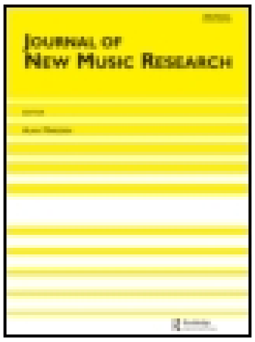

Journal of New Music Research

ISSN: 0929-8215 (Print) 1744-5027 (Online) Journal homepage: http://www.tandfonline.com/loi/nnmr20

# An Information Theoretic Approach to Chord Categorization and Functional Harmony

Nori Jacoby, Naftali Tishby & Dmitri Tymoczko

**To cite this article:** Nori Jacoby, Naftali Tishby & Dmitri Tymoczko (2015): An Information Theoretic Approach to Chord Categorization and Functional Harmony, Journal of New Music Research, DOI: 10.1080/09298215.2015.1036888

**To link to this article:** http://dx.doi.org/10.1080/09298215.2015.1036888

Published online: 22 Sep 2015.

Journal of New Music Research, 2015
http://dx.doi.org/10.1080/09298215.2015.1036888

# An Information Theoretic Approach to Chord Categorization and Functional Harmony

Nori Jacoby[1,2*], Naftali Tishby[1] and Dmitri Tymoczko[3]

[1]*Hebrew University of Jerusalem, Israel;* [2]*Bar Ilan University, Israel;* [3]*Princeton University, USA*

*(Received 1 April 2014; accepted 25 March 2015)*

## Abstract

We present new tools for categorizing chords based on corpus data, applicable to a variety of representations from Roman numerals to MIDI notes. Using methods from information theory, we propose that harmonic theories should be evaluated by at least two criteria, accuracy (how well the theory describes the musical surface) and complexity (the efficiency of the theory according to Occam's razor). We use our methods to consider a range of approaches in music theory, including function theory, root functionality, and the figured-bass tradition. Using new corpus data as well as eleven datasets from five published works, we argue that our framework produces results consistent both with musical intuition and previous work, primarily by recovering the tonic/subdominant/dominant categorization central to traditional music theory. By showing that functional harmony can be analysed as a clustering problem, we link machine learning, information theory, corpus analysis and music theory.

**Keywords:** functional harmony, corpus analysis, cluster analysis, information theory, information bottleneck

## 1. Introducing the framework

### 1.1 Introduction

Western harmony has a complex structure characterized by a large number of building blocks (chords) and a larger number of ways to combine them (chord progressions). This complexity is an expressive boon for composers, but a methodological challenge for theorists and pedagogues. For example, since figured-bass theory uses at least 49 separate chord symbols (seven bass notes plus three triadic and four seventh-chord figures, with additional categories for nondiatonic and nonharmonic configurations) figured-based treatises typically contain long lists of prohibitions, desiderata, and intermediate cases (e.g. Gasparini, 1715; Heinichen, 1728; Niedt, 1706; Praetorius, 1615).

One of music theory's central concerns is describing this structure in an approximate and simplified way. The common denominator among various approaches is the attempt to identify concise principles (such as Schenker's *Ursatz* (see Schenker, 1979); or the concept of *Tonnetz* (see Cohn, 1998)) that can structure the understanding of harmony. However, there is little agreement about specifics and arguments have raged since C.P.E Bach's caustic dismissal of Rameau (Kirenberg, 1774).

While contemporary music theory often associates music with relatively complex objects such as trees (Lerdahl & Jackendoff, 1983), graphs (Schenker, 1925) or orbifolds (Tymoczko, 2006), we will restrict our attention here to the simpler problem of categorizing chords. Given a large corpus of music fully annotated with a particular system of chord representation (surface tokens), our approach evaluates their possible groupings or clusters into a more coarse-grained set of categories. Surface tokens can consist of Roman numerals (Temperley, 2009; Tymoczko, 2003), scale degrees (De Clercq & Temperley, 2011; Huron, 2006) or simultaneous MIDI sonorities (Quinn & Mavromatis, 2011; Rohrmeier, 2005). We take these pre-existing representations as our starting point, without attempting to privilege any particular vocabulary. Though we will mainly be concerned with generalizing the concept of 'functional harmony', our approach is flexible enough to model theoretical concepts such as root motion, scale degrees, or the distinction between passing and stable chords.

*Correspondence:* Nori Jacoby, Hebrew University of Jerusalem, The Edmond & Lily Safra Center for Brain Sciences, Edmond J. Safra Campus, Givat Ram, Jerusalem, 91904 Israel. E-mail: jacoby@mit.edu

Formally:

**Definition 1. Deterministic categorization scheme.** *Let $\mathcal{C}$ be a list of surface tokens. Let $C_1, \ldots, C_N$ be a large corpus of music annotated with the symbols of $\mathcal{C}$.* A deterministic *categorization scheme (or a 'theory') is a mapping from $\mathcal{C}$ to a list of categories $\mathcal{F}$:*

$$F : \mathcal{C} \to \mathcal{F}$$

*The set of all surface token that maps to a single category is often called* a cluster.

Table 1 provides some examples of analysed corpora satisfying this definition.

Our main focus lies in developing a technique for evaluating various categorization schemes. We will always evaluate categories relative to an elementary representation; for example, we can compare theories A and B, or C and D in Table 1, but we cannot compare theories C and F, even though they describe the same music, as they relate to different elementary representations (surface tokens).

## 1.2 Criteria for theories

The ability of a theory to describe a musical surface must be testable. Therefore, *accuracy* is a crucial criterion in the evaluation of classification schemes. Of course, quantifying accuracy is non-trivial, because different theories can be accurate in different ways, and evaluating the match between theory and musical practice is itself theory-dependent, which raises the risk of a circular argument. In the context of our Definition 1, accuracy can be measured as the degree to which a coarse-grained categorization scheme represents a more refined surface structure (the surface tokens).

Our claim, however, is that accuracy is insufficient on its own: two theories might be equally accurate, but one theory could be simpler and thus preferable according to Occam's razor. For example, a strict scale-degree theory categorizes chords into seven categories, one for each diatonic tone, whereas the functional categorization into Tonic, Subdominant and Dominant (TSD) groups chords into just three categories. The key question is how to balance the increased accuracy of a seven-category system against its increased complexity.

In recent years, corpus analysis has been used in music scholarship to evaluate the real-world application of theoretical concepts (for a review see Temperley & VanHandel (2013)). Using statistical measures, one can empirically test how well a given theoretical concept describes a large body of digitally annotated scores. For example, Temperley (2009) showed that some root motions are more common than others in a corpus of harmony textbooks, a statistical relation that is predicted according to some theories (Meeùs, 2000; Sadai, Davis, & Shlesinger, 1980; Schoenberg, 1969). Tymoczko (2011) used a corpus of Bach chorales to determine which harmonic theory best described a given repertoire. Nevertheless, a fully developed methodology to evaluate the accuracy of functional categories has not yet been proposed (for efforts in this direction see Tymoczko (2003, 2011)).

Inspired by well-established methods in machine learning and information theory, we will provide quantifiable measures for these two important properties. One possibility that we will explore is defining accuracy by the amount of information lost in predicting neighbouring tokens when replacing a token with a category label and complexity as the amount of information required to code chords with category labels (Sections 1.6 and 1.7). We then show that these measures correlate with common musical intuitions and can contribute new answers to musically relevant questions (for example, to what extent can we apply nineteenth-century music theories to the analysis of music from earlier periods). We propose a method of comparing these different theories by introducing the 'evaluation plane', a mathematical framework graphically representing the accuracy and complexity of possible theories. Using this purely data-driven methodology, we then derive a class of 'optimal' theories, which reflect a continuum of optimal trade-offs between accuracy and complexity. This optimal class can serve as a baseline for comparing pre-existing harmonic categorization schemes.

## 1.3 Conceptual framework

Figure 1 shows a graph in which every theory is evaluated based on complexity and accuracy. We assume for now that both criteria are quantifiable by real positive numbers, with larger numbers reflecting higher complexity or accuracy. Every theory can then be mapped on a two-dimensional plane, where the complexity is the *x*-axis and accuracy the *y*-axis. Since two theories can have the same degree of complexity but differing degrees of accuracy, we should always prefer the more accurate theory (for example, we should prefer theory $F_B$ to $F_C$ in Figure 1). Conversely, given the same degree of accuracy, we should prefer a more parsimonious theory that reduces the amount of complexity (for example, theory $F_B$ is preferable to theory $F_D$ in Figure 1). This conceptual framework can also shed light on more ambiguous cases: if two theories have different accuracy and complexity measures, preferring one or the other depends on the evaluator's preferences regarding the trade-off between the two properties (theories $F_A$ and $F_E$ in Figure 1).

Furthermore there is a privileged class of theories among all possible theories: those that for a given level of complexity provide the maximal accuracy. This class of theories is indexed by the optimal complexity–accuracy curve, which indicates the maximal achievable accuracy for each complexity (see the black curve in Figure 1). A theory lies on the optimal curve if any other theory with the same or less complexity is less accurate; this means that there are no theories positioned above the optimal black curve in Figure 1. Note that this optimal class is not a single theory, but rather a continuum of possible theories characterized by their complexity. This curve is extremely difficult to calculate using brute-force methods, since it requires scanning an exponentially large number of

Table 1. Examples of corpora, surface tokens and theories according to Definition 1. Our formalism works with musical styles from classical to popular, and with different types of surface tokens, ranging from hand-made Roman analysis to unanalysed MIDI sonorities.

| | Corpus | Surface tokens | Name of theory | Clusters |
|---|---|---|---|---|
| A | Major-mode Bach chorales | 7 Diatonic scale degrees: I,ii,iii,IV,V,vi,$vii^o$ | TSD: Tonic Subdominant Dominant | T: I,vi,iii S: ii,IV D: V,$vii^o$ |
| B | Major-mode Bach chorales | 7 Diatonic scale degrees: I,ii,iii,IV,V,vi,$vii^o$ | Mmd: Categorization according to quality (Major, Minor diminished) | M: I,IV,V m: ii,ii,vi d: $vii^o$ |
| C | Major-mode Bach chorales | 40 Most common Roman numerals | Strict Root | 1: I, $I^6$, I6/4<br>2: ii6/5, ii,$ii^7$, $ii^6$, $ii^2$, ii4/3, ii6/4<br>3: iii, $iii^6$, $iii^7$, iii6/4, iii6/5, iii4/3<br>4: IV, $IV^6$, IV6/4<br>5: V, $V^7$,$V^6$, V6/5, $V^2$, V4/3, V6/4<br>6: vi, $vi^6$, $vi^7$, vi6/4, vi6/5, $vi^2$, vi4/3<br>7: $vii^{o6}$, $vii^{\o 7}$, $vii^o$, $vii^{\o}$4/3, $vii^o$6/4,$vii^{\o}$6/5,$vii^{\o 2}$ |
| D | Major-mode Bach chorales. | 40 Most common Roman numerals | Strict Bass | 1: I, $ii^2$, IV6/4, $vi^6$, vi6/5<br>2: ii, $ii^7$, V4/3, V6/4, $viio^6$, $vii^{\o}$6/5<br>3: iii,$I^6$, $iii^7$, vi6/4, vi4/3<br>4: IV, $V^2$,$ii^6$, ii6/5, $vii^{\o}$4/3, $vii^o$6/4<br>5: V, $V^7$,I6/4, $iii^6$, iii6/5, $vi^2$<br>6: vi, $vi^7$,$IV^6$,ii6/4, ii4/3, $vii^{\o 2}$<br>7: $vii^o$, $vii^{\o 7}$,$V^6$, V6/5,iii6/4, iii4/3 |
| E | 100 Rock songs From De Clerq and Temperley (2011) | 12 Scale degrees (no inversion): I,bII,II,bIII,III, IV,#IV,V,bVI,VI, bVII,VII | Diatonic and non Diatonic chords | Diatonic: I,II,III,IV,V,VI,VII Non-diatonic:bI,bII,#IV,bVI, bVII |
| F | Major-mode Bach chorales | 16 most common simultaneous transposed pitch classes extracted from MIDI renditions of Bach chorales (see Rohrmeier & Cross 2008) | Type of chord (Triads, seventh-chords, non-tertian). | Triads: CEG, GBD, AFD, DFA, DF#A, EGB, EG#B, BDF, AC#E<br>7th chords: GBDF, DF#AC, DFAC, ACEG<br>Non-tertian: CDG |

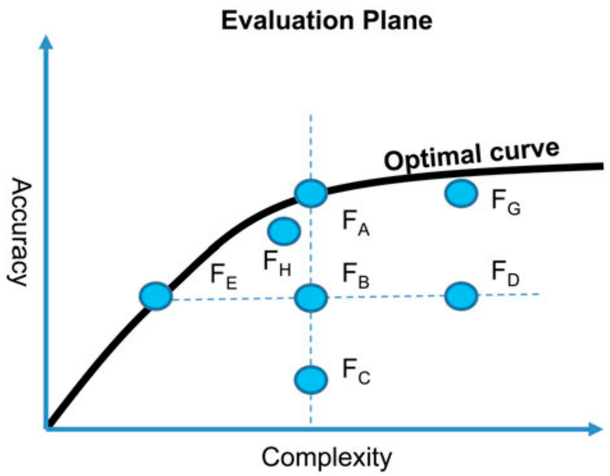

Fig. 1. The conceptual framework: the evaluation plane and the optimal curve.

possible theories. However, we will provide algorithms for solving this problem in a variety of different contexts, drawing on the work of Tishby, Pereira and Bialek (1999).

