE = 11.80$, $p < .001$], Balance 9 [$F(1, 4) = 19.95$, $MSE = 47.76$, $p < .001$], and Balance 10 [$F(1, 4) = 3.02$, $MSE = 16.37$, $p = .017$]. Finally, for the balances on which decade had a significant effect, we used t-tests to identify significant contrasts and the size of their effects. Because we were most interested in changes in harmonic practice over time, we used Helmert coding to create contrasts of the root composition for each decade against the average root composition in all previous decades. Table 1 presents a list of the significant contrasts, again using the Benjamini–Hochberg procedure to control for false discoveries at $p < .05$.

It is clear from Table 1 that there was a particularly significant change in root compositions starting in the 1980s. There is also evidence of a break in the 1970s, but due to space limitations, we restrict our analysis here to two groups: singles released before 1980 ($n = 913$) and singles released in 1980 or afterward ($n = 466$). Table 2 presents some descriptive statistics for these two groups,

[1] `http://www.stat.boogaart.de/compositions/`

Table 1. Effect Sizes of Significant Decade Contrasts for Predicting Balances

Balance & Contrast	p	% increase	95% CI *LL*	95% CI *UL*
Balance 2 (I and IV vs. V)				
1970s vs. prior years	.014	10	0	21
1980s vs. prior years	< .001	10	3	16
Balance 5 (♭VI vs. ♭III)				
1980s vs. prior years	< .001	9	4	15
Balance 7 (♭III and ♭VI vs. ♭VII)				
1980s vs. prior years	< .001	−9	−14	−4
Balance 8 (III and VI vs. II)				
1980s vs. prior years	< .001	10	3	16
Balance 9 (minor borrowings vs. other exotics)				
1970s vs. prior years	< .001	27	14	43
1980s vs. prior years	< .001	25	17	34
Balance 10 (minor tonality vs. major tonality)				
1980s vs. prior years	.009	12	1	24

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit. The CIs have been adjusted to maintain a false coverage rate < .05 [16].

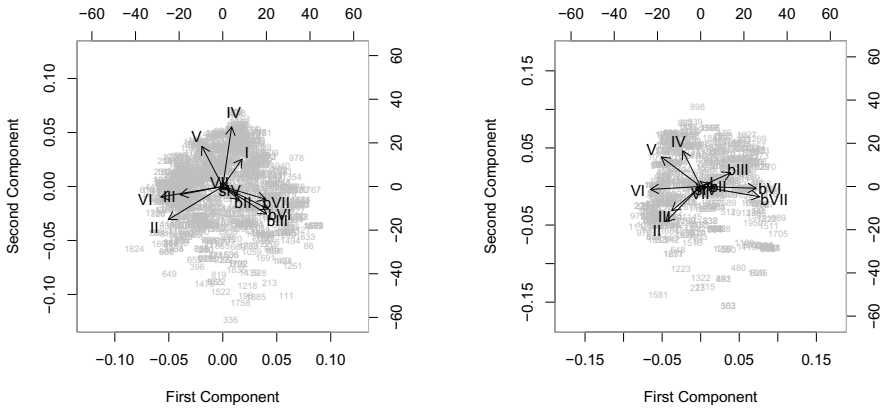
Table 2. Descriptive Statistics for Balances Pre- and Post-1980

