Deconstructing the nPVI

A Methodological Critique of the Normalized Pairwise Variability Index as Applied to Music

Nathaniel Condit-Schultz[1]

[1]McGill University, Canada

Abstract

The normalized pairwise variability index (nPVI) is a measure originally used to compare the rhythms of languages. Patel and Daniele (2003a) introduced the nPVI to music research and it has since been used in a number of studies. In this paper, I present a methodological criticism of the nPVI as applied to music. I discuss the known qualitative features of the nPVI, and illustrate the nPVI's fundamental features and assumptions through its application to a number of musical datasets. My principle criticism regards the application of a linear average (the nPVI) to categorical data (rhythmic notation). I argue that that simpler mathematical characterizations, which are more musically intuitive, can capture the same useful information as the nPVI. Specifically, counting the proportion of successive IOIs that are identical accounts for as much as 98% of variation in nPVIs in musical corpora. I argue that abstract mathematical measures ought to be avoided in preference for more concrete empirical descriptions of specific rhythmic features, and that, rather than focusing on a single measure, multiple measures ought to be used. Finally, I conclude that the usage of nPVI in music research should be limited to specific methodologically justified contexts.

Key words: methodology, rhythm, quantification, computer-based musicology, corpus study

The *normalized pairwise variability index* (nPVI) is a measure of durational contrast between successive rhythmic events. Patel and Daniele (2003a) introduced the nPVI to music research, finding that the greater nPVI of spoken English compared to spoken French is roughly paralleled in the nPVIs of instrumental themes by English and French composers. The nPVI has since become widely used in music research, quantifying variation between nations/cultures, eras, and composers (Daniele & Patel, 2015; Daniele, 2016b; Hansen et al., 2016; Hanson, 2017; Huron & Ollen, 2003; McGowan & Levitt, 2011; Patel & Daniele, 2003b; Patel & Danielle, 2013; Sadakata et al., 2004; VanHandel & Song, 2010). Despite this wide usage, little fundamental critical evaluation has been published concerning the nPVI, and significant questions remain regarding the appropriate methodology for nPVI usage and interpretation. Toussaint (2012, p. 2007) first noted this lack of knowledge regarding the nPVI, writing that it "may be a promising and powerful tool in certain contexts. . . [but] the precise nature of these contexts has yet to be determined." This paper attempts to clarify the nature of the nPVI as applied to music, elucidating its strengths, weaknesses, and assumptions through a critical "deconstruction."

The original use of nPVI in music research had a clear theoretical motivation—to search for parallelism between linguistic and musical rhythm. Though some research has continued to leverage the nPVI's cross-domain applicability (McGowan & Levitt, 2011), more studies have applied it to purely musical data (Daniele & Patel, 2015; Daniele, 2016b; Hansen et al., 2016; Hanson, 2017; Huron & Ollen, 2003; Patel & Daniele, 2003a; Patel & Danielle, 2013; Sadakata et al., 2004; VanHandel & Song, 2010).[1] Of course, the original use of the nPVI need not limit its usage: so long as a measure systematically maps "empirical relational structures of interest" to "numerical relational structures that are useful" (Krantz et al., 1971, p. 9) its original intent is irrelevant. Whats more, Patel and Daniele (2003a, p. B37) argue that the difficult-to-interpret, "dimensionless" property of the nPVI is actually ideal for cross-domain analysis. Still, the continued broad usage of the nPVI in purely musical research has proceeded without any clear articulation of what useful, empirical "structure of interest" the nPVI truly represents.

The nPVI was devised to quantify the distinction between *stress-timed* and *syllable-timed* languages (Grabe & Low, 2002; Low et al., 2000). Stress-timing is characterized by semi-regular agogic accents, articulated through the (rough) alternation of long and short inter-onset-intervals (IOIs).[2] In music, such agogic alternation is associated with "swung" rhythms and, more broadly, triple and compound-duple meter (London & Jones, 2011)—often evoking "bouncing" or "lilting" qualia. The nPVI is an appealing quantification of these qualities, as illustrated in Figure 1. However, the nPVI measures *any* durational contrast between successive events, which can result in unintuitive and unpredictable results when applied to diverse musical rhythms. For instance, the left three rhythms in Figure 2 are musically quite similar yet have very different nPVIs, while the right three rhythms are qualitatively quite different yet have have the same nPVI. These observations are pertinent to the interpretation of several published studies: For example, Hanson (2017) reports a difference in nPVI between Western and

Latin musics, suggesting that this reflects differences between composers' native tongues. However, he also notes that Latin rhythms feature "idiomatic rhythms such as syncopation and hemiola" which are not found in Western-style music (Hanson, 2017, p. 482). Thus, it seems possible that the differences in nPVI observed by Hanson might be attributed to differences in syncopation, hemiola, or other *musical* rhythmic features, rather than any *linguistic* rhythm quality. To date, only one study has directly tested listeners' ability to experience the subjective quality of the nPVI: Hannon (2009, pp. 404–406) found that participants could quickly learn to sort melodies differing in nPVI into two groups with approximately 70% accuracy. What rhythmic qualities participants based their decisions on is not clear.

FIGURES 1 AND 2 HERE.

The Formula

Before continuing the discussion, it is appropriate to review the nPVI calculation itself and consider the formula's internal logic. An nPVI is a continuous numeric value falling in the interval $[0, 200)$. Given any ordered series of IOIs, an nPVI can be calculated as,

$$nPVI = \frac{100}{m-1} * \sum_{k=1}^{m-1} \left| \frac{IOI_k - IOI_{k+1}}{(\frac{IOI_k + IOI_{k+1}}{2})} \right| \tag{1}$$

where k indexes the kth IOI and m is the total number of IOIs. With some algebraic rearranging, we can see that the core of the nPVI equation is a simple calculation applied to each adjacent pair of IOIs, which I call the *normalized pairwise calculation* (nPC):

$$nPC = 200 * \left| \frac{\text{antecedent IOI - consequent IOI}}{\text{antecedent IOI + consequent IOI}} \right| \tag{2}$$

These nPCs are simply averaged to get an nPVI. The division by the sum of each pair controls for absolute duration, providing the "normalization" which accounts for changes in overall pace over the course of a rhythm. However, it should be noted that music notation inherently normalizes IOIs to some extent, as changes of tempo are not reflected in duration symbols. Thus, the chief effect of the pairwise normalization of music is that the ratio between IOIs is all that is considered, not their absolute size.[3] For example, the pairs 𝅗𝅥:♩, ♩:♪, and ♪:𝅘𝅥𝅯 all result in the same nPC.

Musicians typically characterize the relationships between rhythmic IOIs as ratios (2/1, 3/1, etc.). Fortunately, the nPC calculation shown in Equation 2 is a monotonic transformation of the ratio between IOIs, where: $nPC = f(ratio) = 200 * |\frac{ratio-1}{ratio+1}|$.[4] This relationship is illustrated in Figure 3. The effect of the absolute-value signs is to negate the ordering of each pair of IOIs, such that reciprocal ratios are considered equivalent—thus, ♩→♪ = ♪→♩.

FIGURE 3 HERE

Datasets

This paper draws upon four musical corpora to explore and illustrate the nature of the nPVI: (1) The European and (2) Chinese components of the Essen database of folk song; (3) The first violin part from a convenient sample of 58 Haydn string quartets; (4) The author's (Condit-Schultz, 2016) corpus of popular rap transcriptions, the Musical Corpus of Flow (MCFlow). All four datasets are encoded in Humdrum syntax; The Essen and Haydn datasets were accessed through the Kern Scores website while MCFlow is available at rapscience.net. To mimic the application of the nPVI in previous research, these corpora can either be compared to each other or broken into various subgroupings. The European and Chinese corpora can be divided into regions (21 European regions, 4 Chinese regions), which can further be divided into individual songs. The Haydn quartets can be divided by opus, or into individual movements. The MCFlow can be divided by year or by song. Figure 4 shows the distribution of nPVIs across various subdivisions of the four corpora. The variation in nPVI evident in Figure 4 is broadly consistent with the results reported in other research. For instance, the scope of variation in nPVI between European regions is comparable to the scope of regional variation observed by Huron and Ollen (2003).

FIGURE 4 HERE

As is evident in Figure 4, variation in nPVIs is far greater within groups than between groups, a pattern which seems to be present in all published musical nPVI studies (Patel & Daniele, 2003a; Raju et al., 2010; Sadakata et al., 2004; VanHandel & Song, 2010), though not all scholars have reported distributional details. Within-language nPVI variability is also far larger than between-language nPVI variability (Loukina et al., 2011; Wiget et al., 2010). As a result, differentiating or classifying rhythms based on nPVI is essentially impossible (Loukina et al., 2011; Vukovics & Shanahan, 2017; Wiget et al., 2010). To illustrate, I attempted to use nPVI to classify European songs by region, using multinomial regression models fit using the R `nnet` package (R Core Team, 2013; Venables & Ripley, 2002). Since the German region is overwhelmingly overrepresented in the Essen collection (5,265 of 6,043 songs), German songs were excluded from this experiment.[5] The single Hungarian song in the sample was also excluded. For the remaining nineteen regions, the model using song nPVI as a predictor was significantly more predictive than a null model with no predictor (Likelihood-ratio test: $\chi^2 = 35.8$, $DF = 18$, $p < .05$) which always predicts the most frequent region (Yugoslavia). However, this significance reflects a small predictive effect size: The nPVI predictor model predicted the European region correctly 16.4% of the time, compared to an accuracy of 14.8% achieved in the null model.[6] Results for predicting the four Chinese regions are similar: 54.2% accuracy with nPVI as predictor; 52.8% without.[7] [8]