### 1.4 Concrete musical example

Let us give a concrete musical example:

Consider the following sequence of inversion-free Roman numerals:

$$I \to IV \to ii \to vii^o \to V \to vi \to I \to V \to I \qquad (1)$$

An assignment of categories in our formalism maps each surface token to a set of categories (other symbols), as in the following mapping to the set $\{\boldsymbol{T}, \boldsymbol{S}, \boldsymbol{D}\}$ (Tonic/Subdominant/Dominant):

$$F_{TSD}(I) = F_{TSD}(iii) = F_{TSD}(vi) = \boldsymbol{T};$$
$$F_{TSD}(ii) = F_{TSD}(IV) = \boldsymbol{S};$$

$$F_{TSD}(V) = F_{TSD}(vii^o) = \boldsymbol{D} \tag{2}$$

This categorization scheme maps our Roman numeral sequence to the following sequence of symbols

$$\boldsymbol{T} \rightarrow \boldsymbol{S} \rightarrow \boldsymbol{S} \rightarrow \boldsymbol{D} \rightarrow \boldsymbol{D} \rightarrow \boldsymbol{T} \rightarrow \boldsymbol{T} \rightarrow \boldsymbol{D} \rightarrow \boldsymbol{T} \tag{3}$$

thus implementing a version of standard North-American function theory, as for example articulated by Kostka and Payne (1984). This categorization scheme is illustrated in figure 2.

It is illustrative to contrast this theory with a toy theory that categorizes harmonies according to their intrinsic quality, minor, major and diminished (or {**M**, **m**, **d**}).

$$\begin{aligned} &F_{Mmd}(I) = F_{Mmd}(IV) = F_{Mmd}(V) = \mathbf{M};\\ &F_{Mmd}(ii) = F_{Mmd}(vi) = F_{Mmd}(iii) = \mathbf{m};\\ &F_{Mmd}(vii^o) = \mathbf{d} \end{aligned} \tag{4}$$

The sequence in Example 4 therefore maps to:

$$\mathbf{M} \rightarrow \mathbf{M} \rightarrow \mathbf{m} \rightarrow \mathbf{d} \rightarrow \mathbf{M} \rightarrow \mathbf{m} \rightarrow \mathbf{M} \rightarrow \mathbf{M} \rightarrow \mathbf{M} \tag{5}$$

Clearly, Examples 3 and 5 provide different information about the original sequence. While the two assignments use three symbols each, we might intuitively feel that Example 5 contains less information regarding the original musical content of Example 3. As we show later, this is indeed the case.

### 1.5 Graded or 'fuzzy' membership

Definition 1 requires that each surface token be associated with exactly one category. We can relax this requirement by allowing a single surface token to belong to more than one category in a fuzzy or 'graded' way (Figure 2; Agmon, 1995). To understand how this could be done, consider that the mappings of Equation 2 can also be written in probabilistic notation:

| $F_{TSD}$ | $T$ | $S$ | $D$ |
|---|---|---|---|
| $I$ | 1 | 0 | 0 |
| $ii$ | 0 | 1 | 0 |
| $iii$ | 1 | 0 | 0 |
| $IV$ | 0 | 1 | 0 |
| $V$ | 0 | 0 | 1 |
| $vi$ | 1 | 0 | 0 |
| $vii^o$ | 0 | 0 | 1 |

(6)

Each entry in the table represents the weight of a Roman numeral's membership in the appropriate category. For example: $p(F_{TSD} = \boldsymbol{T}|C = I) = 1$, which says that chord (token) *I* is mapped to category *T* (Tonic function) with a weight of 100%; or $p(F_{TSD} = \boldsymbol{D}|C = IV) = 0$, which says that *IV* does not belong to ***D*** at all. The advantage of this notation is that it permits nondeterministic mappings whereby a single chord, such as *iii*, can belong to multiple functional categories with arbitrary weights. We can thus write any functional mapping, deterministic or not, as a matrix $p(F = f|C = c)$, where *c* ranges over our surface tokens and *f* ranges over all possible functions.

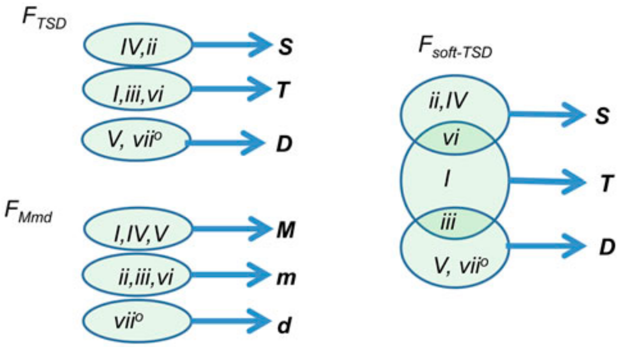

Fig. 2. Categorization schemes. $F_{TSD}$ uses the standard categories tonic, subdominant and dominant while $F_{Mmd}$ categorizes according to triad quality. $F_{soft\text{-}TSD}$ offers a more sophisticated version of function theory where chords can map to multiple functions.

For example, consider the following probabilistic 'soft' clustering:

| $F_{soft\text{-}TSD}$ | $T$ | $S$ | $D$ |
|---|---|---|---|
| $I$ | 1 | 0 | 0 |
| $ii$ | 0 | 1 | 0 |
| $iii$ | 0.5 | 0 | 0.5 |
| $IV$ | 0 | 1 | 0 |
| $V$ | 0 | 0 | 1 |
| $vi$ | 0.5 | 0.5 | 0 |
| $vii^o$ | 0 | 0 | 1 |

(7)

Here the *iii* token (chord) is represented as being 50% tonic and 50% dominant, while the *vi* chord is 50% tonic and 50% subdominant. This is consistent with the familiar idea that *I*, *IV* and *V* are prototypes of the Tonic and Subdominant and Dominant categories, and that the other chords ($ii$, $iii$, $vi$) are more loosely associated with one or more categories (Agmon, 1995)[1]. This leads to the following generalization of Definition 1:

**Definition 2. Probabilistic categorization scheme.** *Let* $\mathcal{C}$ *be a list of surface tokens. Let* $C_1, \ldots, C_N$ *be a large corpus of music annotated with the symbols of* $\mathcal{C}$*.* A probabilistic categorization scheme *on a set of labels* $\mathcal{F}$ *(which we term a 'theory') is a random variable F defined by the conditional probabilities* $p(F = f|C = c)$ *where* $f \in \mathcal{F}$ *and* $c \in \mathcal{C}$*.*

[1] Note that the important idea of graded or 'fuzzy' membership of chords in functional categories (Agmon, 1995) is formalized here by using the language of probabilities and random variables. While this is not the only approach, there is a long tradition of using this particular formalism for this purpose in the machine-learning community (Hastie, Tibshirani, Friedman, & Franklin, 2005) and music (Temperley, 2007). The problem of finding such categorizations is often described in the machine-learning literature as 'distributional clustering' (Pereira, Tishby, & Lee, 1993). As we will see, this definition will be instrumental for the rest of the theory developed here.

### 1.6 Quantifying complexity

Complexity, as its name suggests, is related to the number of categories we use; indeed, an intuitive definition of complexity is given by the number of categories in our theory. However, in some cases we might want to distinguish between frequent and very infrequent categories: for example, if a category appears only extremely rarely in the totality of all our data (say once in a million tokens) we might want to say that our theory is roughly as complex as a theory in which that category is simply not used.

To account for rare tokens, entropy is often used; for equally likely categories, this quantity is simply the logarithm of the number of categories. However, when categories are not equally likely, rare categories contribute less than common categories (for further applications of entropy in music see Temperley (2007)).

$$H(F) = -\sum_{f \in \mathcal{F}} p(F = f) \log_2 p(F = f). \tag{8}$$

Note that this measure of complexity is *data relative* and cannot be inferred simply from the theory itself.

For technical reasons, it is sometimes useful to consider *mutual information*, a quantity that is closely related to entropy and is central to information theory (Shannon, 2001). Formally, mutual information $I(F; C)$ is defined as:

$$I(F; C) = H(F) - H(F|C), \tag{9}$$

where

$$H(F|C) = -\sum_{c \in C} p(C = c) \times \left( \sum_{f \in \mathcal{F}} p(F = f|C = c) \log_2 p(F = f|C = c) \right). \tag{10}$$

Mutual information is simply the entropy minus a term that that is due to the fuzziness of the classification scheme, $H(F|C)$. In the case of deterministic mapping $H(F|C) = 0$ and then entropy and mutual information are identical. Mutual information is measured in bits and captures the amount of relevant information that our functional categories retain from the original scheme. Mutual information is symmetric $I(F; C) = I(C; F)$ and non-negative $I(F; C) \geq 0$ (Cover & Thomas, 2012).

Table 2 lists our three options for defining complexity: counting symbols, simple entropy and mutual information. Note the more compact formula for mutual information in Table 2, case C which can be easily derived from Equations 8–10. This definition requires knowledge of the marginal distribution of the surface chords $p\,(C = c)$, which can be computed from the empirical histogram of chords in the corpus.

All three measures in Table 2 are non-negative, and therefore comply with the requirements discussed in Section 1.3. Since mutual information is well known and easy to work with[2], we will favour it – though we also use the two other alternatives (which often produce similar results).

Note that Mavromatis (2009, 2012) introduced complexity to the music community through the concept of minimum description length as a metric for estimating the number of clusters in a Hidden Markov Model (HMM). In Section 2.6 we compare our method to the HMM approach.

### 1.7 Quantifying accuracy

There are multiple methods and for defining accuracy. In principle we could apply the formalism in Figure 1 to a wide class of accuracy metrics. However, in this paper, we mostly associate accuracy with *prediction* – that is, our ability to infer something about the musical stimulus based only on functional information. The thought here is that functional labels are often used to specify grammatical rules or statistical tendencies: if we know that the current chord in a classical piece is a dominant, say, then we have a pretty good idea that the next chord will be a tonic. Note that in later parts of the work we will compare the predictive approach to other alternatives (Sections 2.5 and 2.6).

Let us illustrate the approach by recalling Example 1:

$$I \rightarrow IV \rightarrow ii \rightarrow vii^o \rightarrow V \rightarrow vi \rightarrow I \rightarrow V \rightarrow I \tag{11}$$

Assume that we replace one surface token (the second *V*) with the category label associated by $F_{TSD}$ in Example 3; the new sequence is:

$$I \rightarrow IV \rightarrow ii \rightarrow vii^o \rightarrow \mathbf{D} \rightarrow vi \rightarrow I \rightarrow V \rightarrow I \tag{12}$$

We might try to measure accuracy by measuring the amount of information lost by this replacement. Formally, let $X$ be the current token $C_n$, let $F$ be the current category $F_n$ and let $Y$ be the random variable associated with the context or 'all other tokens' (see also Figure 3):

$$Y = (C_1, C_2, \ldots, C_{n-1}, C_{n+1}, C_{n+2}, \ldots) \tag{13}$$

and:

$$X \equiv C_n, \tag{14}$$
$$F \equiv F_n. \tag{15}$$

Accuracy would then correspond to the mutual information between $F$ and $Y$:

$$I(F; Y) = H(Y) - H(Y|F). \tag{16}$$

If $F$ is a one-to-one mapping (where each symbol is mapped to itself, and there is no reduction), the mutual information attains the maximal value, or $I(F; Y) = I(X; Y)$. If $F$ maps all surface tokens into one symbol (all information is lost) the mutual information attains the minimal value $I(F; Y) = 0$. All other mappings of surface tokens to categories (as in

[2] As we develop our formalization, we will notice further advantages of using mutual information. For example, this choice significantly simplifies some of the algorithmic steps.

Table 2. Three possible definitions of complexity ($I_C(F)$).

| | Name of complexity measure | Symbol | Formal Definition |
|---|---|---|---|
| A | Number of labels | $I_C(F) = \lvert\mathcal{F}\rvert$ | Number of symbols in $\mathcal{F}$ |
| B | Entropy | $I_C(F) = H(F)$ | $H(F) = -\sum_{f\in\mathcal{F}} p(F = f)\log_2 p(F = f)$ |
| C | Mutual information | $I_C(F) = I(F;C)$ | $(I(F;C) = H(F) - H(F\|C) = H(C) - H(C\|F)$<br>$= \sum_{f\in\mathcal{F},c=\mathcal{C}} p(F = f\|C = c)P(C = c)\log_2 \frac{p(F=f\|C=c)}{p(F=f)}$ |

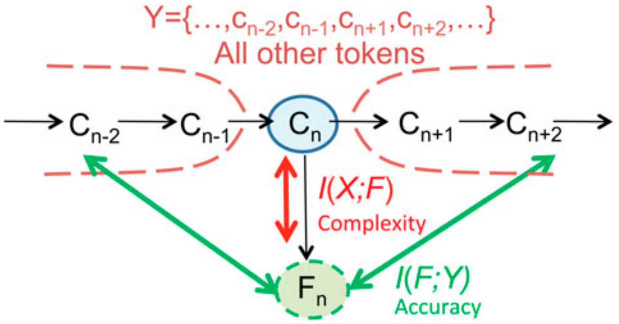

Fig. 3. Accuracy as mutual information between the current category and all other surface tokens $I(F;Y)$.

