Modelling pattern interestingness in comparative music corpus analysis *

Kerstin Neubarth[a] and Darrell Conklin[b,c]

[a]Europa-Universität Flensburg, Flensburg, Germany;

[b]University of the Basque Country UPV/EHU, San Sebastian, Spain;

[c]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Abstract

In computational pattern discovery, pattern evaluation measures select or rank patterns according to their potential interestingness in a given analysis task. Many measures have been proposed to accommodate different pattern types and properties. This paper presents a method and case study employing measures for frequent, characteristic, associative, contrasting, dependent, and significant patterns to model pattern interestingness in a reference analysis, Frances Densmore's study of Teton Sioux songs. Results suggest that interesting changes from older to more recent Sioux songs according to Densmore's analysis are best captured by contrast, dependency, and significance measures.

1 Introduction

Pattern discovery provides powerful and versatile techniques for symbolic music analysis. Patterns in music include *intra-opus patterns*, repeated within a single piece of music, and *inter-opus patterns*, occurring across multiple pieces in a music corpus (Conklin, 2010a). Inter-opus pattern discovery can be applied to unstructured corpora, extracting patterns which describe general features of the represented repertoire (e.g. Conklin and Anagnostopoulou, 2001), or to partitioned corpora, extracting patterns which distinguish classes of music pieces such as different song types, geographic

*This is the Accepted Manuscript of the article published by Taylor & Francis in *Journal of Mathematics and Music*, 15(2):154-167, 2021, available at: https://doi.org/10.1080/17459737.2021.1900436

regions, or composers (e.g. Conklin and Anagnostopoulou, 2011; Collins et al., 2016). Traditionally, pattern mining in music – both intra- and inter-opus analysis – has been dominated by work on discovering *sequential patterns*. For inter-opus pattern mining, *global-feature patterns* offer an alternative pattern representation (e.g. Taminau et al., 2009; Shanahan, Neubarth, and Conklin, 2016). This paper studies inter-opus, global-feature patterns in class-labelled music corpora.

Searching for patterns in music may proceed as *deductive* analysis, which retrieves instances of specified patterns, or as *inductive* analysis, which finds unspecified patterns satisfying certain criteria of pattern interestingness (Conklin, 2010b). Beyond music data mining, many measures for quantifying pattern interestingness have been proposed, originating from different contexts including statistics and information theory (Geng and Hamilton, 2006); relating to different pattern types, such as frequent or contrasting patterns (Dong and Li, 1999; Bay and Pazzani, 2001; Han et al., 2007); and satisfying different properties regarding e.g. their scaling behaviour (Piatetsky-Shapiro, 1991; Tan, Kumar, and Srivastava, 2002; Lenca et al., 2007). Interestingness measures can be employed to distinguish interesting from uninteresting patterns, usually requiring the definition of a measure threshold, or to rank patterns. Measures are used during pattern discovery to prune the search space, or during post-processing to filter or rank the output of discovered patterns (Geng and Hamilton, 2006).

The work presented in this paper explores computational measures for modelling the interestingness of patterns suggested by extant music analyses. Hence it lies at the intersection of deductive and inductive analysis: it shares with deductive analysis the study of given patterns, and for studying these patterns makes use of pattern evaluation criteria usually employed in inductive analysis. More specifically, we report a case study on patterns in Native American music: a meta-analysis of Frances Densmore's analysis of Teton Sioux music, which investigates changes from older to more modern Sioux songs (Densmore, 1918). The case study illustrates core interests in data mining (e.g. Dong and Li, 1999) and computational music corpus analysis (e.g. Jackson, 1970; Broze and Shanahan, 2013): discovering changes in chronologically structured data.

2 Reference analysis: interesting changes in Teton Sioux music

Frances Densmore (1867–1957) was one of the most prolific collectors of North American native music. The case study in this paper focuses on Densmore's analysis of Teton Sioux songs, collected

Table 1: Music content descriptors in Densmore's analysis of Teton Sioux songs (Densmore, 1918).

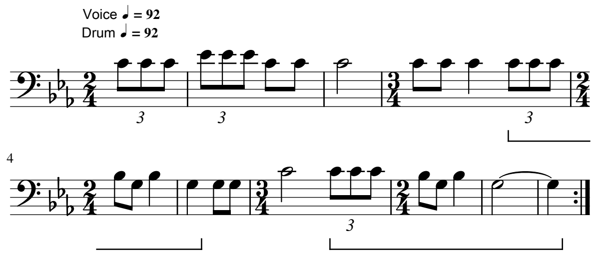
Attribute	Description
tonality	tonality [according to major/minor third above keynote]
firstReKey	first note of song – its relation to keynote
lastReKey	last note of song – its relation to keynote
lastReCompass	last note of song – its relation to compass of song
compass	number of tones comprising compass of song
material	tone material
accidentals	accidentals [chromatic alterations of tones]
structure	melodic structure [relation between contiguous accented tones]
firstProgression	first progression – downward and upward
firstMetricPos	part of measure on which song begins
firstMeasure	rhythm (metre) of first measure
metreChange	change of time (measure-lengths)
rhythmDrum	rhythm of drum
rhythmicUnit	rhythmic unit of song
tempoVoice	metric unit of voice (tempo)
tempoDrum	metric unit of drum (tempo)
tempoVoiceDrum	comparison of metric unit of voice and drum (tempo)

on the Standing Rock and Sisseton Reservations in North and South Dakota between 1911 and 1914 (Densmore, 1918). Like most of her publications with the Bureau of American Ethnology, the study of Teton Sioux music includes quantitative analyses of the documented songs based on global music content features, which capture the "melodic trend and general musical character" of the songs (Densmore, 1910, p. 3).