The Distribution of nPCs

An nPVI is the arithmetic mean of a set of nPCs. However, though a ubiquitous tool for characterizing the central tendency of numeric data, a mean is not always a meaningful value. Means are informative when summarizing unimodally and continuously distributed numbers, particularly when they are normally distributed, as is often assumed. None of these conditions are true of the rhythms found within a musical score, which are drawn from a small set of integer-related IOIs. Thus, though nPCs are in principle continuous, when applied to symbolic music notation—as most studies have[9]—the practical reality is a categorical distribution, with nearly all IOI pairs forming the ratios $1/1$, $2/1$, $3/1$, or $4/1$. To illustrate, the rhythm ♩ ♫ ♩ ♫ consists of the pairwise ratios $\{1/2, 1/1, 1/2, 1/2, 1/1\}$, averaging a ratio of $7/10$. However, the ratio of $7/10$ never actually occurs in the passage, and is thus not descriptive of the rhythm's central tendency. Figure 5 shows a histogram of nPCs (all IOI pairs) within each corpus. The mean of each distribution (i.e. the nPVI) is marked below each histogram as a cross-hair symbol. Figure 6 shows nPC histograms for four individual songs drawn from the European corpus. These four songs represent a range of nPVIs within the European dataset, specifically the 20%, 40%, 60%, and 80% nPVI quantiles of the corpus. (In other words, the first song's nPVI is greater than only one out of five European songs', while the last song's nPVI is greater than four out of five.) Categorical distributions like those evident in Figures 5 and 6 are not effectively described by their mean. Contrast these with Figure 7, which shows the distribution of nPCs in a corpus of linguistic data[10]; as can be seen, a truly continuous distribution of values *is* evident in language, making the mean a more meaningful descriptor of the distribution's center of mass.[11]

FIGURES 5, 6, and 7 HERE

How might we better characterize distributions like those shown in Figures 5 and 6? Jian (2004) proposed using the median nPC rather than the mean for linguistic data. (Figure 7 includes the median and mode[12] of the linguistic nPCs, as an `x` and an `o` respectively.) However, the median of the musical nPC-distributions shown in Figure 5 and the first two songs in Figure 6 are all zero, as in all cases more than half of the pairs form a ratio of $1/1$. The medians of the remaining two songs are $66.6\overline{6}$, and the modes of all eight distributions are the same as their respective medians. Thus, neither the mode nor median is as sensitive as the mean in detecting changes in categorical distributions of nPCs: Though the mean (e.g. the nPVI) doesn't correspond to typical pairwise ratios in a musical passage, it nonetheless reflects a balance between two or three modal "poles" in the distribution, providing more information than the median or mode alone.

Isochrony

One striking feature of Figures 5 and 6 is the concentration of isochronous ($ratio = 1/1$; $nPC = 0$) IOI pairs. This reflects the highly regular, periodic nature of musical rhythm. In fact, it appears that much of the information in these distributions is simply captured by the proportion of isochronous pairs—an observation first articulated by Raju, Asu, et al. (2010, p. 64). To test this observation, a simple linear regression model was created to predict the nPVI of each song in each of the four corpora using the *isochrony proportion* (IsoP) as a predictor.[13] I calculate the IsoP by iterating over every pair of successive IOIs in a rhythm, counting the pairs which are identical, and dividing this count by the total number of pairs (one less than the total number of IOIs). As can be seen in Table 1, 86–92% of variance in nPVI is accounted for by the IsoP. Of course, nPVIs *do* reflect more than IsoP: If the proportion of $2/1$ pairs[14] is added as a second predictor to each regression model, the models' performances are improved substantially, as reported in Table 2. This illustrates that the nPVI largely reflects a combination of IsoP *and* $2/1$ proportions, with other (rarer) pairwise ratios only exerting some small residual influence ($< 5\%$ of variance) on the final value.

Another approach would be to calculate nPVIs *excluding* specific nPC values—for instance, excluding isochronous pairs. By "factoring out" isochrony we get a new measure (the *pairwise anisochronous contrast index*) which is sensitive to changes in the frequencies of $2/1$, $3/1$, or other pairs, without being overwhelmed by isochrony. Unfortunately, the pACI is still extremely variable within groups in my corpora; applying my multinomial region classification model (described above), the pACI performs no better than the nPVI when predicting European regions (15.3% accuracy). Alternatively, we might characterize nPC distributions using Shannon entropy, a convenient measure of the "complexity" of a categorical distribution. Interestingly, this *normalized pairwise entropy index* (nPEI) performs slightly better as a predictor of European regions than the nPVI itself (accuracy = 18.3%).

Van Handel (2010) suggests that duration pairs straddling phrase boundaries ought to be excluded when calculating the nPVI, resulting in what she calls the *phrase-nPVI* (pnPVI).[15] Figure 8 shows the distribution of pnPVIs in three of the four corpora (the Haydn dataset had to be excluded because it contains no phrasing information). If we compare Figure 8 to Figure 4, we can see that pnPVIs scores are generally lower than nPVIs. This illustrates exactly why Van Handel suggested the pnPVI: IOI ratios at phrase boundaries are generally much longer and more varied than ratios within phrases, inflating the nPVI if these boundaries are included. Results of new regression analyses with pnPVIs predicted by phrase-IsoP (excluding pairs which straddle phrase boundaries from the IsoP calculation) are reported in the bottom halves of Tables 1 and 2. As can be seen, if attention is restricted to intra-phrase rhythmic consideration, the nPVI and the IsoP are even more highly correlated.

FIGURE 8 HERE

Reducing complex, multi-dimensional distributions like those shown in Figures 5 and 6 to a single descriptive statistic is inevitably reductive. Thus, though one-dimensional measures (like the nPVI or IsoP) are convenient for statistical comparisons and visualizations, when ever possible it is preferable to consider more complex descriptions of data. For instance, it may be more fruitful to compare and contrast complete nPC distributions, which contain much more information about pairwise IOI relationships. As an example, we can consider the differences between French and English nPC distributions: the proportion of $^1/_1$ pairs in French and English songs are 41.5% and 38.3% respectively—a fairly minor difference. However, French songs in the Essen corpus contain approximately 63% more $^3/_1$ ratios than English songs. Indeed, the proportion of $^3/_1$ ratios does function as a better categorizer of European regions than the nPVI; $^3/_1$ proportions predict the European region more accurately (19.3%) than IsoP or the nPVI.[16] Only by studying the complete distribution of pairwise ratios, can more precise observations such as this be made. As a compromise between a single index value and the complete nPC distribution, we might report a 2–4 dimensional "pairwise IOI profile." For instance, we could present the proportion of $^1/_1$, $^2/_1$, or $^3/_1$ ratios in the data, which account for the vast majority of pairs. Indeed, using main effects for and interactions between $^1/_1$, $^2/_1$, and $^3/_1$ proportions, European regions can be predicted with 22.0% accuracy.[17]

Micro-timing

As we've seen, my major concern with the nPVI is its application to notation-like, quantized IOI data. Even given these concerns, we might still expect the nPVI to be useful when applied to non-categorical rhythmic data measured from human performances (London & Jones, 2011, p. 120). To date, only McGowan and Levitt (2011) have made use of actual performance timing data in an nPVI study. Fortunately, Raju, et al. (2010) conducted a study specifically to compare nPVIs derived from notation to nPVIs derived from human performance timings. They found that performed nPVIs were generally higher than score-based nPVIs, though on closer inspection only three out of twelve songs evinced this difference. This suggests that using scores or performances may result in similar nPVIs in many instances (Raju et al., 2010, p. 63).

To compare nPC distributions of human performances with those of music notation, I draw on the MARG (Heo et al., 2013) and EEP (Marchini et al., 2014) datasets. The MARG dataset contains detailed timing data for the sung performances of three folk tunes by twenty adult singers, serving as an excellent comparison point for the Essen corpora, as the three tunes are identical or similar to tunes which appear in Essen.[18] The EEP dataset contains detailed performance information for a professional performance of segments of Beethoven's fourth String Quartet (Opus 18, No. 4)—to be comparable to the Haydn data, I restrict my analysis to the first violin part. These datasets are not as large, nor structured in the same manner, as the notation-based corpora, but are the best available to me. Figure 9 shows the distribution of nPC values in each corpus—each figure shows the nPC distribution of the notated score in thicker, lighter colored bars, and the nPC distribution of the performance data in thinner, darker colored bars. The nPVIs of the performance data are marked by cross-hairs below each plot—individual dots indicate the nPVI of individual performers in the MARG data—, while the nPVIs of the notated scores are marked by cross-hairs above each plot. Consistent with the observations of Raju, et al. (2010), the performed nPVIs are all slightly higher than the notated nPVIs. As expected, the nPC distributions of the performance data are continuous. However, the performed nPCs cluster around the categorical nPCs seen in the notation, especially in the MARG data. Despite the smoother distribution of values, the global average of these distributions (the nPVI) is still not a very useful summary, as each distribution is clearly multimodal.

FIGURE 9 HERE

The Distribution of nPVIs

Having discussed in detail the distribution of nPCs in real musical data, it is pertinent to briefly discuss the mathematical properties of the nPVI itself. Many papers (Hanson, 2017; Huron & Ollen, 2003; Patel & Daniele, 2003a; Patel et al., 2006; Sadakata et al., 2004) have used the non-parametric Mann-Whitney U-test to compare nPVIs between groups, presumably because authors have been (appropriately) concerned that the nPVI may not be normally distributed. In other cases, scholars have used parametric, normal-distribution assumptions without reservation (Daniele & Patel, 2015; Daniele, 2016b; Hansen et al., 2016; London & Jones, 2011; McGowan & Levitt, 2011; Patel & Daniele, 2003a; Patel et al., 2006; Patel & Danielle, 2013; Raju et al., 2010; VanHandel & Song, 2010; VanHandel, 2016), especially when interested in more complex statistical relationships like ANOVA or linear regression. Technically, nPVI cannot be normally distributed because it is bounded in the range $[0, 200)$. Whats more, it is not clear how linear the nPVI really is—is the nPVI a ratio-, interval-, or ordinal-level scale?[19] Still, statistical tests which "technically" violate normality assumptions are frequently reported (for instance, ANOVA on Likert scales or proportions) as much research suggests that these tests are robust to these violations (Norman, 2010). Indeed, averages of non-normal distributions (like the nPVI) are often themselves distributed normally. In my datasets, the distribution of nPVI residuals is close to normal, though with a slight positive skew (Figure 10).[20] Thus, though treating the nPVI with statistical tests that assume normal distributions is possibly problematic, it is within the norms of statistical reporting.