Definition 2) would have intermediate $I(F;Y)$ values (Cover & Thomas, 2012).

In practice, however, $Y = (C_1, C_2, \ldots, C_{n-1}, C_{n+1}, C_{n+2}, \ldots)$ is a very high-dimensional vector, making it impossible to compute the mutual information $I(F;Y)$ directly. For this reason further approximations are needed. One approach is to consider only those chords in temporal proximity to the current chord $C_n$, since they can be expected to exert greater influence on the music. In this case, we replace $Y$ with $Y'$ representing the local context of the current chord ($X = C_n$). The validity and extent of this assumption is an important question on its own, which we further explore after we fully develop our formalism (see Section 2.5). Note that these definitions generate one number $I_a(F)$, which estimates the average accuracy over all possible chords.

Table 3 presents some possible definitions of local context. Figures 4 and 5 describe these alternatives graphically.

In Figure 4(a) we consider a local context to be the next chord. (This captures one traditional motivation for function theory, namely specifying first-order grammatical tendencies or rules.) For example, if $Y' = Y_{n+1}$ we effectively evaluate the mutual information based on the distribution of chords bigrams and ignore all higher order structures. Table 4(a), (d) and (e) model the assumption that $C_n$ is a first, second and third order Markov chain, respectively (see Tishby et al., 1999).

Figure 4(b) represents a local context as the *previous* token. We can also consider the local context as containing the next token and the previous token (Figure 4(c)), the next two tokens (Figure 4(d)) or the next three tokens (Figure 5(e)). Indeed, there are analogous definitions for any choice of local context $Y'$. Note, however, that the number of states that need to be considered to compute the mutual information is exponential in the length of $Y'$; thus one cannot expect to have enough data to properly evaluate the accuracy for long $Y'$ unless using more sophisticated methods (as suggested by Pearce and Wiggins (2006)).

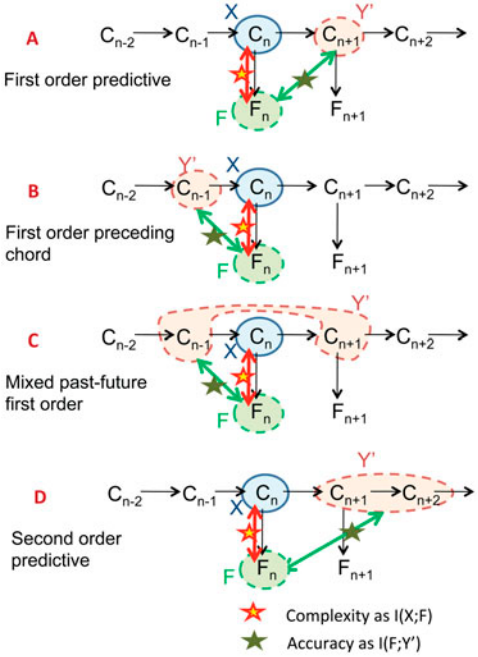

Fig. 4. Possible definitions of the accuracy of a theory (see Table 3, parts A–D).

In Figures 4(a)–(d) and 5(e), the accuracy is defined by $I(F;Y') = H(Y') - H(Y'|F)$, where $Y'$ changes with the context. Figure 5(f) presents a slightly different approach, in which we try to estimate how well a category predicts the other local *categories*. The motivation here is that our earlier approaches could be associated with assertions such as ‘dominant chords tend to go to I chords with a frequency of X%, to I$^6$ chords with a frequency of Y%, …’. By contrast Figure 5(f) provides an alternative that is more aligned with traditional function theories, which often attempt to predict the next function (the next category) rather than the chord (the next token) itself. This approach is associated with statements such as ‘dominant chords tend to go to tonic chords’. As we will see, our formalism works with all of these alternative

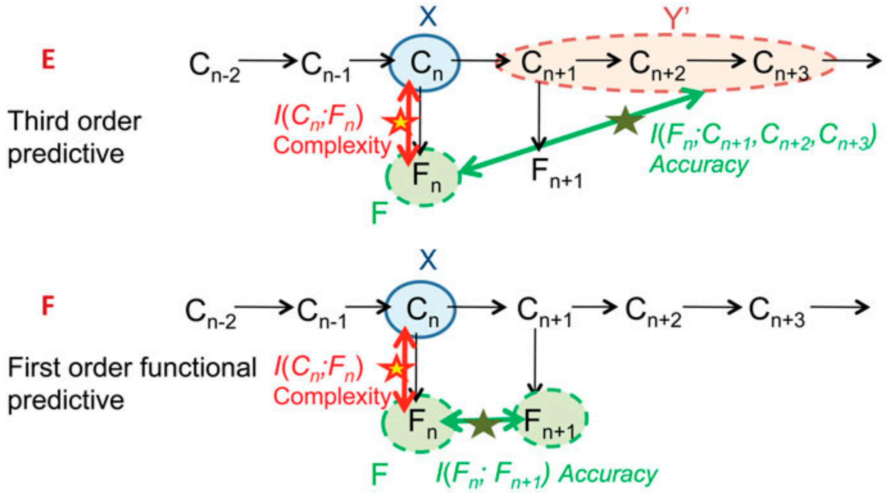

Fig. 5. More possible definitions of the accuracy of a theory (see Table 3, parts E–F).

Table 3. Possible definitions of the accuracy of a theory $I_a(F)$.

| | Name of accuracy measure | Formal definition |
|---|---|---|
| A | First order predictive (‘Predictive power’) | $I_a(F) = I(F; Y')$; where $Y' = C_{n+1}$, $F = F_n$ |
| B | First-order preceding chord (‘Time reversed’) | $I_a(F) = I(F; Y')$; where $Y' = C_{n-1}$, $F = F_n$ |
| C | Mixed past-future first order | $I_a(F) = I(F; Y')$; where $Y' = (C_{n+1}, Cn - 1)$, $F = F_n$ |
| D | Second-order predictive | $I_a(F) = I(F; Y')$; where $Y' = (C_{n+1}, Cn + 2)$, $F = F_n$ |
| E | Third-order predictive | $I_a(F) = I(F; Y')$; where $Y' = (C_{n+1}, Cn + 2, Cn + 3)$, $F = F_n$ |
| F | First-order functional predictive clustering (pairwise clustering) | $I_a(F) = I(F_n; F_{n+1})$ |

definitions, though they require slightly different tools when computing optimal theories.

### 1.8 The evaluation plane

Tables 2 and 3 show multiple ways of defining complexity and accuracy, respectively. Assuming we pick one of the definitions for complexity and one of the definitions for accuracy, we can now formally define the ‘evaluation plane’ of figure 1.

**Definition 3. Evaluation plane.** *The evaluation plane is a two-dimensional graph in which the x- and y-axes represent the complexity and accuracy of all possible theories. Each theory F of Definition 2 can be located on a point on the plane defined by* $(x, y) = (I_c(F), I_a(F))$*, where* $I_c(F)$ *and* $I_a(F)$ *are the complexity and accuracy metrics associated with F, respectively.*

For example, suppose we want to compare the two theories in Definition 2, $F_1$ and $F_2$. We start by acquiring a large corpus of music annotated with surface tokens $\mathcal{C}$. We now pick our favourite definitions of accuracy and complexity from Tables 2 and 3, respectively. We compute $p(C)$, $p(Y')$, $p(F_1)$ and $p(F_2)$ along with $p(F_1|C)$, $p(F_2|C)$, $p(Y'|F_1)$ and $p(Y'|F_2)$. (These last distributions are required for quantifying complexity and accuracy and can be directly computed from the corpus by generating the appropriate histograms.) We can then position $F_1$ and $F_2$ on the information plane by computing $(I_c(F_1), I_a(F_1))$ and $(I_c(F_2), I_a(F_2))$. If $I_c(F_1) \approx I_c(F_2)$, we can choose a theory with a larger $I_a(F)$ (see $F_A$ and $F_B$ in Figure 1 for a schematic representation of this situation). Similarly, if $I_a(F_1) \approx I_a(F_2)$ we can choose the theory with smaller $I_c(F)$ (as in the case of $F_B$ and $F_E$ in Figure 1).

In the general case where all accuracy and complexity scores are different, we can use a combined score or compute the optimal curve.

### 1.9 Using a combined score

If we have a specific relative weighting of accuracy or complexity (parameterized by the constant $\lambda \geq 0$), we can try to identify the theory with maximal combined score:

$$\mathcal{L}(F) = I_a(F) - \lambda I_c(F), \tag{17}$$

where the constant $\lambda$ tells us the relative weight of maximizing accuracy compared to the relative weight of minimizing complexity in the combined score $\mathcal{L}$. This combined score is also often called the Lagrangian of the complexity–accuracy tradeoff (Tishby et al., 1999). Appendix A describes an algorithm for finding a theory with a maximal combined score, which solves the following equation:

$$F = \text{argmax}_{\text{all possible } F} \mathcal{L}(F). \tag{18}$$

### 1.10 Computing the optimal curve

We can now define a privileged class of *optimal theories*, which are maximally accurate for a given complexity.

**Definition 4. Optimal theory.** *$F$ is an* optimal theory *if for every other theory $F'$ such that $I_c(F') \leq I_c(F)$ then $I_a(F') \leq I_a(F)$.*

In Figure 1, theory $F_A$ would be optimal if all theories with less complexity (including $F_B$, $F_C$, $F_H$ and $F_E$) have less accuracy.

**Definition 5. Optimal curve; the optimal curve problem.** *The set of points on the evaluation plane corresponding to all optimal theories is called the* optimal curve. *Finding a theory on the optimal curve with a complexity less or equal to $I_c^0$ is called the* optimal curve problem*:*

$$F = \text{argmax}_{I_c(F) \leq I_c^0} I_a(F).$$

From Equation 18, a theory with a maximal combined score for a given value of λ is always on the optimal curve. If measure complexity as mutual information (Table 2, case C) and accuracy using any of the first five measures in Table 3, then the set of all realizable points under the optimal curve (the set of all points $(a, c)$ in the evaluation plane for which there exists a theory $F'$ such that $(a, c) = (I_a(F'), I_c(F'))$ is convex and dense (Tishby et al., 1999). This means that for any two points representing theories on the evaluation plane, there exist other theories realizing the entire line connecting the two points. The mathematical property of convexity is desirable because it greatly simplifies the optimal curve problem (see Appendix A); this is a core reason why we measure complexity and accuracy using mutual information.

We can use the optimal curve to evaluate theories, rejecting those that are far from the optimal curve ($F_D$, $F_B$ and $F_C$ on Figure 1) in favour of those that are optimal ($F_A$ and $F_E$) or near-optimal ($F_G$ and $F_H$). Alternatively, we might investigate optimal theories in their own right, since they provide *self-emergent* categories (clusters) of chords. Previous writers have proposed various other methods for identifying functional categories from corpus data (Quinn & Mavromatis, 2011; Rohrmeier, 2005; Rohrmeier & Cross, 2008). One of our contributions here is to provide a principled framework that can solve this problem for any surface structure.

### 1.11 Mapping the optimal curve problem to the machine learning literature: solving the optimal curve problem

Finding the optimal curve for a given corpus is a well-known problem in machine learning. If we choose the complexity to be $I(F; X)$ (as in Table 2, case C) and the accuracy to be any of the measures of Table 3, cases A–E, then the problem is known as an 'Information Bottleneck' problem, the evaluation plane is known as the 'information plane' and the optimal curve is known as the 'information curve' (Tishby et al., 1999; for applications see Friedman, Mosenzon, Slonim, & Tishby, 2001; Hecht, Noor, & Tishby 2009; Schneiderman, Slonim, Tishby, de Ruyter van Steveninck, & Bialek, 2002; Slonim & Tishby, 2000). The accuracy measure $I(C_{n+1}; F_n)$ of Table 3, case A is known as the 'first order predictive-power' or simply the 'predictive power'.

In this case, an iterative algorithm proposed by Tishby et al. (1999) can compute the optimal curve effectively for problems with less than a few thousand surface tokens. This algorithm is described in Appendix A, with corresponding code and web interface provided online (cluster.norijacoby.com). The algorithm is highly non-trivial and is far more effective than the naíve approach that computes the accuracy and complexity for the infinitely large family of all possible theories.