2.1 Densmore's collection of Teton Sioux music

Densmore's publication on Teton Sioux music presents transcriptions and analyses of 240 songs. The corpus is organised according to "the age of the songs, this series being divided for analysis into two groups, one comprising songs believed to be more than 50 years old and the other comprising songs of more recent origin" (Densmore, 1918, p. v). The class of older songs contains 147 songs associated with obsolete ceremonies or recorded by old men who had learned or received the songs in their youth. The class of comparatively modern songs comprises 93 songs recorded by young men, linked to modern tribal societies, or referencing a recent custom.

To describe musical properties of songs Densmore applied global features, i.e. song-level attribute–value pairs, which capture melodic and rhythmic-metric aspects of songs (Table 1). For the compu-



firstProgression : up
firstMetricPos : accented
firstMeasure : 2-4
metreChange : yes
rhythmDrum :
quarter_unaccented
rhythmicUnit : one
tempoVoice : slow
tempoDrum : slow
tempoVoiceDrum : same

Figure 1: Densmore's transcription of the modern Teton Sioux song "Song of the Buffalo Hunt (c)" (Cat. No. 545, Densmore, 1918, p. 442) with encoding by selected global features. Angular brackets mark the rhythmic unit.

tational analysis we collated the feature encoding for the 240 songs from Densmore's publication. To allow comparison with Densmore's analysis results, we aggregated attribute values when suggested by Densmore's textual description. For example, with respect to the attribute compass Densmore commented on songs "having a range of 12 or more tones" (Densmore, 1918, p. 24), aggregating values from 12 to 17 tones. For comparing the tempo of old and more modern Sioux songs, the values of attributes tempoVoice and tempoDrum (measured in metronome values) were aggregated into two bins, with a split point at the median, reducing the 30 resp. 27 fragmented and infrequently observed attribute values considered in Densmore's original tabulated analysis to two categorical values slow, covering metronome markings 48 to 96, and fast, covering metronome markings 100 to 192. Figure 1 shows a short example song encoded by selected global features.

2.2 Reference patterns in Densmore's analysis of Teton Sioux music

Densmore's comparative analysis of old and modern Sioux songs studied one attribute at a time, i.e. the analysis reveals single-feature patterns. To identify reference patterns in Densmore's analysis, we first extracted all pairs $\langle$*feature* , *class*$\rangle$ which are mentioned in the textual descriptions accompanying Densmore's quantitative analyses. From these, redundant patterns – due to symmetries between the two classes or in cases of only two attribute values – were removed. For example, the "larger proportion [of songs] having a range of 12 or more tones" (Densmore, 1918, p. 24) among the old songs implies a smaller proportion of wide-range songs in the modern group (Densmore, 1918, p. 25), and a "decrease in the percentage of songs having a change of measure-lengths" from the old to the modern songs implies "an increase [...] in songs without change in time" (Densmore, 1918, p. 25). Of such symmetric pairs we retained the over-represented pattern, in the above examples $\langle \mathsf{compass} : \mathsf{twelve\_or\_more}\, , \mathsf{old} \rangle$ and $\langle \mathsf{metreChange} : \mathsf{no}\, , \mathsf{modern} \rangle$. Where the text mentions only the under-represented pattern, for the current study this was replaced by the corresponding over-represented pattern, e.g. "[t]he modern songs show a smaller proportion of songs in which the final tone is the lowest in the song" (Densmore, 1918, p. 24) was recorded as the complementary group of old songs containing a larger proportion of songs ending on the lowest tone, i.e. by the pattern $\langle \mathsf{lastReCompass} : \mathsf{lowest}\, , \mathsf{old} \rangle$.

In a second step, we assigned the reference patterns a level of interestingness based on Densmore's description. Analysing changes from older to newer Teton Sioux songs, Densmore appears to distinguish different degrees of change: for example, "the newer group shows an increase in the proportion of songs which begin in 2–4 time" (Densmore, 1918, p. 24) but "a *large* increase in the proportion having two or more rhythmic units" (Densmore, 1918, p. 25, our emphasis). Additionally, Densmore's comments on the tabulated analyses are followed by a concluding paragraph – "[s]ummarizing briefly the results of a comparison of the old and the more modern Sioux songs" (Densmore, 1918, p. 25) – which recapitulates a selection of the previously presented observations. Assuming the patterns highlighted in the summary to be of particular interest, we derived an ordinal scale of four *reference levels* of interestingness:

- **A**: patterns covered in the summary and described, in the preceding text, by a qualifier marking a pronounced (e.g. "large" or "decided") change;
- **B**: patterns which are included in the summary but have not been qualified as a pronounced change;
- **C**: patterns mentioned in the text but not in the summary (none of which is explicitly qualified as a pronounced change);
- **D**: patterns which show "no material differences" (Densmore, 1918, p. 25) or for which the proportion "is the same in the two groups" (Densmore, 1918, p. 23).

The analysis results in 33 reference patterns (see Figure 3 in Section 3.2): six patterns at level **A**, five at level **B**, four at level **C**, and 18 patterns (nine patterns potentially associated with either class) at level **D**.

3 Computational analysis of the reference patterns

To model pattern interestingness in Densmore's analysis, the reference patterns are ranked by computational pattern interestingness measures, considering measures for different types of patterns: frequent and characteristic, contrasting, associative, dependent, and significant patterns.