FIGURE 10 HERE

Proceeding with the assumption that parametric models are acceptable, we can note a more serious (though also commonplace) violation of statistical assumptions: the assumption of independence. Published statistical analyses of nPVI data have generally failed to address major sources of dependence in data. For example, in Patel's and Daniele's original nPVI study (2003a, pp. B41–42), their Mann-

Whitney test makes no allowance for variation between composers, despite the fact that large variation between composers is evident in their data. Given the large variations they report between composers, it is entirely plausible that a different random sample of composers would have resulted in difference results. To illustrate using my own data, a simple one-way ANOVA on my four corpora is significant (F test: $F = 9.0$, $DF = (3, 6386)$, $p < .05$), indicating that the nPVI differs significantly between the four corpora. However, if random variation between sub-groups (regions, opuses, etc.) is taken into account—specifying them as random intercepts in a mixed-effects model—the resulting model is not significant (Likelihood-ratio test: $\chi^2 = 7.3$, $DF = 3$, $p > .05$). This analysis should not be taken as definitive—there are more statistical and methodological issues to consider—but illustrates the importance of data dependence issues in nPVI, especially given the repeated observation of large sub-group variation in nPVI values.

Most statistical measures are underpinned by principled conceptual frameworks and probabilistic "assumptions": for instance, Shannon entropy is grounded in information theory. Nonetheless, these same measures are frequently used as convenient heuristics, even when their original conceptual intentions are not valid. The nPVI may too serve as just such a useful heuristic measure of rhythmic style, and many scholars have (implicitly) treated it this way. For instance, though the word "variability" in the nPVI is actually a misnomer (Patel et al., 2006, p. 3035), scholars have often treated the nPVI as a measure of "durational variability" in general (VanHandel & Song, 2010, p. 1). This interpretation is not unreasonable: Patel, Iversen, et al. (2006) found that the *coefficient of variation* (CV) does correlate with nPVI.[21] In my own datasets, the correlation between CV and nPVI is close to the lower boundary observed by Patel et al. ($r = .37$, $p < .05$). Toussaint (2012) investigated the correlation between nPVI and a number of objective and subjective characterizations of rhythmic "complexity," finding that the nPVI performs poorly as a predictor of the subjective complexity of rhythms, but does correlate with some mathematical measures of complexity (Toussaint, 2012, p. 1007). Indeed, Shannon entropy—widely used as a convenient proxy for complexity Cox (2010); Margulis & Beatty (2008)—correlates fairly well with the nPVI in my data ($r = .72$, $p < .05$). Still, unlike entropy or the CV, little work has been done to suggest that the nPVI *is* a particularly useful heuristic, especially when compared to alternative measures.

Conclusion

Empirical musicologists are faced with the difficult task of objectively characterizing and quantifying the plethora of rhythmic features and qualities that appear in music. Many approaches have been defined, each with their own implicit assumptions and biases and each reflecting different facets of rhythmic quality. The nPVI is but one approach to quantifying rhythmic quality, though the recent literature seems to treat it as *the* measure of rhythmic style. For instance, Daniele (2016a) proposes the intriguing prospect of an empirical "rhythmic fingerprint" to describe the rhythmic practices of different composers, but bases his fingerprint entirely on one feature: the nPVI. Such overreliance on the nPVI limits research to single set of methodological assumptions: pairwise, normalized, unordered, etc. None of these assumptions are bad—for instance, pairwise analyses have been fruitful in many areas of musical inquiry (Arthur, 2017; de Clercq & Temperley, 2011; Condit-Schultz, 2016)—yet they offer us only one perspective. In linguistics, several studies have reported the danger of relying solely on the nPVI, advocating the use of multiple rhythmic measures in any study (Loukina et al., 2011; Wiget et al., 2010).

It is up to the scholarly community to critically evaluate all quantitative measures, both in statistical/ mathematical and *musicological* terms. In order to facilitate mathematical evaluation, it is essential that the assumptions underpinning all quantitative measures, and the nature of the data being studied, are explicitly articulated. Indeed, the principle weakness in published descriptions of the nPVI has been the failure to recognize the fundamental differences between musical rhythm data and linguistic rhythm data. It seems that the nPVI may be a useful proxy for rhythmic variance and complexity—but if a measure is used only as a convenient, heuristic, this should always be made clear. In order to facilitate musicological evaluation, computational measures should be related to theoretical characterizations. The nPVI *may* constitute a useful measure of some rhythmic qualities (perhaps "swing" or "lilt"), but these qualities have yet to be established through behavioral psychology research. In contrast, consider Huron and Ommen (2006) or Temperley and Temperley (2011), which utilize simple, transparent, and clearly articulated quantifications of concrete rhythmic features (syncopation and the "Scotch snap" respectively). Taking a similar tack, we might define concrete definitions of rhythmic qualities of interest: we might define "lilt" as an event which is shorter than the previous event *and* the subsequent event. This definition of lilt correlates fairly highly with the nPVI (between $r = .63$ and $r = .82$ in my four corpora), but further research is required to determine if it is an effective measure of the subjective quality of lilt. Fortunately, the most concrete conclusion of this paper is that the nPVI can effectively be exchanged with the more intuitive *isochrony proportion* in many cases. This alternative measure captures most of the same information as the nPVI, but is more methodologically transparent, and easier to intuit.

Many fine studies have been conducted using the nPVI, and there is no reason to think that any flaws in the nPVI undermine their basic conclusions. Indeed significant (in the statistical sense) categorical differences and linear/curvilinear trends in nPVI value have been consistently observed in a number of datasets, suggesting that nPVI is a measure of *something*. However, studies have consistently found that nPVI effect sizes are quite small, with observed variation within groups consistently overwhelming variation between groups. Inversely, these small effect sizes make the nPVI a poor predictor itself: my attempts to train categorical models to use the nPVI to predict a songs' regions found only tiny increases above chance performance. These results are consistent with findings in other linguistic (Loukina et al., 2011; Wiget et al., 2010) and musical (Vukovics & Shanahan, 2017) research.

Though I've offered substantive criticism of the nPVI as applied to musical data, I acknowledge that it may indeed be an effective measure in some situations—the cross-domain comparison of language and music, for instance. Another area where the nPVI might be useful is in the study of performance timing data, especially when the performance practice eschews or blurs rhythm categories. For example, nPVI might be used as a descriptor of the degree of jazz swing, which has been shown to vary continuously without respecting neat rational relationships (Honing & De Haas, 2008).

By no means is the nPVI the only quantitative measure to evade thorough interrogation: It is all too common that complex mathematical functions are treated as "black boxes" without clear qualitative correlates. This paper is intended not just as a critique of the nPVI, but as a case study in quantitative methodological critique. All abstract mathematical quantifiers—including the coefficient of variation and entropy—ought to be regarded with suspicion, especially when used as convenient heuristics outside of their original conceptual framework. For instance, entropy cannot be taken too literally as a measure of information content in music if we only calculate it based on the first-order conditional distributions of a few isolated musical parameters (Krumhansl, 2015; Margulis & Beatty, 2008). My main concern is not with failings of the nPVI, but that important methodological issues regarding the nPVI—e.g. that it is a linear average of discrete categories—and qualitative features—that the nPVI is highly correlated with repeated IOIs—have not been explicitly acknowledged. Readers may not recognize potential issues, or assumptions, of these functions unless they are clearly explained. Likewise, readers cannot form coherent critical interpretations of research if important methodological assumptions of that research are not communicated. It is up to researchers to explicitly articulate why the empirical measure they choose is an appropriate tool for the task at hand, just as Patel and Daniele (2003a) do in their original paper.

REFERENCES

Arthur, C. (2017). Taking harmony into account. *Music Perception: An Interdisciplinary Journal*, *34*(4), 405–423.

Condit-Schultz, N. (2016). *MCFlow: A Digital Corpus of Rap Flow*. Ph.D. thesis, Ohio State University.

Cox, G. (2010). On the Relationship Between Entropy and Meaning in Music: An Exploration with Recurrent Neural Networks. *Proceedings of the 32nd Annual Cognitive Science Society. Austin TX: CSS*.

Daniele, J. R. (2016a). A Tool for the Quantitative Anthropology of Music: Use of the nPVI Equation to Analyze Rhythmic Variability within Long-term Historical Patterns in Music. *Empirical Musicology Review*, *11*(2), 228–233.

Daniele, J. R. (2016b). The "Rhythmic Fingerprint": An Extension of the nPVI to Quantify Rhythmic Influence. *Empirical Musicology Review*, *11*(2), 243–260.

Daniele, J. R., & Patel, A. D. (2015). Stability and Change in Rhythmic Patterning Across a Composers Lifetime: A Study of Four Famous Composers Using the nPVI Equation. *Music Perception: An Interdisciplinary Journal*, *33*(2), 255–265.
URL `http://www.jstor.org/stable/10.1525/mp.2015.33.2.255`

de Clercq, T., & Temperley, D. (2011). A Corpus Analysis of Rock Harmony. *Popular Music*, *30*(01), 47–70.

Dellwo, V., Lemman, A., & Kolly, M. J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. In *Proceedings of Interspeech*. Portland OR, USA.

Grabe, E., & Low, E. L. (2002). *Durational variability in speech and the rhythm class hypothesis*, (pp. 515–546). Berlin: Mouton de Gruyter.

Hannon, E. E. (2009). Perceiving speech rhythm in music: listeners classify instrumental songs according to language of origin. *Cognition*, *111*(3), 404–410.

Hansen, N. C., Sadakata, M., & Pearce, M. (2016). Nonlinear Changes in the Rhythm of European Art Musc: Quantitative Support for Historical Musicology. *Music Perception*, *33*(4), 414–431.

Hanson, J. (2017). Rhythmic Variability in Language and Music of Latino and Latino-Inspired Composers. *Music Perception*, *34*(4), 482–488.