Bal.	1	2	3	4	5	6	7	8	9	10	11
Years 1958–1979											
M	−0.83	0.59	−0.03	0.15	−0.04	0.44	−0.31	−0.42	0.76	−1.26	−4.51
SD	1.00	1.21	0.71	0.88	0.98	1.16	1.30	1.28	1.46	2.16	1.67
Pearson's correlation coefficients											
1	–	.08*	.13*	.02	.00	−.02	−.13*	.12*	−.06	−.15*	−.31*
2	−.12*	–	.02	.05	−.12*	−.14*	−.11*	.09*	.22*	.34*	−.01
3	.05	−.07	–	−.28*	.02	−.02	−.14*	.18*	−.13*	−.18*	−.10*
4	.10	.12*	−.29*	–	.11*	−.06	.18*	−.11*	−.04	.25*	.18*
5	−.29*	.06	−.06	.03	–	.05	.10*	.08*	−.14*	.08*	−.02
6	.18*	−.20*	−.03	−.05	−.07	–	−.03	.12*	−.06	−.17*	.03
7	.18*	−.19*	.02	.18*	−.11*	−.01	–	−.07*	−.18*	.07	.07*
8	.13*	−.03	.09	−.06	.00	.20*	.10*	–	.00	.03	−.12*
9	−.20*	.30*	−.27*	.06	.20*	−.11*	−.31*	−.07	–	.43*	.15*
10	−.29*	.42*	−.23*	.30*	.27*	−.24*	−.13*	.00	.66*	–	−.09*
11	−.46*	.35*	−.11*	.19*	.25*	−.28*	−.15*	−.08	.35*	.28*	–
Years 1980–1991											
M	−0.77	0.84	−0.07	0.12	0.27	0.54	−0.64	−0.12	1.36	−0.85	−4.46
SD	1.24	1.44	0.72	0.80	1.37	1.30	1.37	1.39	1.75	2.62	1.68

Note. *M* = sample mean; *SD* = sample standard deviation. Balances correspond to those in Fig. 1 and should be interpreted according to the principles described there. Inter-correlations between the balances for singles released between 1958 and 1979 ($n = 913$) appear above the diagonal; inter-correlations for singles released between 1980 and 1991 ($n = 466$) appear below.

*$p < .05$, controlling for false discoveries with the Benjamini–Hochberg procedure [17].

specifically the means and standard deviations for each balance as well as the intra-group correlation matrices (i.e., the correlation matrices of the MANOVA error for each group). From the standard deviations, it is possible to derive the *total variation* for each group: the trace of the variance-covariance matrix, or the sum of the MSE for each balance. For pre-1980 singles, the total variation is 18.93 and for post-1980 singles it is 24.92; we trusted ANOVA to be robust to this heteroskedasticity. Roughly two thirds of the inter-correlations are significant for each group, controlling for false discoveries at $p < .05$.



(a) 1958–1979 (46% of total variance) (b) 1980–1991 (53% of total variance)

Fig. 2. Covariance biplots pre- and post-1980. Each song appears with its leading two loadings (bottom and left axes). Each chord root appears with its leading two principal components (top and right axes). Distances between songs on the graphs approximate the corresponding Mahalanobis distances; songs far from the centre are potential outliers. Distances between roots on the graphs approximate the standard deviation of the log-ratio between the chords, and the cosine of the angle between their rays on the graphs corresponds to their approximate degree of correlation. The first component in each plot seems to correspond to major (left) vs. minor (right). The second component seems to correspond to *tonalness*: songs toward the top use more traditional harmonic language than those toward the bottom.

In order to examine the underlying structure of the data and to identify potential outliers, we produced biplots of the chord roots and songs pre- and post-1980 (Fig. 2). Biplots can be difficult to interpret – see [18] for an explanation of some principles for biplots of compositional data – but for our purposes, only a few aspects are important. Using the left and bottom axes, the Euclidean distances between songs on the chart approximates the Mahalanobis distance between the songs under a compositional data model; as a result, songs that appear far from the centre are potential outliers. Using the right and top axes, the Euclidean distances between chord roots approximate the standard deviation of the log-ratio of the frequencies of these roots. Moreover, the cosine of the angle between

two line segments that connect two pairs of roots approximates to the correlation between the corresponding log-ratios. Overall, for both chord roots and songs, the horizontal axis represents the primary underlying organisational pattern and the vertical represents the secondary underlying organisational pattern, although the data are sufficiently complex that these two-dimensional plots are rough approximations: they account for only half of the total variation.