3.1 Pattern interestingness measures

The pattern interestingness measures considered in this paper evaluate the distribution of pattern occurrences in a partitioned corpus. They map observed frequencies onto a numeric interestingness value, computed from a 2×2 contingency table (Figure 2). Here the variables C and X refer to predicates on songs, the class and pattern predicate respectively. A song satisfies a class C if it is annotated with the corresponding class label; a song satisfies a global-feature pattern X if the song has the attribute value indicated by the global feature; a song satisfies the conjunction of the class and pattern predicates, denoted $X \wedge C$, if it satisfies both X and C. We refer to the songs satisfying a specific class C as the *target class* of analysis and the songs not satisfying C as the *background*. To denote observed frequencies, the following notation is used: $n(C)$ is the number of songs in the target class, while $n(\neg C)$ is the number of songs in the background; $n(X)$ is the number of songs in the corpus satisfying the global-feature pattern X, while $n(\neg X)$ is the number of songs in the corpus which do not satisfy pattern X. Further, $n(X \wedge C)$ is the number of songs in the target class which satisfy pattern X, $n(\neg X \wedge C)$ the number of songs in the target class which do not satisfy pattern X, $n(X \wedge \neg C)$ the number of songs in the background which satisfy pattern X, and $n(\neg X \wedge \neg C)$ the number of songs in the background which do not satisfy pattern X. The variable N denotes the total number of songs in the corpus. The relative frequency of the pattern in the corpus, or its empirical probability, is $P(X) = n(X)/N$, and the relative frequency of the pattern in a class, or its conditional probability given the class, is $P(X|C) = n(X \wedge C)/n(C)$. Building on these definitions, Table 2 lists probability-based measures for frequent and characteristic, contrasting, associative,

	C	$\neg C$	
X	$n(X \wedge C)$	$n(X \wedge \neg C)$	$n(X)$
$\neg X$	$n(\neg X \wedge C)$	$n(\neg X \wedge \neg C)$	$n(\neg X)$
	$n(C)$	$n(\neg C)$	N

Figure 2: Contingency table for a pattern $\langle X , C \rangle$.

Table 2: Interestingness measures for patterns $\langle X\,,\, C\rangle$.

Pattern type	Measure	Definition
frequent	coverage	$P(X)$
	support	$P(X \wedge C)$
characteristic	sensitivity	$P(X\|C)$
	IC^{++}	$P(C)[1 - \frac{P(\neg X\|C)}{P(\neg X\|\neg C)}]$ if $0 \leq \frac{P(\neg X\|C)}{P(\neg X\|\neg C)} < 1$; 0 otherwise
contrasting	support difference	$P(X\|C) - P(X\|\neg C)$
	growth rate	$P(X\|C)\ /\ P(X\|\neg C)$
associative	confidence	$P(C\|X)$
dependent	PS	$P(X \wedge C) - P(X)P(C)$
	interest	$P(X \wedge C)\ /\ P(X)P(C)$
	conviction	$P(X)P(\neg C)\ /\ P(X \wedge \neg C)$
significant	p-value (Fisher)	$P_F(X, C)$

dependent, and significant patterns.

Frequent and characteristic patterns Frequent pattern mining is a core task in both wider data mining and music data mining. Patterns are considered frequent if they occur in a data set with frequency above a user-specified threshold (Han et al., 2007). In frequent pattern mining of class-labelled corpora, *coverage* measures the relative frequency of a pattern in a corpus, while *support* measures the relative co-occurrence of a pattern and a specific class (Geng and Hamilton, 2006). *Sensitivity* computes the proportion of instances in a class which satisfy a pattern (Lavrač, Flach, and Zupan, 1999). A pattern which is shared by all or most instances in a class is considered characteristic of the class (Han et al., 1996). Alternative measures prefer characteristic patterns which are also distinctive of the class. As an example this study includes the IC^{++} measure (Kamber and Shinghal, 1996): the more instances in the target class, relative to the background, do not satisfy the pattern, the less characteristic of the class is the pattern.

Contrasting patterns Contrast pattern mining identifies differences between classes in categorically partitioned data or trends in chronologically partitioned corpora (Dong and Li, 1999; Bay and Pazzani, 2001). Measures for contrasting patterns generally compare the observed relative frequencies of a pattern in a target class and in the background. The measure of *support difference*, used in contrast set mining (Bay and Pazzani, 2001), calculates their difference, while *growth rate*, used in emerging pattern mining (Dong and Li, 1999), computes their ratio. Although Densmore does not systematically quantify differences in her tabulated analyses, for both support difference and growth rate corresponding examples can occasionally be found in her textual description. As an example of the first, her comparison of old and modern Sioux songs shows that "the proportion beginning on the octave is 10 per cent greater in the modern songs" (Densmore, 1918, pp. 23-24). On the other hand, her analysis of Sioux, Chippewa, and Ute songs gives an example of growth rate: "The percentage of songs of a mixed form is more than twice as great in the Ute as in the Chippewa and Sioux" (Densmore, 1922, p. 53).

Associative patterns Associations describe frequently co-occuring patterns (Han et al., 1996); class associations relate frequent patterns to classes (Liu, Hsu, and Ma, 1998). Classic methods for mining class association patterns combine the measures of support to ensure sufficiently frequent patterns and *confidence* to assess the strength of the association between a pattern and a class: confidence corresponds to the conditional probability of the class given the pattern (Liu, Hsu, and Ma, 1998). The confidence measure does not take into account the prior probability of the class, thus a class association pattern may be confident when pattern and class are not correlated or even negatively correlated (e.g. Brin, Motwani, and Silverstein, 1997).

Dependent patterns To address problems of the confidence measure, alternative measures have been used, including the *PS* measure (Piatetsky-Shapiro, 1991), *interest* (Brin, Motwani, and Silverstein, 1997), and *conviction* (Brin et al., 1997): the PS measure, by taking the difference, and interest, by calculating the ratio, compare the joint and individual probabilities of the pattern and the target class, indicating to which extent pattern and class are statistically dependent. Conviction considers the pattern's occurrence in the background to quantify the dependence between pattern and target class. Hence, while the contrast measures compare the pattern's observed count in the target class against its observed count outside the class, relative to the size of the class and the background, the dependency measures compare the pattern's observed count in the class against its count expected under the assumption of pattern and class being independent. Rewriting the measures with absolute rather than relative frequencies – by multiplying the two summands or factors by N to give $n(X \wedge C)$ and $N \times P(X) \times P(C)$ for PS and interest resp. $n(X \wedge \neg C)$ and $N \times P(X) \times P(\neg C)$ for conviction – makes explicit the comparison.