Heo, H., Sung, D., & Lee, K. (2013). Note onset detection based on harmonic cepstrum regularity. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*.

Honing, H., & De Haas, W. B. (2008). Swing Once More: Relating Timing and Tempo in Expert Jazz Drumming. *Music Perception*, *25*(5), 471–476.

Huron, D. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press.

Huron, D., & Ollen, J. E. (2003). Agogic Contrast in French and English Themes: Further Support for Patel and Daniele (2003). *Music Perception*, *21*(2), 267–271.

Jian, H.-L. (2004). An improved pair-wise variability index for comparing the timing characteristics of speech. *Proceedings of the 8th International Conference on Spoken Language Processing*, (pp. 4–8).

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement Vol 1: Additive and Polynomial Representations*. San Diego and London: Academic Press.

Krumhansl, C. L. (2015). Statistics, structure, and style in music. *Music Perception: An Interdisciplinary Journal*, *33*(1), 20–31.
URL http://mp.ucpress.edu/content/33/1/20

London, J., & Jones, K. (2011). Rhythmic Refinements to the nPVI Measure: A Reanalysis of Patel & Daniele (2003a). *Music Perception*, *29*(1), 115–120.

Loukina, A., Kochanski, G., Rosner, B., Keane, E., & Shih, C. (2011). Rhythm measures and dimensions of durational variation in speech. *Journal of the Acoustical Society of America*, *129*(5), 3258–3270.

Low, E. L., Grabe, E., & F., N. (2000). Quantitative characterisations of speech rhythm: 'Syllabletiming in Singapore English. *Language and Speech*, *43*(4), 377–401.

Marchini, M., Ramirez, R., Papiotis, P., & Maestre., E. (2014). The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *Journal of New Music Research*, *43*(3), 303–317.

Margulis, E. H., & Beatty, A. P. (2008). Musical style, psychoaesthetics, and prospects for entropy as an analytic tool. *Computer Music Journal*, *32*(4), 64–78.

McGowan, R. W., & Levitt, A. G. (2011). A Comparison of Rhythm in English Dialects and Music. *Music Perception*, *28*(3), 307–314.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632.

Patel, A. D., & Daniele, J. (2003a). An empirical comparison of rhythm in language and music. *Cognition*, *87*, B35–B45.

Patel, A. D., & Daniele, J. R. (2003b). Stress-Timed vs. Syllable-Timed Music? A Comment on Huron and Ollen (2003). *Music Perception: An Interdisciplinary Journal*, *21*(2), 273–276.

Patel, A. D., & Danielle, J. R. (2013). An empirical study of historical patterns in musical rhythm: Analysis of german & italian classical music using the npvi equation. *Music Perception*, *31*(1), 10–18. Document feature - Graphs; ; Musical Scores; Tables; Last updated - 2014-08-04.
URL http://search.proquest.com/docview/1487696324?accountid=9783

Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of british english and french. *Journal of the Acoustical Society of America*, *119*(5), 3034–3037.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/

Raju, M., Asu, E. L., & Ross, J. (2010). Comparison of rhythm in musical scores and performances as measured with the Pairwise Variability Index. *Musicæ Scientiæ*, *14*(1), 51–71.

Sadakata, M., Desain, P., Honing, H., Patel, A. D., & Iversen, J. R. (2004). A cross-cultural study of the rhythm in english and japanese popular music. In *Proceedings of the International Symposium on Musical Acoustics*.

Temperley, N., & Temperley, D. (2011). Music-Language Correlations and the "Scotch Snap". *Music Perception*, *29*(1), 51–63.

Toussaint, G. T. (2012). The pairwise variability index as a tool in musical rhythm analysis. In *Proceedings of the 12th International Conference on Music Perception and Cognition*, (pp. 1001–1008).

VanHandel, L. (2016). The War of the Romantics: An Alternate Hypothesis Using nPVI for the Quantitative Anthropology of Music. *Empirical Musicology Review*, *11*(2), 235–242.

VanHandel, L., & Song, T. (2010). The role of meter in compositional style in 19th century french and german art song. *Journal of New Music Research*, *39*(1), 1–11.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer, fourth ed. ISBN 0-387-95457-0.
URL `http://www.stats.ox.ac.uk/pub/MASS4`

Vukovics, K., & Shanahan, D. (2017). Using nPVI to Examine the Role of National Musical Styles in the Works of Non-Native Composers. San Diego, CA: Society for Music Perception and Cognition, National Conference.

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America*, *127*(3), 1559–1569.

Nathaniel Condit-Schultz
1685 Avenue Bennett
Montréal, Quebec, H1V 2S7
514-267-3823
nathaniel.condit-schultz@mcgill.ca

Notes

[1]To be sure, most of these studies have assumed that, following Patel and Daniele's original results, that musical nPVI correlates with linguistic nPVI. However, few studies have actually applied the nPVI to both musical and linguistic data.

[2]I will follow the common methodological approach of using IOIs rather than durations. This avoids the messy complexity of considering rhythmic onsets *and* offsets. The principle difference between durations and inter-onset-intervals is that the former does not include rests (silence) between events.

[3]London and Jones (2011, p. 120) speculate about the effect of removing this normalization. The resulting PVI measure, which has been used extensively in linguists, represents the absolute magnitude of differences between durations.

[4]For example, the ratio between 𝅗𝅥 and ♩ is 2/1. Therefore, $(\frac{2}{1}) = 200 * |\frac{2-1}{2+1}| = 200 * \frac{1}{3} = 66.6\overline{6} = nPC$.

[5]When German songs are included, the model simply learned to classify every input as German, achieving an 86% accuracy.

[6]The nPVI predictor model gains this small improvement by guessing that higher nPVIs indicate that a song is Dutch, rather than Yugoslav.

[7]The multinomial model is hampered by its need to predict all categories; regions with more extreme nPVI values might be effectively distinguished from each other in more focused tasks. Indeed, training binary (logistic) regression models for each pair of regions revealed that Dutch songs ($nPVI \sim 44$) could be distinguished from Yugoslav, Polish and Russian songs ($nPVI \sim 35$–38) as much as three times as accurately as a null model. However, these findings are entirely *post-hoc*, representing four successes out of a total of 210 pairwise comparisons. Also, note that these models were tested on the training data itself; More rigorous modeling methodology would train and test different subsets of the data, which would certainly reduce model performance—overfitting is likely.

[8]Relative to random guessing, the success rate of Hannon's (2009) participants (∼70%) represents an increase in the odds of successful classification (French or English) of approximately 140%. How might Hannon's participants have succeeded where computational models have failed? First, Hannon's participants may have based their judgements on other rhythmic qualities which correlate with nPVI—in contrast to statistical models which receive *only* the nPVI as input. Second, Hannon represented each group (English and French) using melodies with nPVIs "close" to values of ∼31 or ∼43, but did not precisely describe their spread around these mean values. It is possible that the nPVIs of songs in Hannon's groups didn't overlap dramatically, as they tend to do in other corpora (Figure 4).

[9]Patel and Daniele motivate their use of notated values by arguing that notation represents the only "unambiguous record of [common-practice] composers's choice of relative durations" (Patel & Daniele, 2003a, p. B40).

[10]From the TEVOID dataset (Dellwo et al., 2012), a corpus of 50 Swiss German speakers speaking 256 sentences each.

[11]Of course, the distribution is still not normal, as it is radically skewed and bounded on the left.

[12]The mode of the distribution was estimated using R's built in `density` function.

[13]This approach is similar to the procedure adopted by Patel, Iversen, et al. (2006) when comparing the nPVI to the coefficient of variation.

[14]To calculate this value: for each successive IOI pair, divide the consequent by the antecedent and ask if the result is either 2 or 2/1. Count these matches and divide by the total number of pairs.

[15]London and Jones (2011, p. 118) make a similar suggestion, though they advocate normalizing boundary-straddling IOIs to the tactus, rather than excluding them.

[16]IsoP predicts European regions with comparable accuracy to the nPVI (16.7%), and the 2/1 proportion performs no better.

[17]All of these categorical prediction models should be regarded as somewhat informal, as the differences in sample sizes between different regions (even if we exclude Germany and Hungary) are not ideal for this type of task.

[18] One of the tunes is the ubiquitous "Twinkle, Twinkle, Little Star" (originally "Ah! vous dirai-je, maman"). The other two tunes are of Korean origin, though "the Butterfly" is essentially identical to the German tune "Hänschen klein."

[19] In nPVI, does $\frac{40}{20} = \frac{100}{50}$? Or does $(40 - 20) = (140 - 120)$?

[20] This skew arises because nPVIs below group means are frequently constrained by the measure's lower bound (0), while no values ever approach the upper bound (200).

[21] However, the predictive relationship between the CV and the nPVI is somewhat weak (r between .37–.60), and they conclude that nPVI is distinct from rhythmic variability (Patel et al., 2006, pp. 3039–3041).

Table 1: Results of linear regressions predicting nPVI and pnPVI from IsoP and pIsoP. Each model's adjusted-R^2 is reported, which is commonly interpreted as the "proportion of variance" accounted for by the predictor. The residual σ is the standard deviation of the models' errors. The prediction quantiles 25%–75% indicate the range in which the middle 50% of errors occur. In other words, half of the first model's predictions miss the true nPVI by between -2.97 and 2.10.

		Adjusted R^2	Residual σ	Prediction 25%–75% Quantiles
nPVI				
	Europe	.91	4.74	-2.97–2.10
	China	.86	4.45	-2.97–2.20
	Haydn	.86	3.39	-1.83–1.24
	Rap	.92	2.27	-1.25–1.07
pnPVI				
	Europe	.95	3.88	-2.19–1.35
	China	.89	4.20	-2.55–1.76
	Rap	.98	1.09	-0.44–0.36

Table 2: Results of linear regressions predicting (p)nPVI from (p)IsoP *and* 2/1 pairs. Each model's adjusted-R^2 is reported, which is commonly interpreted as the "proportion of variance" accounted for by the predictor. The residual σ is the standard deviation of the models' errors. The prediction quantiles 25%–75% indicate the range in which the middle 50% of errors occur. In other words, half of the first model's predictions miss the true nPVI between by between -1.57 and 1.51.