3.2 Popularity

We also used a proportional-odds probit model to test the effect of root composition on *peak chart quintile*, the highest-ranking quintile of the chart (top, second, middle, fourth, or bottom) that a single reached at any point during the period of time from which we sampled. Proportional-odds models assume that rankings come from a continuous latent variable (popularity, in our case) but that it is only possible to observe which of a series of adjoining bins, defined by cutoff points between adjacent ranks, contains the actual value of this latent variable. Proportional-odds regression uses maximum-likelihood estimation to derive both the cutoff points and the regression coefficients given a distribution for the latent variable (normal, in our case, to parallel our compositional data model, although logistic variables are also popular). We used a stepwise search procedure to find the best model, progressively adding and removing single variables and selecting the model with the best value of Akaike's Information Criterion (AIC) for the following step. The same model is preferred regardless of whether one starts from an empty model or a fully general model (all balances, decade, and the possible interactions between decade and the individual balances). This model includes Balance 1 (0.14, 95% CI $[0.04, 0.25]$), Balance 5 (-0.07, 95% CI $[-0.17, 0.02]$), and Balance 11 (0.13, 95% CI $[0.07, 0.20]$). The cutoff points are as follows: Bottom vs. Fourth $= -4.5$, 95% CI $[-5.0, -4.0]$; Fourth vs. Middle $= -3.1$, 95% CI $[-3.5, -2.7]$; Middle vs. Second $= -2.1$, 95% CI $[-2.5, -1.7]$; Second vs. Top $= -1.2$, 95% CI $[-1.6, -0.8]$.

4 Discussion

4.1 Tree of Balances

The technique we used to derive balances is a form of clustering that groups chords more closely together the more frequently that they co-occur within individual songs. It is similar to the correlation analysis that de Clercq and Temperley [1] conducted the chord roots in their corpus. In both cases, each chord is represented by a binary vector corresponding to whether or not the chord appears in a song, but our clustering is based on Euclidean distance, whereas theirs is based on correlation coefficients. Our resulting tree of balances is consistent with what de Clercq and Temperley found: two groups of correlations combining II, III and VI on one hand and ♭III, ♭VI, and ♭VII on the other hand, roughly corresponding to our Balance 10, and a correlation between IV and V that is rather uncorrelated with anything else, roughly corresponding to our Balance 2.

Overall, our tree of balances validates many of the traditional groupings of chords from music theory and perceptual experiments. Balance 11 separates the traditional *harmonic core* of I, IV, and V from the remainder of the chords, with Balances 1 and 2 characterising the internal structure of the harmonic core. Balance 10 separates the non-core chords of major tonality, II, III, and VI, from the others, with Balances 6 and 8 capturing the internal structure of this group. Balance 9 separates the non-core chords of minor tonality, ♭III, ♭VI, and ♭VII, from more exotic chords, with Balances 5 and 7 capturing the internal structure of the minor-tonality group. At first glance, it seems strange that these minor-key chords would be grouped so closely with exotic chords – in particular, one might expect Balances 4 and 8 to be exchanged – but the present structure makes sense as an artefact of our policy on the treatment of modulation and enharmonic equivalents. Because we consider all roots with respect to the overall prevailing key of a piece, the minor-key chords can also be major-key chords in a song that modulates up one semitone, which is quite common in this genre; as such, it is quite sensible that these chords group closely with the Neapolitan chord, which is indistinguishable from the tonic after such a modulation, and ♯IV. In future work, we would like to compare this tree (and all other results) with an analysis of the same data but with all roots considered with respect to local key.

4.2 Harmonic Evolution

The results from the MANOVA in Table 1 show above all a greater use of minor tonalities over time. Given the overall predominance of tonic chords, which when transposed up a semitone appear on the negative side of Balance 9, positive values of Balance 9 suggest minor tonality rather than a major tonality with a semitone transposition. The 1970s saw a 27% increase in the frequency of chords on the positive side of Balance 9 relative to the those on the negative side, and the 1980s saw a further 25% increase; there was also a 10% increase in the frequency of chords on the positive side of Balance 10 relative to the major-tonality chords on its negative side. Furthermore, the decrease in Balance 7 that also appears in the 1980s reflects an increase in the use of ♭VII in general.

The other important story from the MANOVA is a decrease in the relative frequency of dominant chords relative to tonic and subdominant chords (Balance 2): 10% in the 1970s and a further 10% in the 1980s. Given that ♭VII–I has long been theorised as an alternative cadential formula in rock music [19], this decline is consistent with the concomitant rise in ♭VII. It is also consistent with Temperley's theory of a cadential IV in rock [20].