Significant patterns Statistical significance tests estimate the likelihood of encountering observed pattern frequencies due to chance alone (Webb, 2007). Tests such as Fisher's test have been applied in class association mining both in combination with other measures, such as growth rate, confidence or PS, and on their own (e.g. Conklin, 2013; Shanahan, Neubarth, and Conklin, 2016; Li and Zaiane, 2017; Neubarth, Shanahan, and Conklin, 2018). The *p-value* computed by Fisher's exact test (right tail) gives the probability of observing $n(X \wedge C)$ or more occurrences of the pattern in the target class given the marginal counts $n(X)$, $n(C)$ and N. The lower the p-value, the more interesting is the pattern.

3.2 Comparison of reference and computational pattern interestingness

The interestingness measures listed in Table 2 were applied for computational evaluation of the reference patterns, adjusting marginal counts in the contingency table for missing values (Neubarth, Shanahan, and Conklin, 2018). For each interestingness measure, the evaluated patterns were ranked from highest to lowest measure value; in the case of Fisher's test, lower p-values indicate a higher degree of interestingness. In a second step, the computationally ranked patterns were mapped onto ordinal levels of interestingness, following the procedure of Ohsaki et al. (2004): based on the categorisation of the reference patterns (6 patterns at level **A**, 5 patterns at level **B**, 4 patterns at level **C**, and 18 patterns at level **D**; see Figure 3, column "Densmore"), for each interestingness measure the six most highly ranked patterns were assigned the interestingness level **A**, the next five patterns were assigned **B**, the following four patterns were assigned **C**, and the remaining patterns were assigned **D**. This mapping then provides a basis for comparing interestingness ratings of human and computational analysis qualitatively, by visualisation, or quantitatively, by determining the number of reference patterns matched in their interestingness levels by the computational evaluation (Ohsaki et al., 2004).

Figure 3 presents the interestingness ratings, for each reference pattern, based on Densmore's analysis (column "Densmore") and assigned by the computational pattern interestingness measures (columns "p-value" to "coverage"). Patterns are grouped into the reference levels suggested from Densmore's analysis, from **A** (top) to **D** (bottom). Rows within levels are ordered according to agreement across measures, while columns are sorted by agreement across patterns. Comparing computationally evaluated pattern interestingness against the reference levels extracted from Dens-

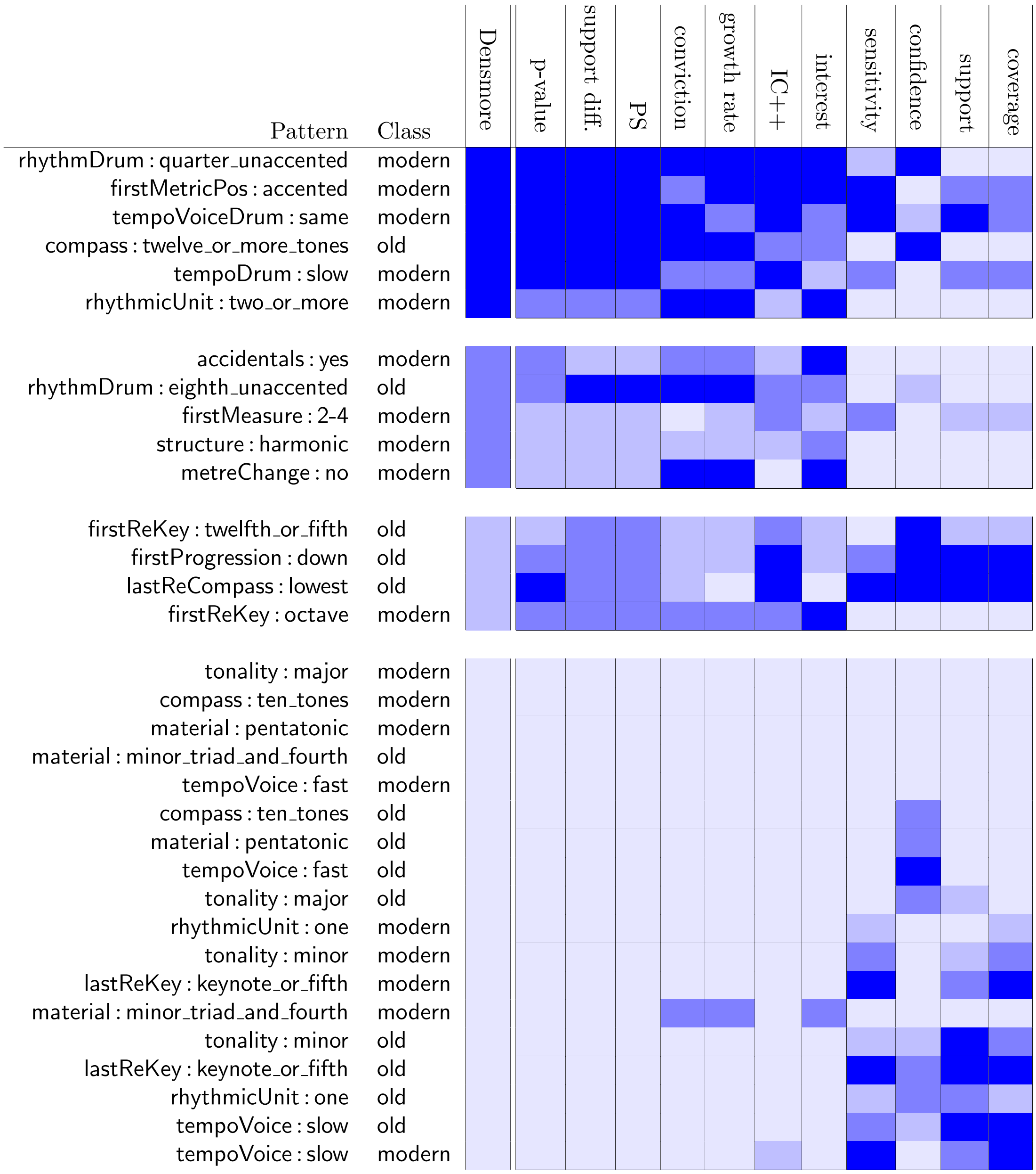


Figure 3: Comparison of pattern interestingness ratings suggested by Densmore's analysis and by computational interestingness measures. Colour legend (interestingness levels): level **A**; level **B**; level **C**; level **D**.