		Adjusted R^2	Residual σ	Prediction 25%–75% Quantiles
nPVI				
	Europe	.96	3.10	-1.57–1.51
	China	.94	3.03	-1.75–1.55
	Haydn	.95	2.00	-1.17–0.56
	Rap	.94	2.01	-0.84–0.88
pnPVI				
	Europe	.97	2.67	-1.04–1.05
	China	.95	2.82	-1.33–1.19
	Rap	.98	1.11	-0.44–0.36

Figure 1: Illustration of rhythmic patterns with different degrees of agogic alternation ("lilt" or "swing"), and their corresponding nPVIs.

Figure 2: Illustration of unintuitive variation (or lack of) in the nPVI. The left examples show significant variation in nPVIs between three musical rhythms which feature agogic contrasts that are much more complicated then those found in language. The right examples show three rhythms with vastly different patterns yet identical nPVIs.

Figure 3: Relationship between IOI ratios and nPC. The ratios should be interpreted as unordered, meaning that $1/2$ receives the same nPC as its reciprocal $2/1$. The absolute size of IOIs is irrelevant, so any IOI pair shown here would receive the same nPC in augmentation or diminution.

Figure 4: Distribution of nPVI scores across four corpora. The overall average of each corpus is represented by the height of each bar. In the European and Chinese corpora cross-hairs indicate regions (sorted from lowest to highest nPVI), while dots indicate songs—the one extremely dense region represents German songs. In the Haydn corpus, cross-hairs indicate opuses (sorted from lowest to highest nPVIs) and dots indicate individual movements. In the Rap corpus, cross-hairs indicate years (in order from 1980 to 2014, skipping 1983 and 1984), and dots indicate individual verses.

Figure 5: Histograms of nPCs in each of four corpora. The X axis indicates nPCs and equivalent pairwise ratios. The height of each bar indicates the proportion of pairs in the corpus which form the ratio (or nPC) represented by that position on the X axis. The cross-hair symbol below each histogram marks the average of the values (i.e. the nPVI).

Figure 6: Histograms of nPCs in four songs drawn from the European corpus. The fours songs were selected based on the position of their nPVI within the distribution of nPVIs in the European corpus. The top left histogram describes a song with a relatively low nPVI, greater than only 20% of European songs. At the other extreme, the bottom right graph plots a song with a relatively high nPVI, greater than 80% of European songs. The X axis indicates nPCs and equivalent pairwise ratios. The height of each bar indicates the proportion of pairs in the corpus which form the ratio (or nPC) represented by that position on the X axis. The cross-hair symbol below each histogram marks the average of the values (i.e. the nPVI).

Figure 7: Histograms of nPCs in the TEVOID corpus of spoken Swiss German. The X axis indicates nPCs and equivalent pairwise ratios. The height of each bar indicates the proportion of pairs in the corpus which form the ratio (or nPC) represented by that position on the X axis. The cross-hair (`+`) symbol below the histogram marks the average of the values (i.e. the nPVI), while the ex (`x`) and circle (`o`) represent the median and mode respectively.

Figure 8: Distribution of pnPVI scores across three corpora—the Haydn corpus is excluded because it lacks phrasing information. The overall average of each corpus is represented by the height of each bar. For each corpora, the distribution of larger subgroups are indicated by cross hair symbols while smaller subgroups are plotted as dots, randomly "jittered" across the X axis so that individual points are visible. In the European and Chinese corpora cross-hairs indicate regions (sorted from lowest to highest pnPVI), while dots indicate songs—the one extremely dense region represents German songs. In the Rap corpus, cross-hairs indicate years (in order from 1980 to 2014, skipping 1983 and 1984), and dots indicate individual verses.

Figure 9: Distribution of nPCs in each song in MARG corpus, and the EEP corpus. The X axis indicates nPCs and equivalent pairwise ratios. The height of each bar indicates the proportion of pairs in the corpus which form the ratio (or nPC) represented by that position on the X axis. Thinner darker bars indicate the distribution of nPCs derived from human performance data, while the wider lighter bars indicate the distribution of nPCs in the music notation data. The darker cross-hair symbol below each histogram marks the average (i.e. the nPVI) of the performance-derived distribution (in the three MARG plots, additional dots indicate the nPVI of individual singers in the data). The lighter cross-hair above the bulk of each plot indicates the average of the notation-derived distribution. (The Y axis includes a separate proportion scale for the notation-derived (larger, normal font) and performance-derived (smaller, italic font) distributions. The absolute height of bars in the performance-derived distribution is much lower because there are far more bins.)

Figure 10: Distribution of nPVI residuals in the four corpora. At the top of the figure, each individual dot represents the nPVI residual of a single song, movement, or rap verse from the dataset (8,166 in total), randomly "jittered" across the Y axis so that individual points are visible. This figure is like Figure 4 turned on its side, and with each dot centered relative to the group's mean (one of 74 cross-hairs in Figure 4—$DF = 74$). The grey histogram is a different representation of the scatter dots, with all dots counted within bins of width four. As can be seen, the distribution evinces a positive skew. The dashed line overlaid on the histogram shows the distribution of nPVI residuals with respect to the single grand mean of the whole dataset ($DF = 1$). This distribution (using only a single degree of freedom) is *slightly* wider than the residuals from smaller group means—illustrating again that nPVI variation across groups is tiny compared to nPVI variation within groups.

nPVI = 0 nPVI = 40 nPVI = $66.6\overline{6}$ nPVI = 100 nPVI = 150

Figure 1

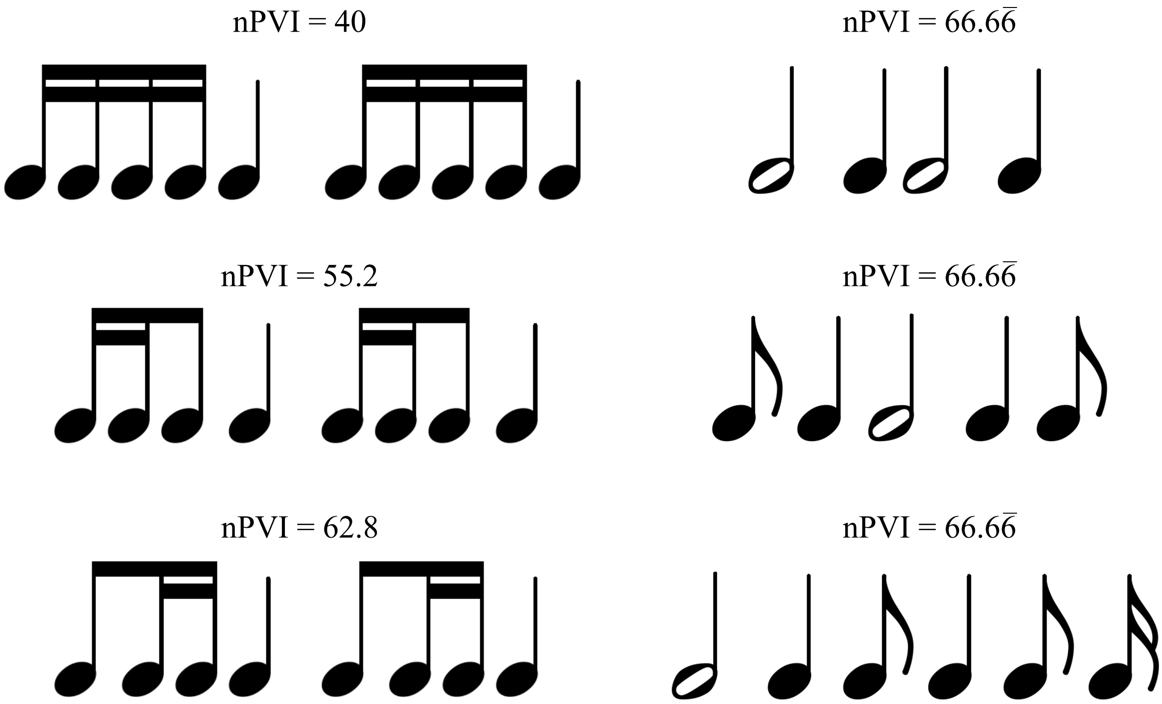
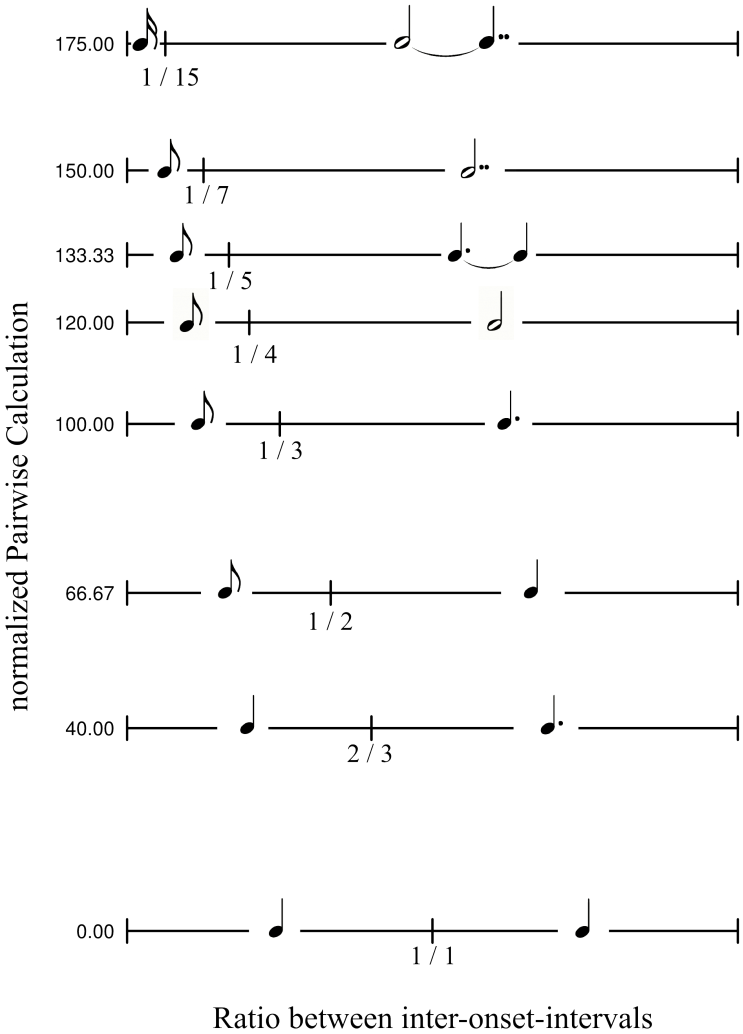

nPVI = 40
nPVI = $66.6\overline{6}$
nPVI = 55.2
nPVI = $66.6\overline{6}$
nPVI = 62.8
nPVI = $66.6\overline{6}$