If, on the other hand, we use the number of categories (Table 2, case A) as our complexity measure, using any accuracy metrics of Table 3, cases A–E for the accuracy (metric 3F requires special treatment), then the problem can be solved by modifying the Tishby et al. (1999) algorithm. In this variant, we apply the same iterative process but further constrain the solution space so that it has the desired number of categories. We refer to this procedure as finding 'deterministic' optimal theories; the full details are described in Appendix A. Finally, if we use the complexity measure $I(C_n; F_n)$ of Table 2, case C and the accuracy measure $I(F_n; F_{n+1})$ of Table 3, case F, the problem is known as 'pairwise-clustering' (Friedman & Goldberger, 2013). Our suggested algorithm can be found in Appendix A.

### 1.12 Summary: computing optimal curves

To summarize, the steps involved in computing the optimal curve are:

(a) We choose a corpus of music (such as Bach chorales).
(b) We choose some elementary representation of surface tokens (for example Roman numerals with or without inversions, or raw MIDI sonorities).
(c) We choose complexity and accuracy metrics from Tables 2 and 3, respectively. For example, we choose the complexity to be $I(X; F) = I(C_n; F_n)$ and the accuracy to be $I\left(X; Y'\right) = I(C_{n+1}; F_n)$.
(d) We compute the joint probability $p(X, Y')$. For example, for the choices in (c) this is computable from the pairwise histogram $Y' = (C_n, C_{n+1})$.
(e) We compute a large sample of optimal functional theories (probabilistic categorization schemes) using the algorithms in Appendix A, plotting the accuracy and complexity of these theories on the information plane (as in Figure 1).

Table 4. Categories obtained from a simple MIDI dataset.

| | Name of category | Type of category | Categories |
|---|---|---|---|
| A | Optimal deterministic theory with 2 cluster | Self-emergent | Category 1: C\|CEG, E\|CEG, A\|CFA, A\|CEA, F\|CDFA, G\|CDG, G\|CEG, F\|CEFA, C\|CEGBb, B\|CEGB, Bb\|CEGBb, C\|CE<br>Category 2: G\|DGB, F\|CFA, G\|DFGB, D\|DFB, B\|DGB, C\|CFG, D\|DFA, B\|DFGB, F\|DFGB, G\|DGBb |
| B | Optimal deterministic theory with 3 cluster | Self-emergent | Category 1: C\|CEG, E\|CEG, A\|CEA, G\|CEG, C\|CEGBb, B\|CEGB, Bb\|CEGBb, C\|CE<br>Category 2: F\|CFA, A\|CFA, F\|CDFA, G\|CDG, D\|DFA, F\|CEFA, F\|DFGB<br>Category 3: G\|DGB, G\|DFGB, D\|DFB, B\|DGB, C\|CFG, B\|DFGB, G\|DGBb |
| C | Optimal deterministic theory with 7 cluster | Self-emergent | Category 1: C\|CEG<br>Category 2: E\|CEG, G\|CEG, C\|CEGBb , C\|CE<br>Category 3: D\|DFA,F\|CFA, D\|DFB, F\|DFGB, G\|DGBb<br>Category 4 : F\|CDFA, G\|CDG, F\|CEFA<br>Category 5 : G\|DGB<br>Category 6: G\|DFGB, B\|DGB, C\|CFG, B\|DFGB<br>Category 7: A\|CFA, A\|CEA, B\|CEGB, Bb\|CEGBb |
| D | $F_{TSD}$. Tonic/Dominant/ Subdominant | Pre-determined | Category T: C\|CEG, E\|CEG, A\|CEA, C\|CFG, G\|CEG, B\|CEGB, C\|CE<br>Category S: F\|CFA, A\|CFA, F\|CDFA, D\|DFA, F\|CEFA<br>Category D: G\|DGB, G\|DFGB, D\|DFB, B\|DGB, G\|CDG, B\|DFGB, F\|DFGB, G\|DGBb<br>Category other: C\|CEGBb, Bb\|CEGBb |
| E | $F_{Mmd}$ Major/Minor/ diminished triads or others | Pre-determined | Category Major tirads: C\|CEG, G\|DGB, F\|CFA, E\|CEG, A\|CFA, B\|DGB, G\|CEG, C\|CE<br>Category minor triads: A\|CEA, D\|DFA, G\|DGBb<br>Category diminished triads: D\|DFB<br>Category other: G\|DFGB, F\|CDFA, C\|CFG, G\|CDG, B\|DFGB, F\|CEFA, C\|CEGBb, F\|DFGB, B\|CEGB, Bb\|CEGBb |
| F | $F_{\mathrm{root}}$ Root based categorization | Pre-determined | Category 1: C\|CEG, E\|CEG, C\|CFG, G\|CEG, C\|CEGBb, B\|CEGB, Bb\|CEGBb, C\|CE<br>Category 2: F\|CDFA, D\|DFA<br>Category 3: F\|CFA, A\|CFA, F\|CEFA<br>Category 4: G\|DGB, G\|DFGB, B\|DGB, G\|CDG, B\|DFGB, F\|DFGB, G\|DGBb<br>Category 5: A\|CEA<br>Category 6: D\|DFB |
| G | $F_{\mathrm{bass}}$ Categorization according to bass | Pre-determined | Category 1 A: A\|CFA, A\|CEA, Bb\|CEGBb<br>Category 2 B: B\|DGB, B\|DFGB, B\|CEGB<br>Category 3 C: C\|CEG, C\|CFG, C\|CEGBb, C\|CE<br>Category 4 D: D\|DFB, D\|DFA<br>Category 5 E: E\|CEG<br>Category 6 F: F\|CFA, F\|CDFA, F\|CEFA, F\|DFGB<br>Category 7 G: G\|DGB, G\|DFGB, G\|CDG, G\|CEG, G\|DGBb |

(f) We plot pre-existing categorizations of interest (for example $F_{TSD}$), and measure their distance from the optimal curve, or compare them to each other.

(g) Using an algorithm from Appendix A, we compute the self-emergent deterministic optimal theories that use $k$ categories (optimal deterministic $k$-categorization schemes) and position them on the evaluation plane. These deterministic optimal theories are usually found very near the unconstrained optimal theories of the optimal curve.

## 2. Corpus results and comparisons with alternative methods

### 2.1 Using the framework on a simple corpus

The following examples are intended to show that our framework can model important music-theoretical concepts in the context of real-world corpora. We focus mainly on surface tokens that are manually annotated Roman numerals. However, in Sections 2.3–2.4 we consider a broader spectrum of data including MIDI-based corpora. In the current section we also focus on the Information Bottleneck accuracy and complexity measures (Table 2, case C and Table 3, case A), $I(F; X)$, $I(Y'; F)$ with $Y' = C_{n+1}$ and $X = C_n$, since these choices are standard in the machine-learning community. A comparison of different accuracy and complexity measures is provided in Section 2.5.

Let us now return to our earlier examples (Sections 1.4–1.5), and plot them on a curve computed from a dataset of actual music. We apply our framework to corpus data compiled from Tymoczko (2011), which records harmonic progressions in major-mode passages from 70 Bach chorales. (Note that unlike some of the later cases we will consider, this dataset ignores chord inversions and uses just seven surface tokens – the familiar Roman numerals – to label major-mode harmonic progressions; note also that in constructing the dataset, Tymoczko regarded I6/4 chords as V, unlike David Huron in the dataset in Section 2.4 below, who regarded I6/4 chords as I.) The algorithm of Appendix A only needs the empirical distribution of consecutive chords $p(C_{n+1}, C_n)$ as input which is computable from Tymoczko's (2011, p. 230) Figure 7.1.6.

Figure 6 shows the evaluation plane and the optimal curve. The black curve represents the optimal trade-off between complexity and accuracy, computed using the iterative algorithm in Appendix A. For every possible theory $F$ (satisfying Definition 2) the point $(I_c(F), I_a(F)) = (I(F; X), I(F; Y')) = (I(F_n; C_n), I(F_n; C_{n+1}))$ lies *below* this curve. The horizontal line at the top of the curve represents the mutual information between the current and following chord, which is the upper limit of $I\left(X; Y'\right) = I(C_n; C_{n+1})$. Figure 6 shows three points associated with optimal deterministic theories with two, three and four categories (clusters). This optimal categorization was computed using a variant of the algorithm where we limit the number of categories (see the Appendix and Slonim and Tishby (2000)). These clusters are entirely self-emergent, in that they are determined solely by the probabilities $p(C_n, C_{n+1})$ in the Bach corpus.

Figure 6 is interesting for several reasons. First, many familiar ways of thinking about harmony lie at or are extremely close to the optimal curve. The optimal categorization into two categories corresponds to 'dominant' and 'not dominant'. Even more remarkably, the optimal assignment to three categories coincides with $F_{TSD}$, the textbook Tonic, Subdominant, and Dominant classification of Equation 2. Note by contrast, that the $F_{Mmd}$ of Equation 4 is positioned significantly below the optimal curve; indeed it is significantly less accurate than the optimal two-symbol clustering ($F_{AB}$). Furthermore, the classification into four categories, while relatively familiar, contains an interesting music-theoretical wrinkle, grouping *I* and *iii* as tonics, *V* and $vii^o$ as dominants, *IV* and *vi* as subdominants, while leaving *ii* in its own category. It is surprising that *vi* resembles *IV* more than *ii* does, suggesting an interesting topic for further music-theoretical research. (One thought is that *vi* and *IV* both can move to *I* in progressions like $vi \rightarrow I^6$ or $IV^6 \rightarrow I$, while $ii \rightarrow I$ progressions are quite rare.) Note that the categorization $F_{soft\text{-}TSD}$, shown in Equation 7, performs similarly to $F_{TSD}$, and is only slightly off the optimal curve.

## 2.2 Using the framework: categorization according to root and bass

This section explores another application of our framework. Figure 7 depicts two simple theories with similar degrees of complexity: the first classifies triads and seventh chords according to their roots, while the second classifies them according to their bass note (see table 1c and 1d). We use a new dataset (dataset 12 from Table B3) drawn from Tymoczko's handmade analyses of all 371 Bach chorales, where the surface tokens combine Roman numerals and figured bass symbols, so that $I^6$ and $I5/3$ are distinct. Somewhat surprisingly, the fundamental-bass theory is slightly more accurate than the root-functional theory, whereas the root-functional approach is significantly simpler. The accuracy of fundamental-bass theory reflects the fact that there are significant regularities in tonal bass lines not captured by functional information (for instance, bass lines tend to move stepwise or by fifth). The simplicity of the root-functional theory is related to the rarity of the *iii* chord, which constitutes only 0.8% of the chords in the corpus. Crucially, however, one gains only a modest amount of accuracy when moving from root-function to fundamental-bass (0.047 bits, from 0.62 to 0.66, which is 3.94 % of the 1.2 bits, the total mutual information). However, the change in the complexity is significantly greater: 0.44 bits (from 2.3 to 2.7) or 11% of the maximal complexity (the entropy $H(X) = I(X; X)$). Furthermore, both classifications lie quite a bit below the optimal curve, far from the optimal deterministic seven-category scheme:

| | |
|---|---|
| class T1: | I, vi, vi$^6$, vi6/4, vi$^7$, vi$^2$, iii, iii$^6$, iii$^7$, iii6/5, iii4/3 |
| class T2: | I$^6$, I6/4, ii6/5 |
| class S1: | IV, ii, , ii$^6$, ii6/4, ii$^7$, vi6/5, vi4/3 |
| class S2: | IV$^6$, IV6/4, ii4/3, ii$^2$, vii$^{ø2}$ |
| class D1: | V |
| class D2: | V$^6$, V$^7$, V6/5, vii$^o$, vii$^ø$7, iii6/4 |
| class D3: | V6/4, V4/3, V$^2$, vii$^{o6}$, vii$^{o6/4}$, vii$^ø$6/5, vii$^ø$4/3 |

At first glance, this classification might seem to show that information-theoretic ideality diverges from musical intuition. But on further reflection one can see the outline of familiar functional ideas: classes T1 and T2 are tonic chords, as our labels suggest, with the T1 containing the root position tonic, and most triadic and seventh-chord inversions of vi and iii. T2 contains the other inversions of the tonic chord (perhaps suggesting that from an information-theoretic point of view I6/4 is more tonic than dominant suspension), and – rather surprisingly – the ii6/5 chord. Class S1 and S2 are basically subdominants, with S1 containing the prototypical subdominants ii, ii$^7$, IV, ii$^6$ and S2 containing only chords with $\hat{6}$ or $\hat{1}$ in the bass. (Note again the surprising presence of vii$^{ø2}$ among the subdominants of S2; this is likely because it often progresses to V$^7$ by way of I6/4.) Finally, D1, D2, and D3 are basically dominant chords with V the sole occupant of its category, chords on $\hat{5}$ and $\hat{7}$ occupying D2, and chords on $\hat{2}$ and $\hat{4}$ occupying D3. In one sense, then, this solution is telling us something we already knew, namely that an ideal seven-category system would group chords using *both* root-functional and fundamental-bass principles. More interesting is the fact that the algorithm actually shows us how to do this, producing a set of categories that no human would devise, yet which make a certain amount of retrospective sense. It could therefore prompt analytical work that helps us appreciate the virtues of this particular functional scheme.