more's analysis, the measures roughly fall into two groups. The measures for frequent patterns – coverage, support, and sensitivity – are largely unsuccessful in distinguishing uninteresting from interesting patterns, according to the reference analysis, matching only ten out of 18 patterns at reference level **D** and at most two out of five patterns at reference level **A**. A larger number of corresponding ratings is achieved by the IC^{++} measure, which is biased towards characteristic patterns that also distinguish the target class from other classes (Kamber and Shinghal, 1996). On the other hand, measures for contrasting and for dependent and significant patterns show overall high agreement with reference ratings, with p-value, support difference and PS matching reference levels slightly better (all except one reference pattern matched at level **A** and all reference patterns matched at level **D**) than growth rate, conviction and interest (three – in the case of interest two – out of five patterns matched at level **A** and 17 out of 18 patterns matched at level **D**). At levels **B** and **C**, for several patterns the ratings are reversed with respect to the reference ratings, suggesting that the inclusion of a pattern in Densmore's summary may be less indicative of quantitative interestingness ratings than her differentiation between strong and neutral changes (e.g. "large increase" vs. "increase"), and that Densmore's selection of patterns included in the summary may be partly based on other, musical or contextual, rather than statistical considerations.

3.3 Analysis of interestingness ratings by computational measures

This section further analyses the differences in interestingness ratings by different computational measures, discussing selected patterns (Table 3). To generalise observations beyond example patterns, we refer to established measure properties which define a measure's behaviour for varying contingency tables. In particular, we build on two well-known properties (Piatetsky-Shapiro, 1991; Tan, Kumar, and Srivastava, 2002): the first covers two scenarios affecting the relative pattern frequencies in the target class and background, the second supports comparison of interestingness ratings for frequent and infrequent patterns, and for different class distributions (Table 4).

Frequency measures Of the measures for frequent and characteristic patterns, support and sensitivity do not consider the background, while coverage quantifies pattern frequency across all classes in a corpus. Hence, sensitivity (coverage) assigns high ranks to patterns which are frequent in the target class (corpus) even if they are similarly or more frequent in the background. For example, the feature lastReKey : keynote_or_fifth occurs in 84% of the old and 85% of the modern

Table 3: Distribution of selected patterns in old and modern Teton Sioux songs (ordered according to their reference in the text). The pattern marked by an asterisk was evaluated taking into account missing values (87 of the old and 31 of the modern songs were recorded without drum, cf. Densmore, 1918, p. 21, footnote 1).

	old		modern	
	$n(X \wedge C)$	$P(X\|C)$	$n(X \wedge C)$	$P(X\|C)$
lastReKey : keynote_or_fifth	124	0.84	79	0.85
compass : twelve_or_more_tones	47	0.32	20	0.22
rhythmicUnits : two_or_more	14	0.09	16	0.17
firstReKey : twelfth_or_fifth	63	0.43	33	0.35
firstMetricPos : accented	75	0.51	69	0.74
tempoDrum : slow*	31	0.52	43	0.69
metreChange : no	9	0.06	9	0.10
firstReKey : octave	30	0.20	27	0.29
lastReCompass : lowest	133	0.90	76	0.82

songs (see Table 3): both associations are ranked among the top-6 patterns (level **A**) by sensitivity and coverage, while they are ranked in agreement with the reference analysis (level **D**) by measures which compare pattern occurrence in the target class and in the background, such as support difference and growth rate, as well as IC^{++}. On the other hand, coverage, sensitivity, and support penalise infrequent features, such as compass : twelve_or_more_tones or rhythmicUnits : two_or_more (assigning level **D**), even if they are distinctive for one of the classes (listed at reference level **A** and assigned levels **A** or **B** by contrast, dependency, and significance measures).

Formally, these observations are captured by measure property M1.1 in Table 4. The property describes a measure's behaviour when, given the same pattern frequency in the target class, a higher number of pattern occurrences is observed in the background: with increasing pattern frequency in the corpus, $n(X)$, but constant pattern frequency in the target class, $n(X \wedge C)$, more pattern occurrences are found in the background, both in absolute frequency $n(X \wedge \neg C)$ and – with constant $P(C)$ and therefore $P(\neg C)$ – in relative frequency $P(X|\neg C)$. Hence, for patterns over-represented in the target class the difference between a pattern's frequency in the class and in the background decreases. Contrast measures, as well as dependency and significance measures, accordingly decrease their value. Support and sensitivity, on the other hand, remain constant while coverage increases (Table 4, column M1.1).

Association measures Confidence quantifies the proportion of pattern occurrences observed in the target class, without taking into account the prior probability of the class (Brin, Motwani, and

Silverstein, 1997). For illustration, pattern $\langle \textsf{firstReKey} : \textsf{twelfth\_or\_fifth} , \textsf{old} \rangle$, included at reference level **C**, is ranked higher by confidence (level **A**) than by the other measures. Of the 96 songs which begin on the twelfth or fifth above the keynote, 63 songs are old songs (see Table 3), giving a confidence of 66%, which is only slightly above the proportion of old songs in the corpus (61%). On the other hand, pattern $\langle \textsf{firstMetricPos} : \textsf{accented} , \textsf{modern} \rangle$, listed at reference level **A**, is ranked lower by confidence (level **D**) than any other measure, despite its confidence being higher than expected given the prior probability of the class modern (39%): of the 144 songs starting with an accented tone, 69 songs are modern songs (see Table 3), giving a confidence of 48%. In terms of change in relative frequency from background to target class, the beginning on the twelfth or fifth above the keynote shows a difference of only 8% (see Table 3: 0.35 to 0.43), leading to the lower ranking by contrast as well as dependency and significance measures. In comparison, the beginning with an accented note increases by 23% from older to more modern songs (see Table 3: 0.51 to 0.74), leading to the higher ranking by contrast as well as dependency and significance measures. More generally, most of the patterns ranked high by confidence relate to class old.