Figure 2



466
467 **Figure 3**

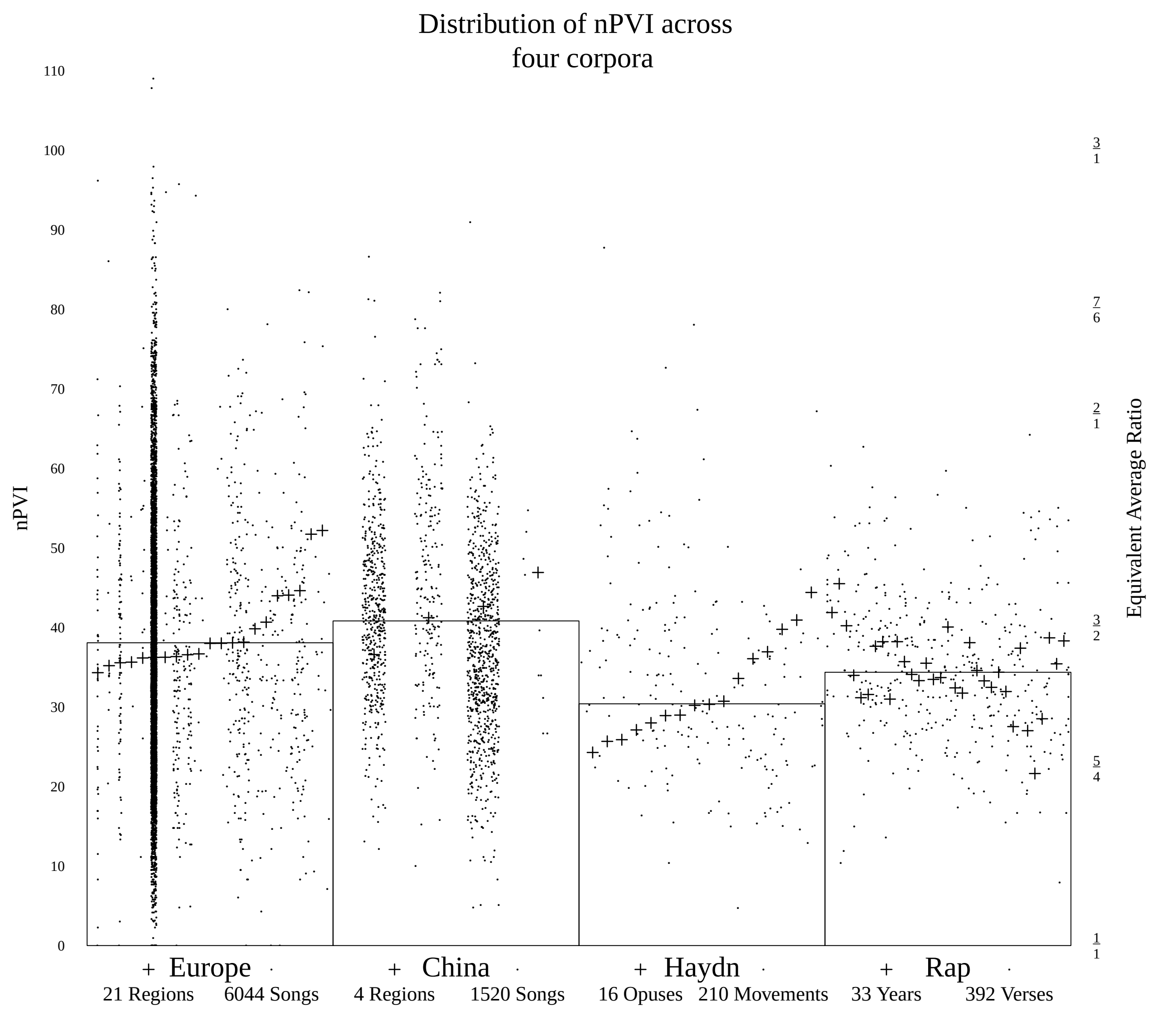
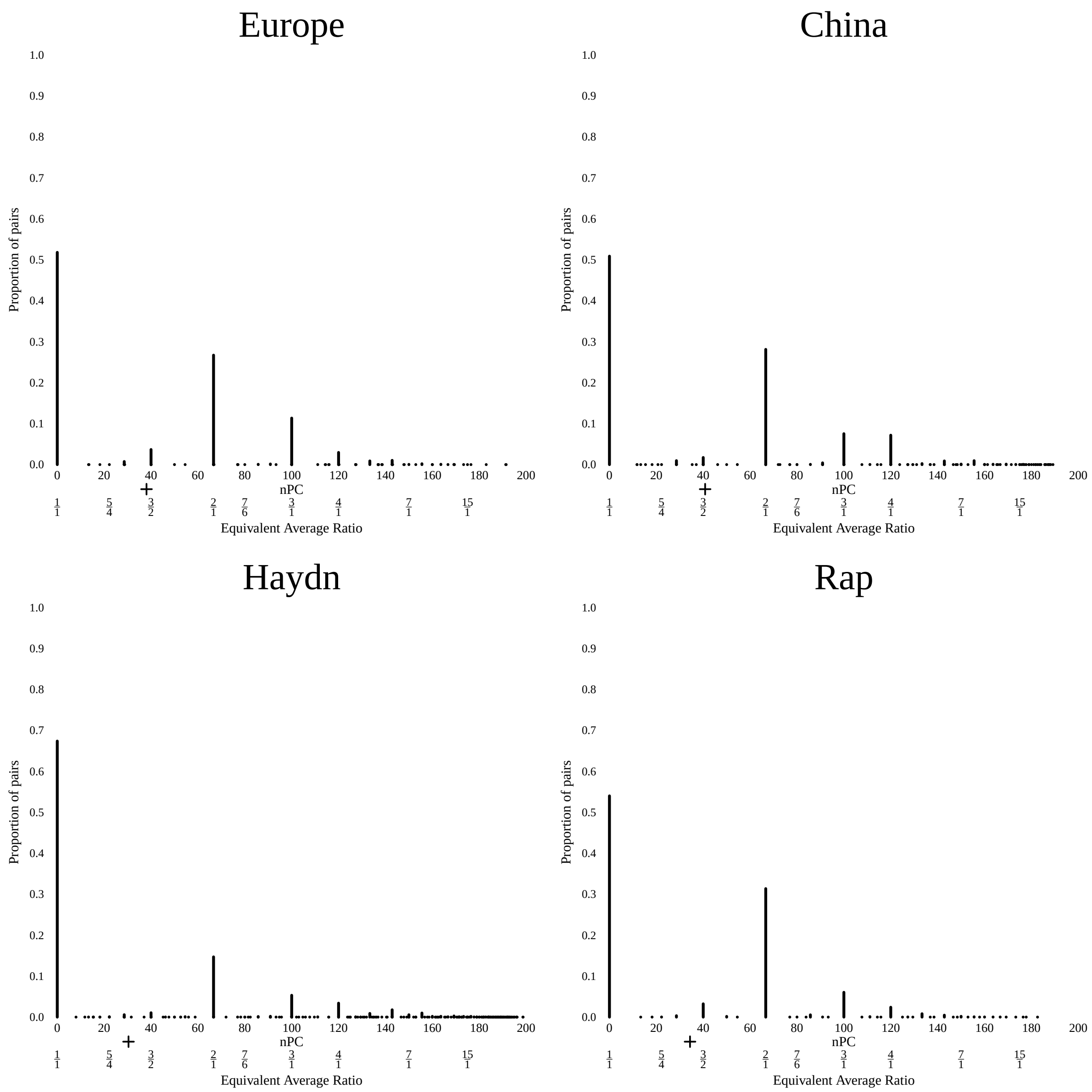

Distribution of nPVI across four corpora
nPVI
Equivalent Average Ratio
0
10
20
30
40
50
60
70
80
90
100
110
3/1
7/6
2/1
3/2
5/4
1/1
+ Europe ·
21 Regions 6044 Songs
+ China ·
4 Regions 1520 Songs
+ Haydn ·
16 Opuses 210 Movements
+ Rap ·
33 Years 392 Verses