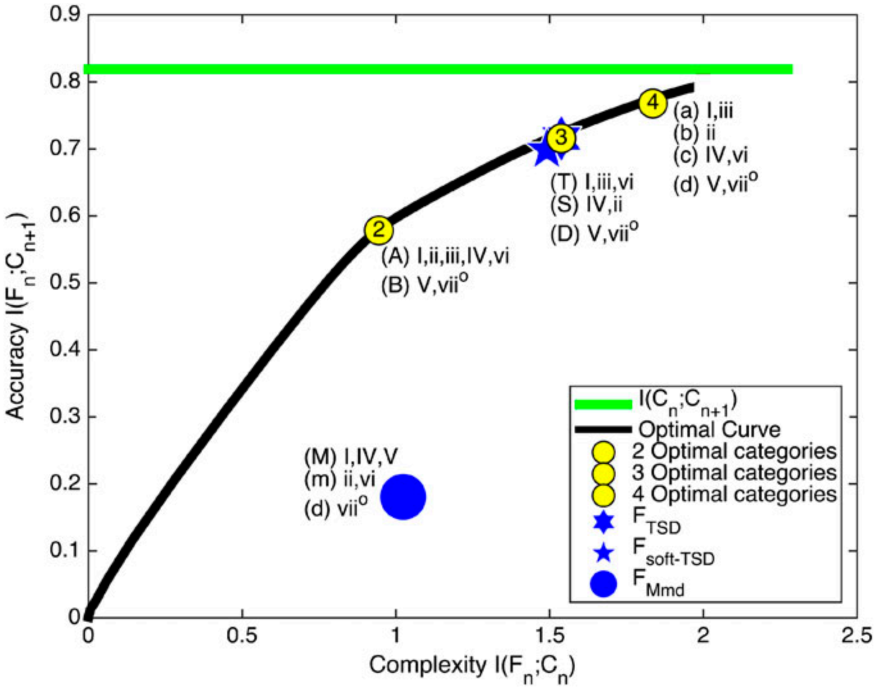

Fig. 6. Comparison of different categorization schemes in a dataset of major-mode passages from 70 Bach chorales in Tymoczko (2011a, §7. 1).

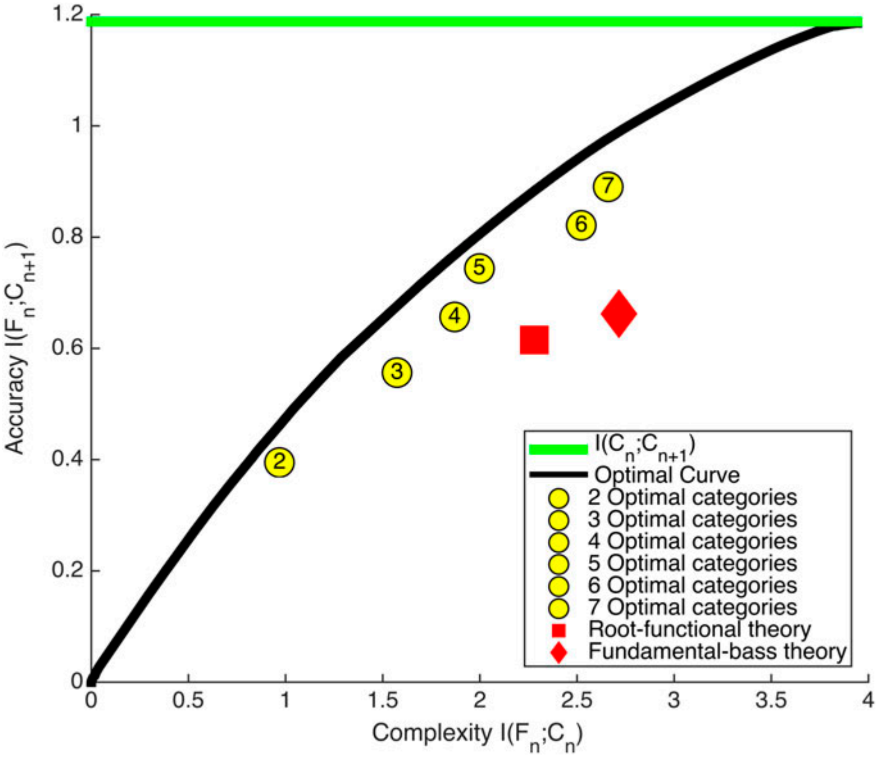

Fig. 7. Root function and bass theories compared on the evaluation plane with deterministic optimal theories, based on a new dataset of all major-mode passages in all Bach chorales (analysed by Tymoczko).

### 2.3 Using the method for analysing fully automatic MIDI data

So far we have worked with manually labelled data. The following example shows how one can use our methods in the context of fully automatic analysis. The point is, first, to show that our methodology can work with minimal assumptions, and second, to compare the results with those of the previous section.

Beginning with MIDI rendition of the Bach chorales, we reduced each chorale to a sequence of vertical sonorities or 'slices'. (The same procedure was used also by Quinn and Mavromatis (2011) and White and Quinn (2014) and implemented in *music21* (Ariza & Cuthbert, 2010).) We adopted the key-finding methodology of White and Quinn (2014), who used a Krumhansl–Schmuckler key-finding algorithm on a windows of eight slices, so that each slice was analysed eight times. If a slice was assigned to multiple keys, we selected the key with the highest score. For simplicity, we only analysed major segments. Each slice was then transposed to C major. For each transposed slice we recorded its pitch classes together with the pitch class of the bass note. Therefore a non-inverted tonic chord (*I* in the local key) would be recorded as 'C| C E G'; this says that the bass is C and that the pitch classes in this chord are C, E and G. We focused on the 22 most common sonorities (those that comprise more than 1% of the dataset). The results of this analysis (including optimal and pre-existing categorizations) are detailed in table 4 and figure 8.

The optimal three-category deterministic theory was, when translated to Roman numerals:

| | |
|---|---|
| Category 1: | I, I$^6$, vi, I6/4, V/IV, I$^2$, V$^2$/IV, I (no fifth) |
| Category 2: | IV, IV$^6$, ii6/5, V$^4$, ii, IV$^7$, V$^2$ |
| Category 3: | V, V$^7$, viio$^6$, V$^6$, I$^4$, V6/5, v |

This is clearly very similar to the standard tonic–subdominant–dominant classification, with prototypical chords such as I and I$^6$ belonging to the tonic category, IV, IV$^6$ and ii belonging to the subdominant category and V, V$^7$ and V$^6$ belonging to the dominant category. We emphasize that this categorization scheme is fully self-emergent, requiring only very minimal assumptions about the structure of the underlying music. To be sure, a few categorizations deviate from those of theory textbooks: for example, the non-tertian sonority I$^4$ (C|CFG) was categorized as dominant, probably because it shares with dominant the tendency to be followed by I. But the correspondence between standard functional classifications and the results of machine learning is quite striking. This can also be seen in Figure 8, which shows that the standard TSD is very close to the optimal three-category deterministic cluster. Note that we again see that fundamental bass categorization is more complex and more accurate than a categorization based on the root of the chord.

These simple and preliminary results are encouraging, since they show that our methods can be used on purely automatic datasets, and that some of the handmade results are robust to the analysis procedure. The next section provides a much more detailed analysis of 16 datasets, mostly but not all handmade. The different corpora span a range of different extraction procedures and musical materials. We have also provided our code, as well as an applet allowing others to test their own datasets with our methods (cluster.norijacoby.com).

### 2.4 Analysing different surface representations: computing optimal curves and optimal deterministic categorizations for 16 datasets

We now apply our approach to 16 datasets, 11 of which are drawn from published works (de Clercq & Temperley, 2011; Huron, 2006; Rohrmeier & Cross, 2008; Temperley, 2009; Tymoczko, 2011) and five of which are new datasets constructed by author Tymoczko.

Again, we focus on the accuracy and complexity measures in Table 2, case C and Table 3, case A: $I_c(F; X)$, $I_c(Y'; F)$ with $Y' = C_{n+1}$ and $X = C_n$. Note that some of the published works provide only the conditional distribution $p(Y|X)$, whereas our algorithm requires the joint distribution $p(C_{n+1}, C_n) = p(C_{n+1}|C_n)p(C_n)$. However $p(C_n)$ can be estimated from $p(C_{n+1}|C_n)$ as the first left eigenvector of the matrix $p(C_{n+1}|C_n)$ (Feller, 1950). This estimation was only done when the marginal distributions were not available (Huron, 2006; Tymoczko, 2011). In all other cases, the distributions were available, and we used $p(C_n)$ directly. In cases where the marginal distribution was available we used the eigenvalues estimation of the marginals and the marginals themselves to verify that the two methods produced similar results, thereby validating the usage of the eigenvalue method in the cases where this approximation is necessary.

Tables B1 and B2, Appendix B, present datasets from published works: the corpora range from manual analyses with seven diatonic scale degrees (datasets 1–5, 8), manual analyses with twelve chromatic scale degrees (datasets 6, 7), chord bigrams extracted from manual analyses with 12 possible scale degrees (dataset 9, based on de Clercq and Temperley (2011)), and machine-constructed datasets of simultaneous notes expressed using standard pitch names (datasets 10 and 11, from Rohrmeier and Cross (2008)), where the chords were extracted from a MIDI file and transposed to C major or C minor, with the original key identified using a key-finding algorithm. These two latter datasets are very different from the others as the raw data include all sonorities, and not just harmonic triads and seventh chords. However, Rorhmeier and Cross (2008) simplified their dataset by keeping only the most common sonorities, which eliminated all but one non-tertian chord (the 'suspension' chord {C, D, G}). Thus although their original dataset is quite different from the others, the reduced data are fundamentally similar, since there is a direct translation between a Roman numeral such as '*ii* in C major' and an octave-free set of letter names such as {D, F, A}. Their analysis is similar to our analysis in Section 2.3, with the main differences being (a) that Rohrmeier and Cross (2008) used sophisticated and rhythm-dependent methods for eliminating passing chords; (b) that they used a wider window for key-finding (assigned one key to all the chorale); and (c) that they recorded transposed pitch classes but not the bass.

Finally, in order to facilitate comparison between the rock dataset 7 (Table B2) and the others, we made the following

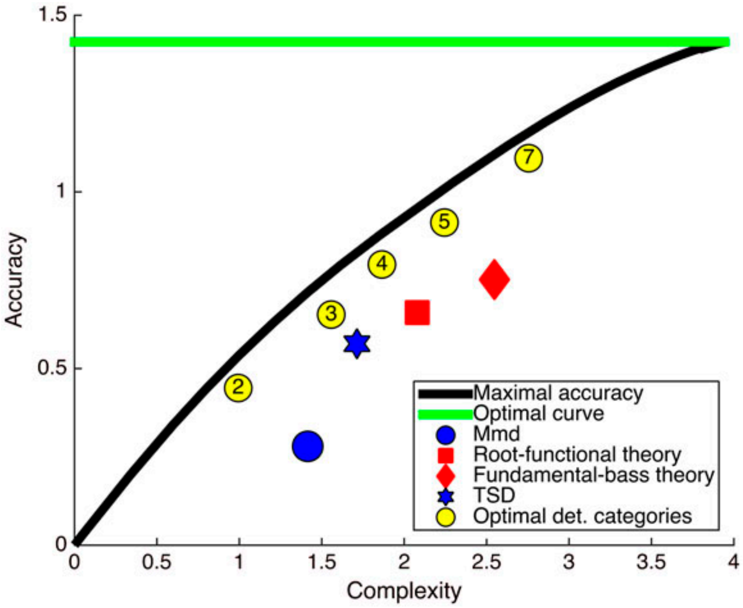

Fig. 8. Evaluation plane for a corpus of Bach chorales analysed from MIDI renditions. The results show high agreement between automatic and human labelled data. In particular, the standard tonic/subdominant/dominant ($F_{TSD}$) classification is quite close to the optimal three-category deterministic theory. Furthermore, strict bass categorization is more complex and more accurate than a categorization that uses the root of the chord.

reductions: we deleted the very rare chords: *bII* and #*IV*, then merged the chords *III* and *bIII*, *VI* and *bVI*, and *VII* and *bVII*, thus generating seven surface tokens with diatonic scale degrees *I*–*VII*. This reduction constitutes dataset 8 in Table B2. Note that we analysed the same dataset twice: once with this reduction (dataset 8) and once without it (dataset 7). As we will shortly see (see Figures 10(c)–(d) later on), our main conclusions were similar in both cases.