Confidence differs from measures for contrasting, dependent, and significant patterns with respect to measure property M1.2 in Table 4. Here the absolute numbers of pattern occurrences in the target class and in the background remain the same, but their relative frequencies change: with increasing $n(C)$, and therefore decreasing $n(\neg C)$, the pattern's relative frequency in the target class $P(X|C)$ decreases, while its relative frequency in the background $P(X|\neg C)$ increases. Thus again the degree to which a pattern is over-represented in the target class relative to the background decreases. While contrast, dependency, and significance measures decrease, confidence remains constant (Table 4, column M1.2).

Contrast measures Support difference and growth rate directly compare pattern occurrence in the target class and in the background, thus they capture changes in relative pattern frequency between older and more modern Sioux songs. The two measures differ in their ranking of frequent and infrequent patterns: for infrequent patterns a clear increase in relative frequency measured as ratio (growth rate) is easier to achieve than when measured as difference (support difference). In fact, contrast pattern discovery employing growth rate is specifically designed for also detecting changes in data when patterns are rare (Dong and Li, 1999). For example, pattern $\langle \textsf{tempoDrum} : \textsf{slow} , \textsf{modern} \rangle$ is ranked higher by support difference (level **A**) than by growth rate (level **B**), while

Table 4: Properties of interestingness measures (for details see text). * Direction of change for over-represented patterns, i.e. for $P(X|C) > P(X|\neg C)$; opposite direction of change for under-represented patterns. † For p-value, with lower values indicating higher interestingness, entries refer to change in interestingness.

Pattern type	Measure	Measure properties			
		M1.1	**M1.2**	**M2.1**	**M2.2**
frequent	coverage	increases	constant	increases	constant
	support	constant	constant	increases	increases
characteristic	sensitivity	constant	decreases	increases	constant
	IC^{++}	decreases	decreases	increases	increases
contrasting	support difference	decreases	decreases	increases	increases*
	growth rate	decreases	decreases	constant	increases*
associative	confidence	decreases	constant	constant	increases
dependent	PS	decreases	decreases	increases	increases*
	interest	decreases	decreases	constant	constant
	conviction	decreases	decreases	constant	increases*
significant	p-value	decreases†	decreases†	increases†	increases†

M1.1 Change with increasing $P(X)$ when $P(X \wedge C)$ and $P(C)$ remain the same
M1.2 Change with increasing $P(C)$ when $P(X \wedge C)$ and $P(X)$ remain the same
M2.1 Change with scaling the first row of the contingency table by a positive factor
M2.2 Change with scaling the first column of the contingency table by a positive factor

pattern ⟨metreChange : no , modern⟩ is ranked higher by growth rate (level **A**) than by support difference (level **C**). The feature tempoDrum : slow describes 69% of the modern and 52% of the old songs which are recorded with drum (see Table 3). On the other hand, the feature metreChange : no describes only 18 songs (7.5%) in the corpus, with nine songs in each of the two classes (6% and 10% respectively in the old and in the modern songs, see Table 3).

This difference between support difference and growth rate can be formally described with reference to measure property M2.1 in Table 4. Scaling the first row of the contingency table by a positive factor increases a pattern's frequency in the corpus but preserves the ratio between $P(X|C)$ and $P(X|\neg C)$. Growth rate therefore remains constant; support difference, on the other hand, increases for more frequent patterns (Table 4, column M2.1). In other words, for the same growth rate value a larger difference between $P(X|C)$ and $P(X|\neg C)$ is required at higher pattern frequencies.

Dependency and significance measures Among the measures for dependent and significant patterns, the interest and conviction measures are more sensitive to deviations in infrequent patterns than the PS measure and p-value. Rewriting PS as weighted relative accuracy, $P(X) \times [P(C|X) - P(C)]$, more obviously exposes the measure's bias towards frequent patterns: in subgroup discovery, weighted relative accuracy is applied to discover distinctive patterns which are as frequent as possible (Kavšek and Lavrač, 2006). Accordingly, the pattern $\langle$lastReCompass : lowest , old$\rangle$ – involving feature lastReCompass : lowest observed for 87% of all songs (see Table 3: 133 old and 76 modern songs, that is 209 of 240 songs) – is ranked higher by the PS measure (level **B**) than by conviction (level **C**) and interest (level **D**). On the other hand, the infrequent pattern $\langle$metreChange : no , modern$\rangle$ is ranked higher by conviction and interest (level **A**) than by PS (level **C**). In turn, conviction and interest differ in their interestingness ratings for different class sizes: in mining imbalanced data sets, interest has been found more suitable to discover patterns for the minority class than conviction (Abdellatif, Ben Hassine, and Ben Yahia, 2019). In the analysis of changes from older to more recent Teton Sioux songs the different bias of the interest and conviction measures is reflected in the respective ratings of patterns such as $\langle$compass : twelve_or_more_tones , old$\rangle$ and $\langle$firstNoteReKey : octave , modern$\rangle$. Both mentioned patterns occur in around 30% of songs in the target class and just over 20% of songs in the background (see Table 3). In the first case, the target class old is the majority class (61%) in the corpus, while in the second case the target class modern is the minority class (39%). The majority-class pattern is ranked higher by conviction (level **A**) than by interest (level **B**), while the minority-class pattern is ranked higher by interest (level **A**) than by conviction (level **B**).

In terms of formal measure properties, interest is invariant to scaling the first column of the contingency table by a positive factor while conviction increases (Table 4, column M2.2). Scaling the first column increases the size of the target class, both in terms of its absolute count and also, with N remaining constant, in terms of its proportion in the corpus.