Figure 4



470

471 **Figure 5**

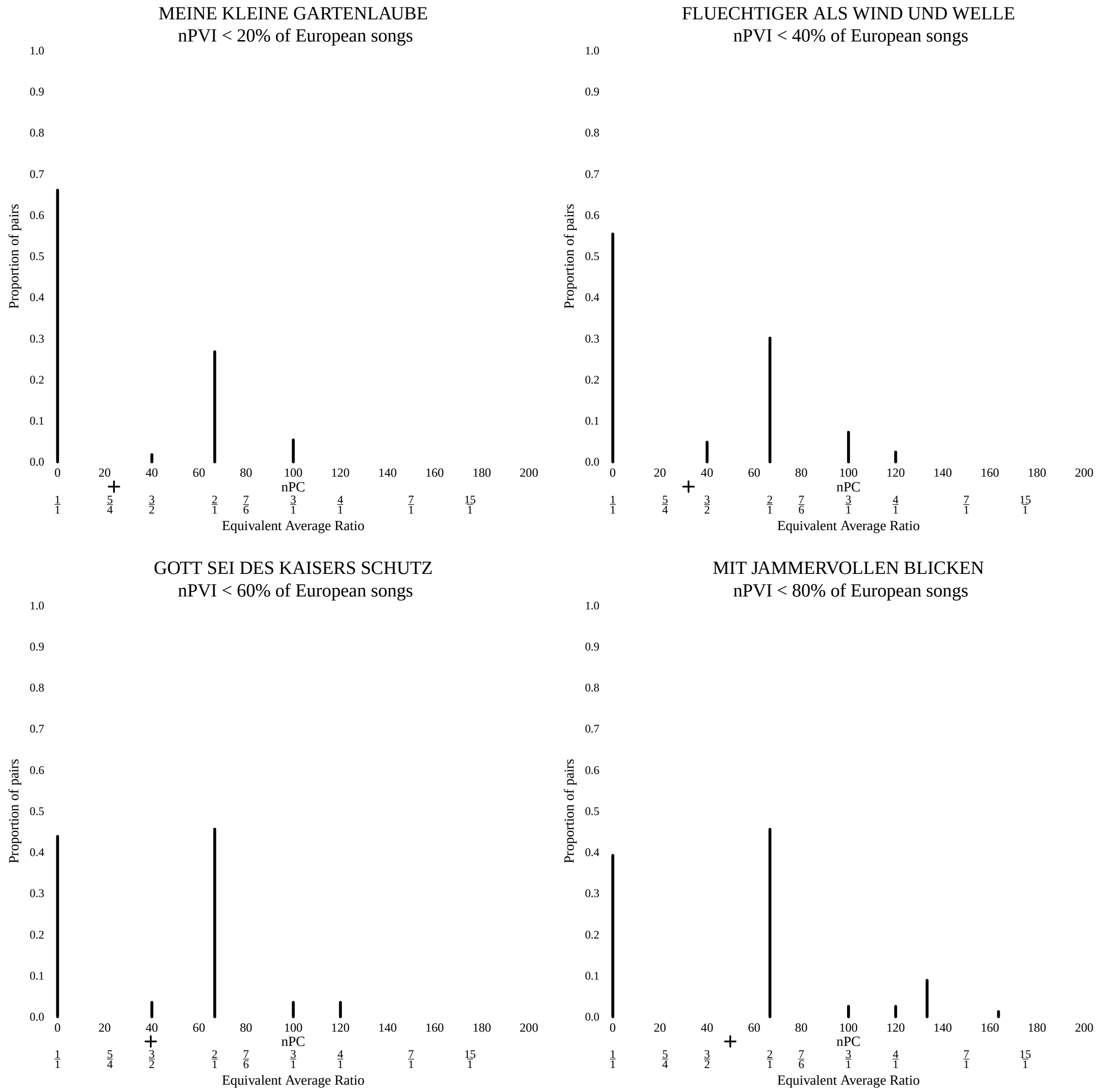

MEINE KLEINE GARTENLAUBE
nPVI < 20% of European songs
FLUECHTIGER ALS WIND UND WELLE
nPVI < 40% of European songs
GOTT SEI DES KAISERS SCHUTZ
nPVI < 60% of European songs
MIT JAMMERVOLLEN BLICKEN
nPVI < 80% of European songs
Proportion of pairs
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
0 20 40 60 80 100 120 140 160 180 200
nPC
1/1 5/4 3/2 2/1 7/6 3/1 4/1 7/1 15/1
Equivalent Average Ratio

Figure 6

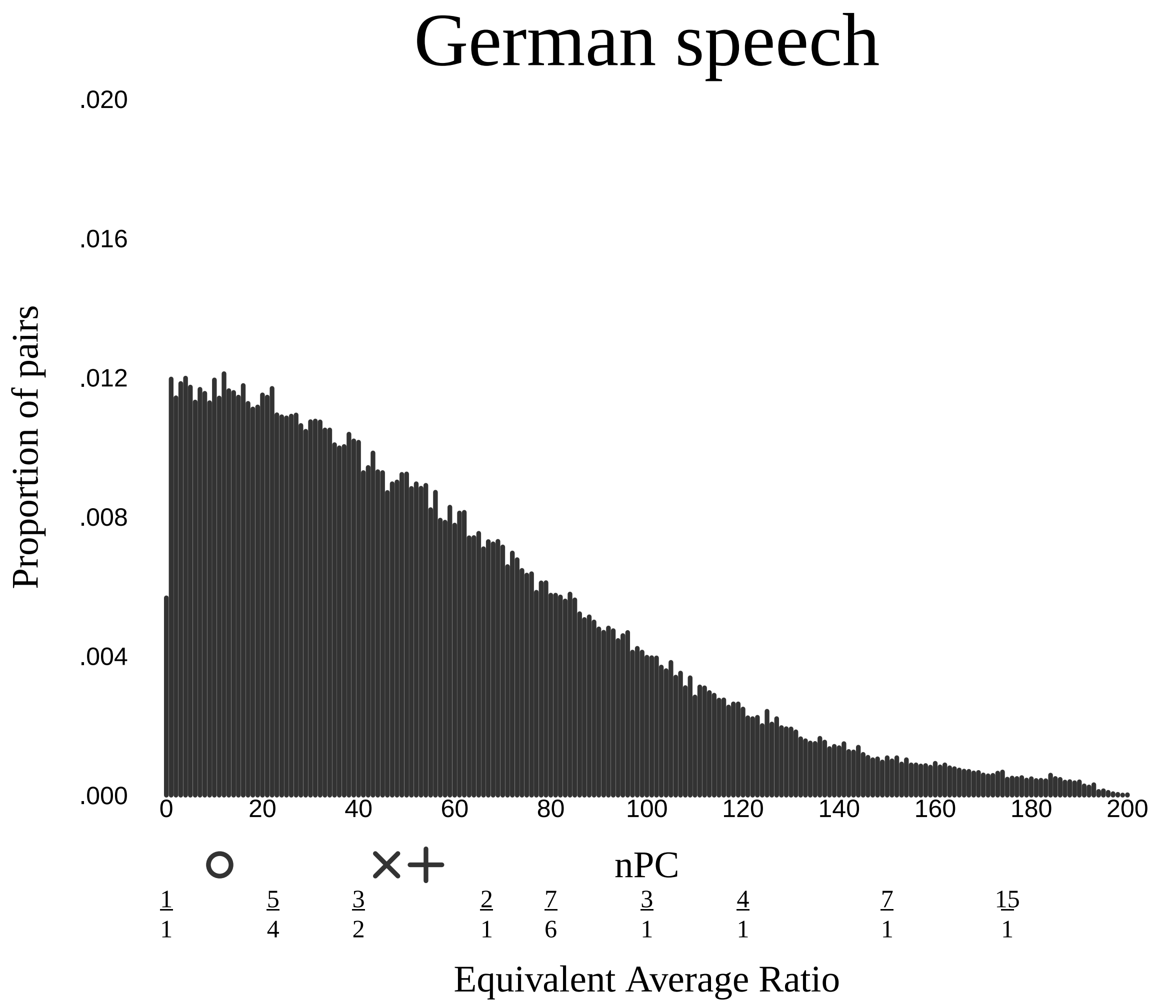

German speech
Proportion of pairs
.020
.016
.012
.008
.004
.000
0 20 40 60 80 100 120 140 160 180 200
nPC
1/1 5/4 3/2 2/1 7/6 3/1 4/1 7/1 15/1
Equivalent Average Ratio