The inter-correlation matrices in Table 2 are the richest results of this study, and because space limitations allow us to present only the most salient patterns here, the reader is encouraged to study these matrices in more depth. Each entry in the matrix describes the degree to which two balances are correlated, which are defining aspects of harmonic style. Roughly two thirds of the inter-correlations are significant for each group, controlling for false discoveries at $p < .05$., which reflects the rather tight constraints of harmonic style in this repertoire. For example, before and after 1980, Balances 9 and 10 are highly

positively correlated, which corresponds to the major-minor modal division: higher incidences of II, III, and VI (the left-hand branch of Balance 10 and also the defining harmonies of the major mode) relative to other non-core harmonies correspond to lower incidences of ♭III, ♭VI, and ♭VII (the right-hand branch of Balance 9 and also the defining harmonies of the minor mode) relative to exotic borrowings like the Neapolitan, and vice-versa. Balances 4 and 10 are also highly positively correlated before and after 1980, reflecting the importance of modulation up one semitone: where there are ♯I and ♯IV chords (the right-hand branch of Balance 10) there are also likely to be ♯II, ♯V, and ♯VI chords (other important chords on the right-hand branch of Balance 10).

In some cases, there are significant inter-correlations before and after 1980, but the direction of the correlation changes after 1980. Such cases are even more interesting because they reflect important changes in harmonic usage. For example, Balances 10 and 11 are negatively correlated before 1980 but positively correlated after 1980, which reflects the growing importance of minor tonalities in popular music. In earlier pop music, the more non-core chords in use (the right-hand branch of Balance 11) relative to the harmonic core, the more major-tonality chords (the left-hand branch of Balance 10) among them; after 1980, the pattern is the reverse. Where there were minor-tonality chords, these were more likely to be ♭III and ♭VI before 1980, when Balances 7 and 11 were positively correlated, whereas as after 1980, when Balances 7 and 11 were negatively correlated, the more likely chord was ♭VII, consistent with other results.

Overall, as the standard deviations presented in Table 2 show, Balances 10 and 11 are the two balances that capture the most variation both before and after 1980. This result is consistent with the biplots in Fig. 2, which seem to organise around a major-minor axis, roughly corresponding to Balance 10, and a core–non-core axis, roughly corresponding to Balance 11. We examined all of the potential outliers in these figures, and although none seemed so far off the mark that it should be excluded from the sample, their deviations are illustrative. Song 336, for example (Rita Coolidge's 1977 cover of 'Your Love Has Lifted Me Higher'), has a striking tonic pedal throughout the entire song, shifting harmonically only between I and II^{4}_{2}; Song 352 (Marky Mark and the Funky Bunch's 'Good Vibrations' from 1991), similarly is based on a harmonic vamp that decorates the tonic with no true harmonic function. Song 1824 (George Harrison's 'My Sweet Lord' from 1977) ingeniously denies resolution to tonic harmony throughout the great majority of the song; song 1581 (Evelyn 'Champagne' King's 'Love Come Down' from 1982) is similar in this regard. It is also telling that both before and after 1980, the songs cluster not into the circle one would expect but into a rainbow shape, with few potential outliers to the top, left, or right; this pattern is consistent with the importance of the harmonic core.

4.3 Popularity

One of the more surprising results of the study was that there was no evidence that the effect of the root composition of a song on the popularity of that song changed from decade to decade. Furthermore, contrary to the stereotype of the

'four-chord song', it seems that throughout the period we studied, the richer the harmony, the more popular the song. The positive coefficient for Balance 11 in the proportional-odds model implies that songs with relatively too many core chords (the left-hand branch of Balance 11 in Fig. 1) relative to other chords (the right-hand branch of Balance 11) were not as popular; moreover, the positive coefficient for Balance 1 implies that even within the harmonic core, too many tonic chords (again, the left-hand branch of the balance) were harmful for popularity. The confidence interval for the remaining variable in the model, Balance 5, spans zero, and thus one cannot conclude definitively which branch of this balance helped or harmed popularity; because the value of this balance is only non-zero in pieces that contain either a ♭III or a ♭VI (i.e., minor-mode pieces), it is weak evidence that minor-mode songs were more popular when they contained a relatively large number of ♭III chords, perhaps reflective of a modulation to the relative major.