We also present five new datasets constructed by Tymoczko (Tables B3 and B4). Two of these derive from manual analyses of the 371 Bach chorales in the Riemenschneider edition; dataset 12 consists of transition frequencies between the 49 major-mode diatonic triads and sevenths (with bass notes) dataset 13 consists of the analogous transition frequencies for the 91 triads and seventh chords most common in the minor mode. (Keys here are determined locally rather than by the global tonic of the piece.)[3] The Mozart datasets contain the analogous information derived from analyses of all the Mozart piano sonatas, compiled with the assistance of more than 30 music theorists (dataset 14 contains major-mode passages, dataset 15 minor-mode). Finally, the Palestrina dataset contains handmade analyses of all Ionian, Mixolydian and 'Lydian' passages in seven Palestrina masses, two in Ionian and one in each of the remaining modes. This dataset explores the limitations of standard Roman numerals in the analysis of late Renaissance repertoire.

Tables B1–B4 provide a comprehensive comparison between optimal deterministic categorization obtained for the sixteen datasets. Figures 6–11 show the evaluation plane and optimal curves associated with some of these datasets. (Note again that the optimal curve itself is computed without the assumption of deterministic categorization; thus it is notable when deterministic categories are found very near the curve. ) The categorizations introduced in Equations 2, 4 and 7 ($F_{TSD}$, $F_{Mmd}$, $F_{soft\text{-}TSD}$) are indicated by the stars on the plane. The first few optimal deterministic categories are indicated on the plane.

Taken together, these datasets provide clear evidence for the syntactical reality of the tonic/subdominant/dominant classification. In datasets 1–8 in Tables B1 and B2, the three-cluster deterministic optimal categories always place I, IV and V in different clusters. This is consistent with the familiar idea that I, IV and V are 'prototypes' of tonic, subdominant, and dominant categories, with the other chords (II, III, VI and VII) more loosely associated with one or more categories (Agmon, 1995). Furthermore, Figures 6, 9 and 10 show that $F_{TSD}$ and $F_{soft\text{-}TSD}$ were nearly optimal on datasets 1–8,

[3] The first 70 chorales were compiled with the help of undergraduates in Tymoczko's MUS306 course at Princeton University, as well as several graduate students (including Hamish Robb and Luis Valencia). For the remaining 301 chorales, Tymoczko corrected analyses produced by Heinrich Taube's 'Chorale Composer' software, as improved by Simon Krauss, an undergraduate thesis student of Tymoczko's. All data were then thoroughly cross-validated using Michael Cuthbert's music21 toolkit, with all discrepancies further analysed to locate possible errors. The 49 major-mode chord forms include three triadic and four seventh-chord inversions for each of the 7 scale degrees. The 91 minor mode chords include three inversions of the 13 triadic forms residing in the natural, harmonic, and melodic minor scales (two triadic forms for every chord except the tonic), and all the corresponding seventh chords except for i$^{\text{maj}7}$, the minor triad with a major seventh, which does not often appear in classical music.

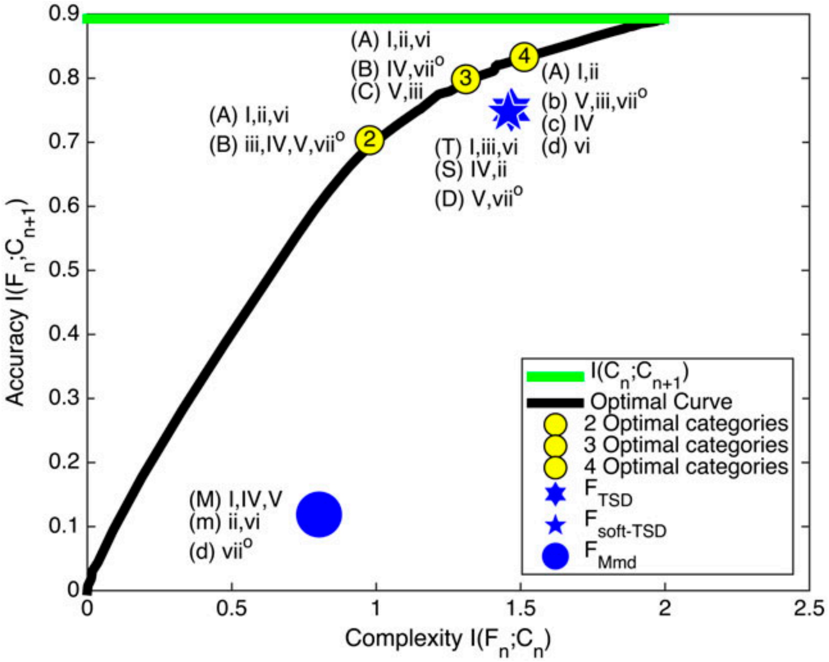

Fig. 9. Information curve (black) for major Mozart passages based on the Tymoczko datasets (2011a, §7. 1) plotted on the evaluation plane. The maximal accuracy ($I(C_n; C_{n+1})$) is indicated by the upper line. The figure also shows the comparison between optimal clustering to 2, 3 and 4 clusters and $F_{TSD}$, $F_{soft\text{-}TSD}$, $F_{Mmd}$ of Equations 2, 4 and 7. The figure demonstrates that two *different* categorization schemes (in this example $F_{TSD}$ and the optimal categorization to three clusters) can perform similarly.

lying very close to the optimal curve. By contrast, we can see that the clustering $F_{Mmd}$ (even though it contains the same number of symbols as $F_{\mathrm{TSD}}$) performed poorly on all the relevant datasets. In datasets 1 and 2, the textbook clustering of Equation 2 was the optimal three-cluster categorization. In keeping with the functional tradition, datasets 1, 2, 6, 8, and 10 often group chords V and vii° together. Similarly, the first cluster in the two-cluster division of dataset 10 is {{D, G, B}, {D, F, G, B}, {D, F, B }}, or $V$, $V^7$ and $vii^o$ – the same clustering produced in many of the handmade corpora. To be sure, there are often deviations from traditional categorization schemes: for example, chords $I$ and ii are sometimes grouped together (datasets 3, 4 and 6; Figure 9); this is probably due to the fact that both $I$ and ii tend to progress to $V$. However the generally close alignment between the functions of traditional theory and our self-emergent clusters suggest that listeners could infer traditional tonal functions directly from statistical properties of the musical surface.

More specifically, the close alignment suggests that *local predictions* play an important role in traditional functional categorization. We can largely recover the terminology of traditional harmonic theory by categorizing chords so as to maximize our ability to *anticipate the next chord*, rather than, for example, focusing on shared pitch-class content, levels of dissonance, or the intrinsic 'sound' of each sonority. This suggests that to be a dominant chord is to a significant extent a matter of *acting like* a dominant chord; that is, to move authentically to $I$ and $I^6$ and, less often, deceptively to $IV^6$ and $vi$. (Of course, there is more to traditional functional categories than simply predicting the next chord, as we discuss in the next section.) This is significant insofar as traditional accounts do not emphasize the predictive utility of tonal functions.

The two non-classical datasets deserve extra attention. As expected from de Clercq and Temperley (2011), chord $IV$ is highly important in the rock corpus: it categorized alone in the optimal 3- and 4-cluster solutions (see dataset 8 of Table B2). The categories in dataset 9, which distinguish progressions whose roots belong to the major scale (cluster 1) from those whose roots belong to natural minor (cluster 2) are also interesting in that they it suggests a kind of oscillation or modulation between two different harmonic regions. Note also that the common-practice division into tonic, subdominant and dominant ($F_{TSD}$ and $F_{soft\text{-}TSD}$) worked quite well (see Figure 10(d)), indicating that the classical theory of functional categories continues to apply to popular music, at least at a coarse-grained level. A similar phenomenon is evident in the Palestrina dataset, which consists of music written before the widespread adoption of functional harmony. Here again, we see that the self-emergent optimal three-category clustering is similar to tonic/subdominant/dominant, suggesting a kind of 'proto-functionality' already at work in this purportedly 'modal' music. Our approach thus supports the intuition that the boundaries between 'functional tonality' and other styles of music is a fuzzy one, with aspects of functionality present in music outside the 'common practice period' of 1680–1850 (Tymoczko, 2011). This in turn supports the analytical project of attempting to understand this proto-functionality in greater detail.

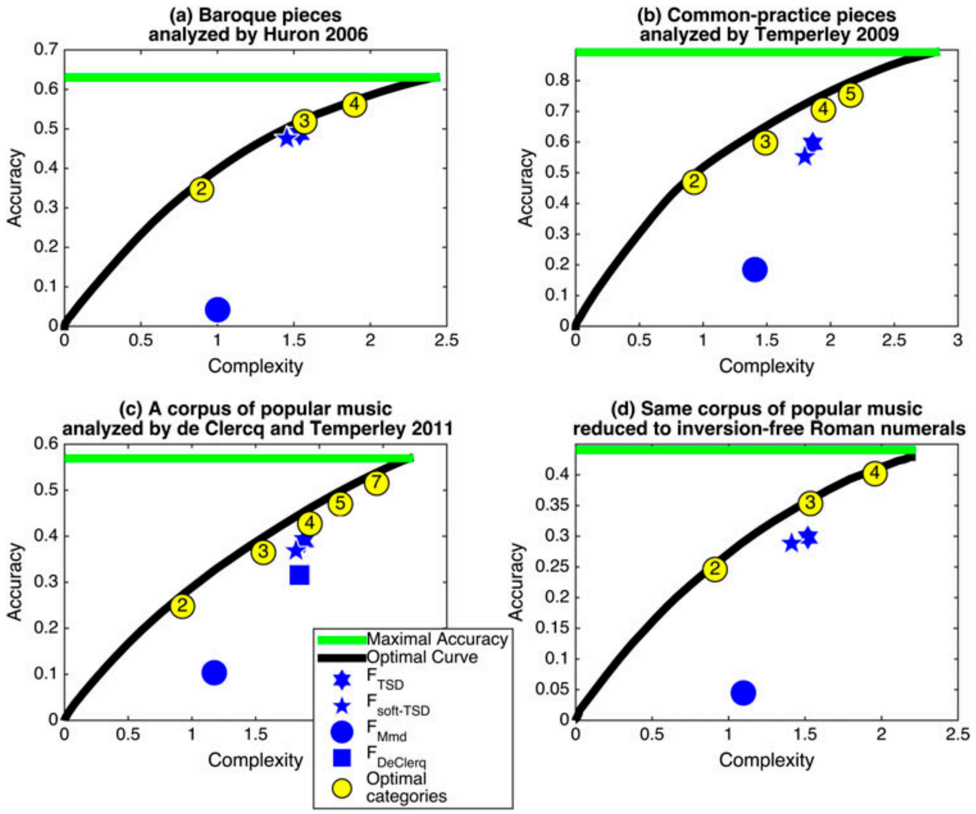

Fig. 10. Optimal curve plotted on the evaluation plane for datasets of (a) Baroque music from Huron (2006); (b) 46 excerpts of common practice examples from a textbook (Kostka & Payne 1995) analysed by Temperley (2009); a corpus of rock music analysed by de Clercq and Temperley (2011) with two versions, one with chromatic inversion-free Roman numerals (c), and one reduced to 7 diatonic scale degrees (d). This reduction was done in order to compare this dataset directly with the other common practice corpora. In each diagram we marked the accuracy and complexity for the first few optimal deterministic categories and $F_{TSD}$ (I, vi, iii/IV, ii/V, vii$^{o}$/all other), $F_{soft\text{-}TSD}$ (I, iii(50%), vi(50%)/IV, ii, vi(50%)/V, vii$^{o}$, iii(50%)/all other), $F_{Mmd}$ (I, IV, V/ii, vi, iii/vii$^{o}$/all other) of Equations 2, 4 and 7. Note that in spite of the large stylistic range, $F_{TSD}$ and $F_{soft\text{-}TSD}$ are always near-optimal, while $F_{Mmd}$ is suboptimal. In the rock corpus (d), we also compare the categorization based on the connected components of the graph in Figure 5(c) of de Clercq and Temperley (I, bii, #IV, vii$^{o}$/ii, iii, vi/biii, bvi, bvii$^{o}$/V, IV; see 2011, p. 66, Figure 4). We see that categorization based on our method is comparable and slightly better than the one based on the similarity graph presented in their paper. The full list of categories marked on this graph can be found in Appendix B (Tables B1 and B2).

It is worth re-emphasizing that our analysis involves very few assumptions: we simply ask how best to compress surface tokens – whatever they are – to predict the near future. Thus, unlike Rohrmeier and Cross (2008), we do not impose the requirement of hierarchical clustering; in general, the process of increasing the number of clusters is not simply a process of splitting one cluster into two components. Nor do we need to resort to intuitive similarity metrics based on statistics of chords, as in Tymoczko (2003) or Rohrmeier and Cross (2008). These metrics, like our approach, use transition probabilities to develop a notion of chord classification or similarity, and to this extent deliver results closely related to our own. (Indeed, Figure 10 shows high agreement between our clusters and the hierarchical clustering in Rohrmeier and Cross (2008).) The difference is that our approach is conceptually minimalist, and is grounded in established techniques from information theory.