In summary, referring to the visual comparison of computational against reference interestingness ratings (Figure 3), these differences between the computational measures and their properties are reflected in high ranks (shown as dark blue cells) assigned to patterns at reference level **A** based on differences in relative frequencies between target class and background (dark blue cells for e.g. p-value, support difference, and PS); at reference level **B** biased towards infrequent contrast patterns (dark blue cells for growth rate, conviction, and interest); at reference level **C** biased towards patterns for the majority class (dark blue cells for confidence and IC++); and at reference level **D** based on pattern frequency without contrasting target class and background (dark blue cells for coverage, support, and sensitivity).

4 Conclusions

In pattern mining, pattern interestingness measures distinguish potentially interesting from uninteresting patterns. In this paper we have presented a strategy and case study of exploiting interestingness measures and their properties to analyse patterns suggested by given music corpus studies. The contribution of this work is threefold.

First, computational pattern evaluation can support the meta-analysis of extant music corpus analyses. In the case of Densmore's comparison between old and modern Teton Sioux songs, the results of the computational analysis confirm Densmore's interest in differences between the two classes of songs – which is also reflected in the definition of the reference levels – complemented by a potential slight preference for more frequent patterns: "The first *important* point of *difference* is that the older songs show a much larger proportion having a range of 12 or more tones", while in "perhaps the *least important* of the tables [...] the groups show *no marked differences*" (Densmore, 1918, p. 24, our emphasis). The concluding sentences of Densmore's summary refer to "*contrasts* between the two groups" (Densmore, 1918, p. 25, our emphasis). Besides measures for contrasting patterns, including growth rate or support difference (Dong and Li, 1999; Bay and Pazzani, 2001), measures for dependent patterns, such as the PS measure and interest, have also been explicitly used for contrast mining, the latter in combination with a statistical significance test (Webb, Butler, and Newlands, 2003; Novak et al., 2009).

Second, the method can be useful in analysing ground-truth, or reference, patterns employed for assessing pattern discovery algorithms. Quantitative studies of music pattern discovery using metrics such as precision and recall with respect to ground-truth patterns assess discovery algorithms equally on all patterns in the ground-truth set, thus implicitly assuming all ground-truth patterns to be of the same pattern type (e.g. van Kranenburg and Conklin, 2016, Nuttall et al., 2019 for inter-opus pattern discovery in partitioned corpora; see also de Reuse and Fujinaga, 2019). An analysis of the ground-truth patterns by computational interestingness measures can provide insights into the pattern types and the homogeneity or potential heterogeneity of the ground-truth pattern set. Existing studies on music pattern discovery algorithms have focused on sequential patterns; the pattern evaluation strategy presented in this paper can be easily extended from global-feature to sequential inter-opus patterns by defining a suitable pattern predicate (Conklin, 2010a). A challenge, however, lies in the definition of the ground-truth patterns: if score annotations only identify patterns in the respective target class (e.g. van Kranenburg, Volk, and Wiering, 2012), pattern interestingness measures which consider pattern counts in the background cannot be computed.

Third, insights from studying pattern interestingness measures and their behaviour in the context of extant music corpus analyses can inform inductive pattern discovery, more specifically the selection of pattern evaluation measures depending on e.g. desired pattern types, expected or required pattern frequency, or the class distribution in the corpus. Regarding Densmore's studies, computational pattern mining also supports extending analysis beyond Densmore's single features to feature-set patterns (Neubarth, Shanahan, and Conklin, 2018). The comparison of older and more recent Sioux songs focuses on distinctive patterns, which can be discovered by contrast, dependency, or significance measures, but other pattern types, such as characteristic or associative patterns, may equally be of interest (Densmore, 1922, 1923).

Acknowledgements

The authors would like to thank the two anonymous reviewers and the journal editors for their valuable and helpful comments.

References

Abdellatif, Safa, Mohamed Ali Ben Hassine, and Sadok Ben Yahia. 2019. "Novel interestingness measures for mining significant association rules from imbalanced data." In *Web, Artifical Intelligence and Network Applications (WAINA 2019),* edited by Leonard Barolli, Makoto Takizawa, Fatos Xhafa, and Tomoya Enokido, Cham, Switzerland, 172–182. Springer.

Bay, Stephen D., and Michael J. Pazzani. 2001. "Detecting group differences: mining contrast sets." *Data Mining and Knowledge Discovery* 5 (3): 213–246.

Brin, Sergey, Rajeev Motwani, and Craig Silverstein. 1997. "Beyond market baskets: generalizing association rules to correlations." *ACM SIGMOD Record* 26 (2): 265–276.

Brin, Sergey, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. 1997. "Dynamic itemset counting and implication rules for market basket data." In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'97),* Tucson, Arizona, USA, 255–264.

Broze, Yuri, and Daniel Shanahan. 2013. "Diachronic changes in jazz harmony: a cognitive perspective." *Music Perception* 31 (1): 32–45.

Collins, Tom, Andreas Arzt, Harald Frostel, and Gerhard Widmer. 2016. "Using geometric symbolic fingerprinting to discover distinctive patterns in polyphonic music." In *Computational Music Analysis,* edited by David Meredith, 445–474. Springer.

Conklin, Darrell. 2010a. "Discovery of distinctive patterns in music." *Intelligent Data Analysis* 14: 547–554.

Conklin, Darrell. 2010b. "Distinctive patterns in the first movement of Brahms' string quartet in C minor." *Journal of Mathematics and Music* 4 (2): 85–92.

Conklin, Darrell. 2013. "Antipattern discovery in folk tunes." *Journal of New Music Research* 42 (2): 161–169.

Conklin, Darrell, and Christina Anagnostopoulou. 2001. "Representation and discovery of multiple viewpoint patterns." In *Proceedings of the International Computer Music Conference (ICMC 2001),* Havana, Cuba, 479–485.