4.4 Implications for Future Research

Many musical questions of interest are best framed as statistics over compositional data. Compositional data have unique statistical behaviours, however, and failing to account for them properly can lead researchers to erroneous conclusions. This paper provides an example of how to use compositional data analysis to avoid such mistakes. The bridge between compositional data analysis and traditional analysis is the ILR transformation for a specific choice of *balances*. Although there are automatic techniques to generate trees of balances, researchers can also devise customised balances that are tailored to their questions.

Our case study validates some traditional understandings about harmony in popular music. Our tree of balances, based on the co-occurence of harmonies, is consistent with previous findings and encapsulates the traditional grouping of a harmonic core and the major-minor mode system. Our model identifies a key inflection point in harmonic usage around 1980, with usage becoming more varied and more focused on IV and ♭VII harmonies; moreover, the inter-correlation matrices in Table 2 are a handbook of harmonic usage pre- and post-1980.

The 'compositions' package we used for these analyses is open-source and freely available, and our case study could form a template for many other types of style analysis: for example, distributions of pitches, melodic or rhythmic patterns, or longer chord progressions, as contrasted among different composers, schools, or time periods. As interest grows in corpus-based musical analysis, so should our field's sophistication using these techniques.

Acknowledgement. We thank the Social Sciences and Humanities Research Council of Canada (SSHRC), who funded this study under their ITST programme.

References

1. de Clercq, T., Temperley, D.: A corpus analysis of rock harmony. Popular Music 30(1), 47–70 (2011)
2. Burgoyne, J.A., Wild, J., Fujinaga, I.: An expert ground-truth set for audio chord recognition and music analysis. In: Leider, C., Klapuri, A.P. (eds.) Proceedings of the 12th International Conference on Music Information Retrieval, Miami, FL, pp. 633–638 (2011)
3. Pearson, K.: Mathematical contributions to the theory of evolution: On a form of spurious correlation which arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London 60, 489–498 (1897)
4. Egozcue, J.J., Pawlowsky-Glahn, V.: Basic concepts and procedures. In: [21], ch. 2, pp. 12–28
5. Aitchison, J.: The statistical analysis of compositional data. Journal of the Royal Statistical Society, Series B 44(2), 139–177 (1982)
6. Burgoyne, J.A.: Stochastic Processes and Database-Driven Musicology. PhD thesis, McGill University, Montréal, QC (2012)
7. Snyder, J.L.: Entropy as a measure of musical style: The influence of a priori assumptions. Music Theory Spectrum 12(1), 121–160 (1990)
8. Aitchison, J.: The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, London (1986)
9. Fry, T.R.L.: Applications in economics. In: [21], pp. 318–326
10. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Monographs on Statistics and Applied Probability, vol. 37. Chapman & Hall/CRC, Boca Raton, FL (1989)
11. Egozcue, J.J., Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis. Mathematical Geology 37(7), 795–828 (2005)
12. Aitchison, J.: On criteria for measures of compositional difference. Mathematical Geology 24(4), 365–379 (1992)
13. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figuera, G., Barceló-Vidal, C.: Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3), 279–300 (2003)
14. Mateu-Figuera, G., Pawlowsky-Glahn, V., Egozcue, J.J.: The principle of working on coordinates. In: [21], pp. 29–42
15. Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A.: Dealing with zeros. In: [21], pp. 43–58.
16. Benjamini, Y., Yekutieli, D.: False discovery rate—Adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association 100(469), 71–81 (2005)
17. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 1(57), 289–300 (1995)
18. Aitchison, J., Greenacre, M.: Biplots of compositional data. Journal of the Royal Statistical Society, Series C 51(4), 375–392 (2002)
19. Moore, A.: The so-called 'Flattened Seventh' in rock. Popular Music 14(2), 185–201 (1995)
20. Temperley, D.: The cadential IV in rock. Music Theory Online 17(1) (2011)
21. Pawlowsky-Glahn, V., Buccianti, A. (eds.): Compositional Data Analysis: Theory and Applications. Wiley, Chichester (2011)