### 2.5 Comparison with different accuracy metrics

The examples in the previous section depart from traditional theory by focusing on the near-future. Traditional harmonic functions might be thought to include the past as well: two chords are thought to have the same harmonic function if they both *proceed to* and *are preceded by* the same harmonies. (Thus chords *IV6/4* and *V6/5* are said to have different harmonic functions, even though both chords overwhelmingly tend to proceed to *I*.) However we can easily use our methodology to categorize chords based on both their ability to ‘retrodict’ the past, or even a combination of prediction and retrodiction; it is simply a matter of choosing a different accuracy metric (Table 3). Tables 5–6 provide optimal deterministic categories for two of our datasets. These tables show that all variants yield very similar results. The only substantial difference is that retrodiction causes the tonic chord (*I*) to be categorized separately, largely because tonic chords are very likely to be preceded by dominant chords. (In the

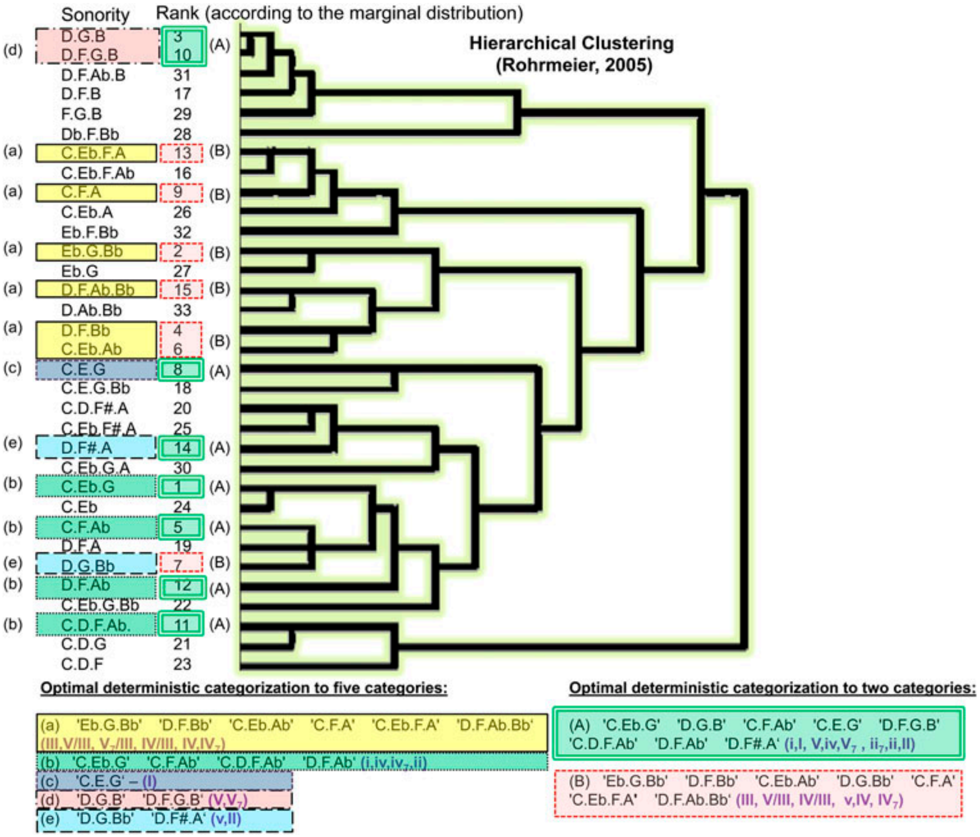
Sonority
Rank (according to the marginal distribution)
Hierarchical Clustering (Rohrmeier, 2005)
(d) D.G.B 3 (A)
D.F.G.B 10
D.F.Ab.B 31
D.F.B 17
F.G.B 29
Db.F.Bb 28
(a) C.Eb.F.A 13 (B)
C.Eb.F.Ab 16
(a) C.F.A 9 (B)
C.Eb.A 26
Eb.F.Bb 32
(a) Eb.G.Bb 2 (B)
Eb.G 27
(a) D.F.Ab.Bb 15 (B)
D.Ab.Bb 33
(a) D.F.Bb 4
C.Eb.Ab 6 (B)
(c) C.E.G 8 (A)
C.E.G.Bb 18
C.D.F#.A 20
C.Eb.F#.A 25
(e) D.F#.A 14 (A)
C.Eb.G.A 30
(b) C.Eb.G 1 (A)
C.Eb 24
(b) C.F.Ab 5 (A)
D.F.A 19
(e) D.G.Bb 7 (B)
(b) D.F.Ab 12 (A)
C.Eb.G.Bb 22
(b) C.D.F.Ab. 11 (A)
C.D.G 21
C.D.F 23
Optimal deterministic categorization to five categories:
(a) 'Eb.G.Bb' 'D.F.Bb' 'C.Eb.Ab' 'C.F.A' 'C.Eb.F.A' 'D.F.Ab.Bb' (III,V/III, V7/III, IV/III, IV,IV7)
(b) 'C.Eb.G' 'C.F.Ab' 'C.D.F.Ab' 'D.F.Ab' (i,iv,iv7,ii)
(c) 'C.E.G' – (I)
(d) 'D.G.B' 'D.F.G.B' (V,V7)
(e) 'D.G.Bb' 'D.F#.A' (v,II)
Optimal deterministic categorization to two categories:
(A) 'C.Eb.G' 'D.G.B' 'C.F.Ab' 'C.E.G' 'D.F.G.B' 'C.D.F.Ab' 'D.F.Ab' 'D.F#.A' (i,I, V,iv,V7 , ii7,ii,II)
(B) 'Eb.G.Bb' 'D.F.Bb' 'C.Eb.Ab' 'D.G.Bb' 'C.F.A' 'C.Eb.F.A' 'D.F.Ab.Bb' (III, V/III, IV/III, v,IV, IV7)

Fig. 11. Comparison of our optimal deterministic categories and the hierarchical clustering of minor Bach chorales in Rohrmeier and Cross (2008). The figure shows high agreement between the two approaches.

Table 5. Deterministic optimal categories of structural variants in dataset 1A.

| | Variant (Table 3) | Compressed Variable | Predicted Variable | Categories |
|---|---|---|---|---|
| A | First-order predictive (standard) | $X = C_t$ | $Y = C_{t+1}$ | **2 Categories:** $V, vii^o/I, IV, ii, vi, iii$<br>**3 Categories:**$V, vii^o/I, vi, iii/IV, ii$<br>**4 Categories:**$V, vii^o/I, iii/IV, vi/ii$ |
| B | First-order preceding chord (‘time reversed’) | $X = C_t$ | $Y = C_{t-1}$ | **2 Categories:**$\boldsymbol{I/V, IV, vii^o, ii, vi, iii}$<br>**3 Categories:**$\boldsymbol{I/V, vii^o/IV, ii, vi, iii}$<br>**4 Categories:**$\boldsymbol{I/V, vii^o/IV, vi, iii/ii}$ |
| C | Mixed past-future first order | $X = C_t$ | $Y = (C_{t-1}, C_{t+1})$ | **2 Categories:**$V, vii^o/I, IV, ii, vi, iii$<br>**3 Categories:**$\boldsymbol{V, vii^o/I, iii/IV, ii, vi}$<br>**4 Categories:**$V, vii^o/I, iii/IV, vi/ii$ |
| D | Second-order predictive | $X = C_t$ | $Y = (C_{t+1}, C_{t+2})$ | **2 Categories:**$V, vii^o/I, IV, ii, vi, iii$<br>**3 Categories:**$V, vii^o/I, vi, iii/IV, ii$<br>**4 Categories:**$V, vii^o/I, iii/IV, vi/ii$ |

forward-oriented accuracy metric of Table 3, case A, *I* and *iii* can categorized together because they tend to move in the same way; when we focus on retrodiction, it becomes relevant that *I* is more likely to be directly preceded by a dominant.) This suggests that the clusters of functional harmony can be simultaneously understood as indicating *how chords tend to move* and *how chords tend to be approached*. Table 3 shows the deterministic optimal categorization into three categories obtained using dataset 1A. (Dataset 1A is drawn from all 371 chorales rather than the 70 in dataset 1, and also uses the accurate marginal distribution.) The second-order variant produces results identical to those in the standard method, which is why we focus on first-order statistics in this article. This supports the hypothesis that harmonic categorization is primarily dependent on very local structure.

Table 7 calculates the optimal three-category deterministic categorization with the functional predictive clustering accuracy metric (in Table 3, case F) and with the standard

Table 6. Deterministic optimal categories of structural variants in dataset 12.

| | Variant (Table 3) | Optimal 3 Categories for dataset 12 |
|---|---|---|
| A | First-order predictive | **Category 1:** $I, vi, iii, V^2, vi^6, V^7/IV, V6/5/IV, V^6/vi, iii^6, V2/IV, V6/5/vi, vii^{\varnothing}4/3, vii^{o}6/4, vii^{\varnothing}6/5, V^2/V, Imaj^7, V4/3/IV, vii^{o7}/vi, V/vi, v$<br>**Category 2:** $IV, IV^6, ii6/5, ii, ii^7, I6/4, ii^6, V6/5/V, IVmaj^7, vii^{o6}/V, IVmaj6/5, vi^7, ii^2, V/V, V^6/V, vii^{\varnothing7}/V, IVmaj^2, V^7/V, vi6/4, vii^{\varnothing}4/3/V, V6/5/ii, I^6$<br>**Category 3:** $V, V^7, V^6, vii^{o6}, vii^{o}, V6/5, vii/o^7, V4/3, V6/4, IV6/4$ |
| B | Time Reversed | **Category 1:** $I, V^7, V^2, iii^6, V6/vi, Imaj^7$<br>**Category 2:** $I^6, vi, IV^6, iii, vi^6, V4/3, V^2/IV, vi^7, vii^{o6}/V, IVmaj6/5, V6/5/vi, V6/4, vi6/4, vii^{\varnothing}4/3/V, vii^{o7}/vi, V/vi, V4/3/IV, IV6/4$<br>**Category 3:** $V, IV, V^6, vii^{o6}, ii6/5, V6/5, ii, ii^7, I6/4, ii^6, vii^{\varnothing7}, V6/5/V, V^7/IV, IVmaj7, V6/5/IV, vii^{o}, ii^2, vii^{\varnothing}4/3, V/V, V^6/V, vii^{\varnothing7}/V, IVmaj^2, V^7/V, viio6/4, V^2/V, vii^{\varnothing}6/5, V6/5/ii, v$ |
| C | Mixed past-future | **Category 1:** $I, I^6, vi, V^7/IV, vi^6, iii^6, vii^{o6}/V, V^2/IV, V6/5/IV, V^6/vi, V^7/V, vi6/4, V4/3/IV, Imaj^7, V6/5/ii, IV6/4, v$<br>**Category 2:** $IV, V^6, IV^6, ii6/5, vii^{o6}, V^2, ii, ii^7, I6/4, ii^6, iii, V6/5/V, IVmaj^7, vi^7, IVmaj6/5, viio, ii^2, vii^{\varnothing}4/3, V6/5/vi, V/V, V^6/V, vii^{\varnothing7}/V, IVmaj^2, vii^{o}6/4, vii^{\varnothing}4/3/V, vii^{\varnothing}6/5, V/vi$<br>**Category 3:** $V, V^7, V6/5, vii^{\varnothing7}, V4/3, V6/4, V^2/V, vii^{o7}/vi$ |
| D | Second-order predictive | **Category 1:** $I, I^6, vi, IV^6, iii, IVmaj6/5, vi^6, vi^7, V6/5/IV, V^7/IV, iii^6, vii^{o6}/V, ii^2, V^2/IV, V6/5/vi, V6/V, vii^{\varnothing7}/V, vii^{\varnothing}6/5, V^7/V, V4/3/IV, V^2/V, Imaj^7, v$<br>**Category 2:** $IV, ii6/5, ii, V^2, ii^7, I6/4, ii^6, V6/5/V, IVmaj^7, vii^{\varnothing}4/3, V/V, vii^{o}6/4, vii^{\varnothing}4/3/V, IVmaj^2, vi6/4, V6/5/ii$<br>**Category 3:** $V, V^6, V^7, vii^{o6}, V6/5, vii^{\varnothing7}, V4/3, V6/vi, vii^{o}, V6/4, V/vi, vii^{o7}/vi, IV6/4$ |

first-order predictive metrics (Table 3, case A), using two versions of 371 Bach chorales (datasets 1A and 12). The table shows that the categories obtained by the two methods are once again similar: in the case of dataset 1A they are identical. In the case of dataset 12 there are some important differences, however, with *V* and *IV* clustered together in the functional predictive approach, contrary to musical intuition. Furthermore, the algorithm for computing the functional predictive clustering is usually much slower and more sensitive to the problem of local minima (see Appendix A).