Figure 7

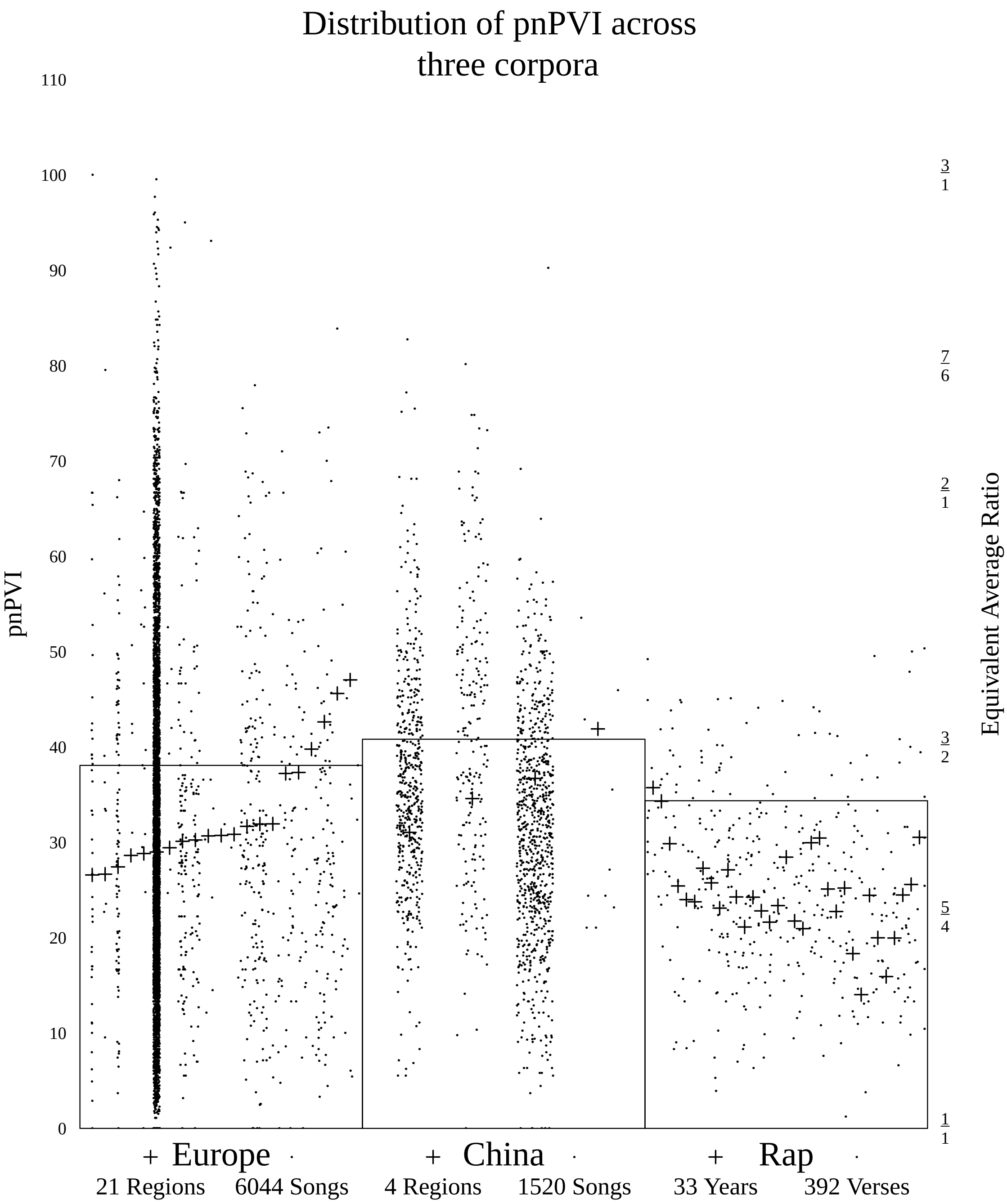

Distribution of pnPVI across three corpora
pnPVI
110
100
90
80
70
60
50
40
30
20
10
0
Equivalent Average Ratio
3/1
7/6
2/1
3/2
5/4
1/1
+ Europe ·
21 Regions 6044 Songs
+ China ·
4 Regions 1520 Songs
+ Rap ·
33 Years 392 Verses

Figure 8

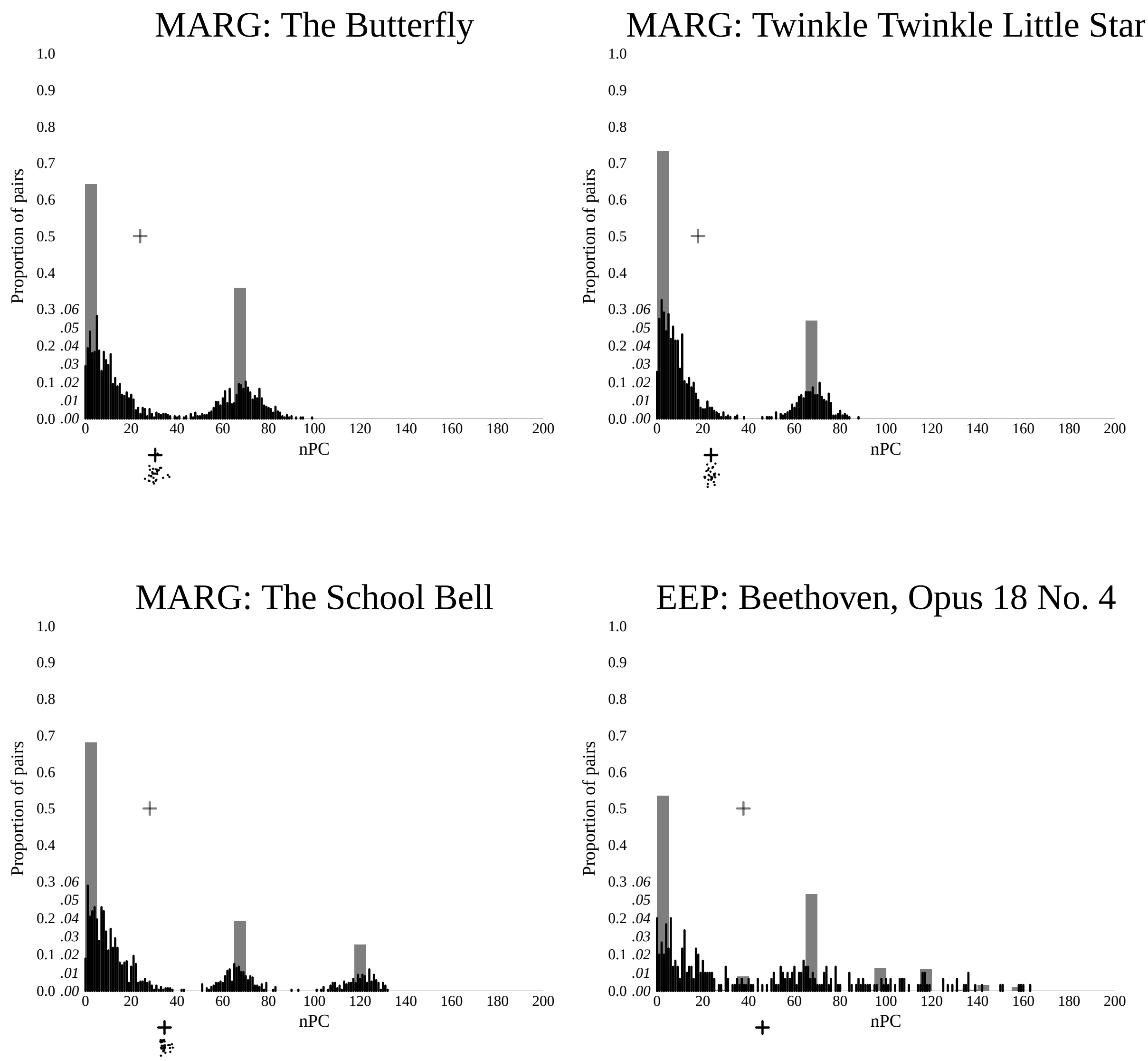

MARG: The Butterfly
MARG: Twinkle Twinkle Little Star
MARG: The School Bell
EEP: Beethoven, Opus 18 No. 4
Proportion of pairs
nPC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3 .06
.05
0.2 .04
.03
0.1 .02
.01
0.0 .00
0 20 40 60 80 100 120 140 160 180 200

Figure 9

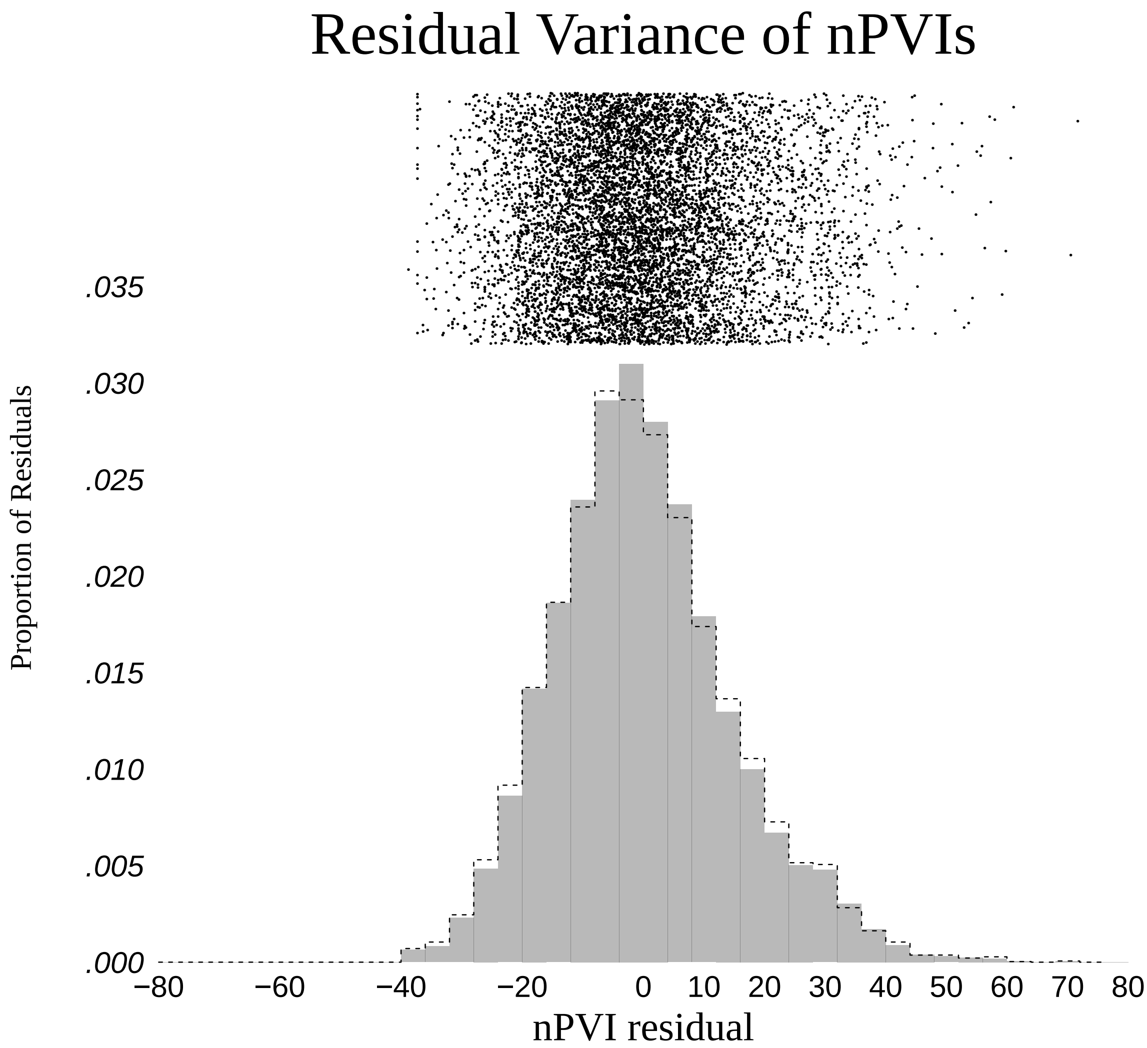

Residual Variance of nPVIs
Proportion of Residuals
.035
.030
.025
.020
.015
.010
.005
.000
−80
−60
−40
−20
0
10
20
30
40
50
60
70
80
nPVI residual

Figure 10