Conklin, Darrell, and Christina Anagnostopoulou. 2011. "Comparative pattern analysis of Cretan folk songs." *Journal of New Music Research* 40 (2): 119–125.

de Reuse, Timothy, and Ichiro Fujinaga. 2019. "Pattern clustering in monophonic music by learning a non-linear embedding from human annotations." In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019),* Delft, Netherlands, 761–768.

Densmore, Frances. 1910. *Chippewa Music.* Washington, DC: Smithsonian Institution, Bureau of American Ethnology, Bulletin 45.

Densmore, Frances. 1918. *Teton Sioux Music.* Washington, DC: Smithsonian Institution, Bureau of American Ethnology, Bulletin 61.

Densmore, Frances. 1922. *Northern Ute Music.* Washington, DC: Smithsonian Institution, Bureau of American Ethnology, Bulletin 75.

Densmore, Frances. 1923. *Mandan and Hidatsa Music.* Washington, DC: Smithsonian Institution, Bureau of American Ethnology, Bulletin 80.

Dong, Guozhu, and Jinyan Li. 1999. "Efficient mining of emerging patterns: discovering trends and differences." In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99),* San Diego, CA, USA, 43–52.

Geng, Liqiang, and Howard J. Hamilton. 2006. "Interestingness measures for data mining: a survey." *ACM Computing Surveys* 38 (3): 1–32.

Han, Jiawei, Hong Cheng, Dong Xin, and Xifeng Yan. 2007. "Frequent pattern mining: current status and future directions." *Data Mining and Knowledge Discovery* 15: 55–86.

Han, Jiawei, Yongjian Fu, Wei Wang, Krzysztof Koperski, and Osmar Zaiane. 1996. "DMQL: a data mining query language for relational databases." In *Proceedings of the ACM SIGMOD Workshop on Data Mining and Knowledge Discovery (DMKD-96),* Montreal, Canada, 27–33.

Jackson, Roland. 1970. "Harmony before and after 1910: a computer comparison." In *The Computer and Music,* edited by Harry B. Lincoln, 132–146. Ithaca, New York: Cornell University Press.

Kamber, Micheline, and Raijan Shinghal. 1996. "Evaluating the interestingness of characteristic rules." In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96),* Portland, Oregon, USA, 263–266.

Kavšek, Branko, and Nada Lavrač. 2006. "APRIORI-SD: adapting association rule learning to subgroup discovery." *Applied Artificial Intelligence* 20: 543–583.

Lavrač, Nada, Peter Flach, and Blaz Zupan. 1999. "Rule evaluation measures: a unifying view." In *Proceedings of the 9th International Workshop on Inductive Logic Programming (ILP-99),* Bled, Slovenia, 174–185.

Lenca, Philippe, Benoît Vaillant, Patrick Meyer, and Stéphane Lallich. 2007. "Association rule interestingness measures: experimental and theoretical studies." In *Quality Measures in Data Mining,* edited by Fabrice Guillet and Howard J. Hamilton, 51–76. Berlin, Heidelberg: Springer.

Li, Jundong, and Osmar R. Zaiane. 2017. "Exploiting statistically significant dependent rules for associative classification." *Intelligent Data Analysis* 21 (5): 1155–1172.

Liu, Bing, Wynne Hsu, and Yiming Ma. 1998. "Integrating classification and association rule mining." In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98),* New York, USA, 80–86.

Neubarth, Kerstin, Daniel Shanahan, and Darrell Conklin. 2018. "Supervised descriptive pattern discovery in Native American music." *Journal of New Music Research* 47 (1): 1–16.

Novak, Petra Kralj, Nada Lavrač, Dragan Gamberger, and Antonija Krstačić. 2009. "CSM-SD: methodology for contrast set mining through subgroup discovery." *Journal of Biomedical Informatics* 42 (1): 113–122.

Nuttall, Thomas, Miguel García Casado, Víctor Núñez Tarifa, Rafael Caro Repetto, and Xavier Serra. 2019. "Contributing to new musicological theories with computational methods: the case of centonization in Arab-Andalusian music." In *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR 2019),* Delft, Netherlands, 223–228.

Ohsaki, Miho, Shinya Kitaguchi, Kazuya Okamoto, Hideto Yokoi, and Takahira Yamaguchi. 2004. "Evaluation of rule interestingness measures with a clinical dataset on hepatitis." In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2004),* Pisa, Italy, 362–373.

Piatetsky-Shapiro, Gregory. 1991. "Discovery, analysis and presentation of strong rules." In *Knowledge Discovery in Databases,* edited by Gregory Piatetsky-Shapiro and William Frawley, 229–248. Cambridge, MA: MIT Press.

Shanahan, Daniel, Kerstin Neubarth, and Darrell Conklin. 2016. "Mining musical traits of social functions in Native American music." In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016),* New York, USA, 681–687.

Taminau, Jonatan, Ruben Hillewaere, Stijn Meganck, Darrell Conklin, Ann Nowé, and Bernard Manderick. 2009. "Descriptive subgroup mining of folk music." In *2nd International Workshop on Machine Learning and Music at ECML/PKDD 2009 (MML 2009),* Bled, Slovenia.

Tan, Pang-Ning, Vipin Kumar, and Jaideep Srivastava. 2002. "Selecting the right interestingness measure for association patterns." In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002),* Edmonton, Alberta, Canada, 32–41.

van Kranenburg, Peter, and Darrell Conklin. 2016. "A pattern mining approach to study a collection of Dutch folk songs." In *6th International Workshop on Folk Music Analysis (FMA 2016),* Dublin, Ireland.

van Kranenburg, Peter, Anja Volk, and Frans Wiering. 2012. "On identifying folk song melodies employing recurring motifs." In *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for Cognitive Sciences of Music (ICMPC/ESCOM 2012),* Thessaloniki, Greece, 1057–1062.

Webb, Geoffrey I. 2007. "Discovering significant patterns." *Machine Learning* 68: 1–33.

Webb, Geoffrey I., Shane Butler, and Douglas Newlands. 2003. "On detecting differences between groups." In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003),* Washington, DC, USA, 256–265.