The similarity of the approaches can be underscored by returning to the example in Section 2.2 (Table 1, cases C and D), which compares root and bass using the accuracy metric in Table 3, case A. Somewhat surprisingly, when we move to the accuracy metric of Table 3, case F, the results do not differ materially: once again the root-oriented theory is significantly simpler (2.3 bits versus 2.7 bits or 11% of the total entropy) whereas the bass-oriented theory is slightly more accurate (0.37 bits versus 0.33 bits or 2.8% of the maximal mutual information, an even smaller difference than the 3.9% in first-order predictive variance of Section 2.2). This shows that the greater accuracy of the bass-oriented theory is not simply a result of the fact that the first-order predictive variant uses functions to predict chords themselves. (One might have thought, as we in fact did, that the superiority of the bass-oriented theory was an artifact of using functions to predict the *specific inversion* of the next chord.) Even when we closely model the procedures of traditional music theory, in which functions are used to predict functions, the bass-oriented theory proves to be slightly more accurate.

Further research is needed to compare the advantages and disadvantages of these variants. One natural direction for a generalization would be to use variable length Markov chains. For example, Conklin (2010) and Pearce and Wiggins (2006) developed an approach in which predictions are based on ‘multiple viewpoints’ or variable length Markov chains. If we apply this approach here, we can try to estimate the variable $Y$ with a ‘mixed-order’ variable $Y'$ (instead of a fixed order, as is the case for all of the variants in Table 3, cases A–E). This promising direction calls for further investigation.

## 2.6 Comparison with HMM

Recent work has applied Hidden Markov Models (HMM) to corpus analysis (Mavromatis, 2009, 2012; Raphael & Stoddard, 2004; Temperley, 2007). Although both models come from the domain of machine learning, HMM is a *generative* process where surface tokens are emitted from hidden states; by contrast, our method is an *analytical* process that generates functional states from surface tokens. Figure 12 shows the similarities and differences between these two formalisms. The HMM model can be related to the compositional process where the composer has some desired functional progression in mind, which is then expressed by the appropriate surface tokens. Our formalism, on the other hand, can be likened to the experience of a listener who reconstructs a functional progression as musical information unfolds in real time. In our formalism, once a categorization scheme has been acquired, the listener simply computes her estimate of the current function from the current surface token. In a HMM model, deciphering functional labels from surface tokens is non-direct and requires applying a complex algorithm requiring high memory capacity (Viterbi decoding; see Rabiner & Juang, 1986). Clearly there will be some situations where HMM approaches are preferable, including those where we wish to simulate a composer's behaviour; in this sense the methods are complementary.

The main advantage of our method is that ours has significantly fewer degrees of freedom. In HMM one needs to specify $p(C_n|F_n)$, which is comparable to $p(F_n|C_n)$ in our approach. However in HMM one also needs to specify $p(F_{n+1}|F_n)$ which is a matrix of size $|F| \times |F|$. These extra degrees of freedom do not necessarily correspond to knowledge possessed by musical listeners. (For example, a listener might know that tonics follow dominants, but not have a very specific quantitative hypothesis about the frequency of this progression.) In our approach we simply specify clusters ('*I* and *vi* are tonics') and let the algorithm derive probabilities such as $p(F_{n+1}|F_n)$. Thus it is easy to compare pre-existing categorization schemes (Table 1), whereas this is not completely natural using the HMM approach.

Table 8 compares the two methods as applied to the same corpus, modelling functional categories in two different datasets: (a) dataset 1A, inversion-free two-chord Roman-numeral progressions drawn from major-mode passages in all 371 chorales; and (b), dataset 12, containing two-chord major-mode progressions drawn from the 371 chorales, but including figured-bass symbols as well as inversions. In both cases, we cluster each chord to the most likely function associated with it. Formally, we choose for each chord $c$ the function $f$ that maximizes the likelihood:

$$P(F_n = f|C_n = c) = p(C_n = c|F_n = f) \times \quad p(F_n = f)/p(C_n = c). \tag{19}$$

For HMM we used Matlab's *hmm_train* function with three hidden states. Although both techniques reproduce the TSD classification in the 7-category dataset (1A), the HMM has more trouble with more categories: here, *IV* and *V* are categorized in the first cluster and *I* and $V^7$ in the second. This suggests that our framework is more consonant with traditional functional ideas. In retrospect, this is not surprising, since we are using methods specifically designed for clustering (see Friedman et al., 2001; Hecht et al., 2009; Slonim & Tishby, 2006).

Further research is needed to determine whether the advantage of our approach derives simply from the reduction in degrees of freedom or from deeper structural differences. It is suggestive that our approach originates from a model of how the brain uses perceptual categories to screen out irrelevant sensory information (Tishby & Polani, 2011). Any ear-training teacher will recognize this familiar musical situation: multiplicity often overwhelms beginning students, who struggle to distinguish closely related chords such as *V4/3* and $vii^{o6}$, $IV^6$ and *vi*, or $I^6$ and *iii*. Many pedagogues recommend that students begin with simplified categories such as tonic, predominant and dominant, or these same categories augmented with bass notes (Quinn, 2005). Once these perceptual categories are firmly in place students can then turn to the finer differentiations within categories. The mathematics of our approach closely mimic this process of perceptual simplification (see Tishby & Polani, 2011), with the plausibility of its results suggesting that prediction constitutes one important feature of traditional harmonic functions.

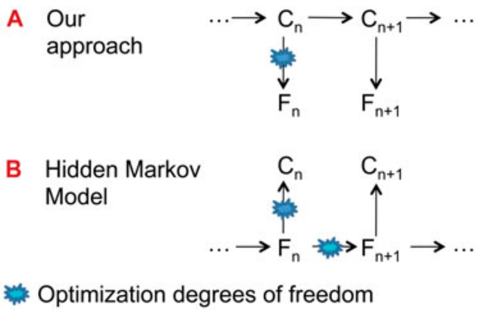
A
Our approach
··· → C<sub>n</sub> → C<sub>n+1</sub> → ···
F<sub>n</sub>
F<sub>n+1</sub>
B
Hidden Markov Model
C<sub>n</sub>
C<sub>n+1</sub>
··· → F<sub>n</sub> → F<sub>n+1</sub> → ···
Optimization degrees of freedom

Fig. 12. Structural similarities between our approach and HMM. The arrows in the diagram represent a graphical model (Feller, 1950) of the statistical relation between the random variables $C_n$ and $F_n$. In the HMM all other distributions are determined by $p(F_{n+1}|F_n)$ and $p(C_n|F_n)$ but in our approach all other distributions are determined from $p(F_n|C_n)$. In both cases we assume the empirical distribution of $C_n$ is computable from a large corpus.

## 2.7 Discussion

The strength of our framework is that it is a unified, fairly assumption-free approach where harmonic categories emerge naturally from data. The twin notions of the evaluation plane and optimal curve help to focus attention on the inherent tradeoffs between complexity and accuracy. This gives a new way to consider the gains that can be obtained by altering the resolution of harmonic theories (i.e. adding or subtracting additional categories or symbols). As we have seen, our framework reproduces traditional classifications into tonic, subdominant and dominant categories, while suggesting several avenues for more detailed music-theoretical research – such as the similarity of *IV* and *vi* in Figure 6, or the presence of attenuated harmonic functionality in Palestrina and rock. In this sense, corpus data and machine learning can provide the impetus for more traditional and detailed music-theoretical explorations.

The strength of this method is also a weakness: our method uses only probability distributions while ignoring the psychological perceptual similarities of chords. This can produce categories that make sense based on local statistics, but are less intuitive in musical terms, grouping chords according to behaviour rather than sound. (Recall, in this context, the difference between *IV6/4* and *V6/5*, which both tend to proceed to *I*.) That said, since perceptual similarity can be expected to influence composer choice, and hence the historical

Table 7. The deterministic optimal categorization into three categories of the first order predictive (Table 3A) and first order pairwise (Table 3F) variants.

| | Variant (Table 3) | Deterministic optimal categorization to three categories: dataset 1A | Deterministic optimal categories: dataset 12 |
|---|---|---|---|
| A | First-order predictive | **Category 1:** $I, vi, iii$ | **Category 1:** $I, vi, iii, V^{2}, vi^{6}, V7/IV, V6/5/IV, V6/vi, iii^{6}, V^{2}/IV, V6/5/vi, vii^{\varnothing}4/3, vii^{o}6/4, vii^{\varnothing}6/5, V^{2}/V, Imaj^{7}, V4/3/IV, vii^{o7}/vi, V/vi, v$ |
| | | **Category 2:** $IV, ii$ | **Category 2:** $IV, IV^{6}, ii6/5, ii, ii^{7}, I6/4, ii^{6}, V6/5/V, IVmaj^{7}, viio6/V, IVmaj6/5, vi^{7}, ii^{2}, V/V, V^{6}/V, vii^{\varnothing}7/V, IVmaj^{2}, V^{7}/V, vi6/4, vii^{\varnothing}4/3/V, V6/5/ii, I^{6}$ |
| | | **Category 3:** $V, vii^{o}$ | **Category 3:** $V, V^{7}, V^{6}, vii^{o6}, vii^{o}, V6/5, vii^{\varnothing7}, V4/3, V6/4, IV6/4$ |
| F | First-order functional predictive | **Category 1:** $I, vi, iii$ | **Category 1:** $I, I^{6}, vi, iii, vi^{6}, iii^{6}, V6/5/vi, vii^{\varnothing7}/V, vi6/4, Imaj^{7}, IV6/4, V/vi$ |
| | | **Category 2:** $IV, ii$ | **Category 2:** $V7, vii^{o6}, V6/5, V^{2}, vii\varnothing7, V4/3, V^{6}/vi, vii^{o}, vii^{\varnothing}4/3, vii^{o}6/4, vii^{\varnothing}6/5, V6/4, vii^{o7}/vi$ |
| | | **Category 3:** $V, vii^{o}$ | **Category 3:** $V, IV, V^{6}, IV^{6}, ii6/5, ii, ii^{7}, I6/4, ii^{6}, V6/5/V, IVmaj^{7}, vii^{o6}/V, V^{7}/IV, IVmaj6/5, V6/5/IV, vi^{7}, V^{2}/IV, ii^{2}, V/V, V6/V, IVmaj^{2}, V^{7}/V, vii^{\varnothing}4/3/V, V^{2}/V, V4/3/IV, V6/5/ii, v$ |

Table 8. A comparison of our first-order predictive variant approach with the Hidden Markov Model (HMM) on two datasets 1A and dataset 12. For both methods, for each chord $c$ we chose the function $f$ that maximizes the likelihood $p(F_n = f|C_n = c) = p(C_n = c|F_n = f)p(F_n = f)/p(Cn = c)$.

| | Method | 3 Clusters: Dataset 1A | 3 Clusters: Dataset 12 |
|---|---|---|---|
| A | First-order predictive. | **Category 1:** $I, vi, iii$ | **Category 1:** $I, vi, V^{2}, iii, vi^{6}, V^{7}/IV, V6/5/IV, V6/vi, iii6, V^{2}/IV, V6/5/vi, vii^{\varnothing}4/3, vii^{o}6/4, vii^{\varnothing}6/5, V^{2}/V, Imaj^{7}, V4/3/IV, vii^{o7}/vi, V/vi, v$ |
| | | **Category 2:** $IV, ii$ | **Category 2:** $IV, IV^{6}, ii6/5, ii, ii^{7}, I6/4, ii^{6}, V6/5/V, IVmaj^{7}, vii^{o6}/V, IVmaj^{6}/5, vi^{7}, ii^{2}, V/V, V^{6}/V, vii^{\varnothing7}/V, V^{7}/V, IVmaj^{2}, vi6/4, vii^{\varnothing}4/3/V, V6/5/ii, I^{6}$ |
| | | **Category 3:** $V, vii^{o}$ | **Category 3:** $V, V^{7}, V^{6}, vii^{o6}, V6/5, vii^{\varnothing7}, V4/3, vii^{o}, V6/4, IV6/4$ |
| B | HMM | **Category 1:** $I, iii$ | **Category 1:** $V, IV, V^{6}, V6/5, ii, ii^{7}, ii^{6}, vii^{\varnothing}7, vii^{o}, vii^{o}6/4, IVmaj^{2}, v$ |
| | | **Category 2:** $IV, ii, vi$ | **Category 2:** $I, V^{7}, viio^{6}, V^{2}, V^{6}/vi, iii^{6}, vii^{\varnothing}4/3, vii^{\varnothing}6/5, V6/4, vii^{o7}/vi$ |
| | | **Category 3:** $V, vii^{o}$ | **Category 3:** $I^{6}, vi, IV^{6}, ii6/5, I6/4, iii, V6/5/V, V4/3, IVmaj7, vi^{6}, vii^{o6}/V, V^{7}/IV, IVmaj6/5, V6/5/IV, vi^{7}, V^{2}/IV, ii^{2}, V6/5/vi, V/V, V^{6}/V, vii^{\varnothing7}/V, V^{7}/V, vi6/4, vii^{\varnothing}4/3/V, V^{2}/V, Imaj^{7}, V4/3/IV, V/vi, V6/5/ii, IV6/4$ |

development of musical syntax, it should be reflected in the distribution of chords. This could perhaps explain why we were able to recover a considerable amount of musical structure even when we ignored everything but local chord distributions. But again, the question of relative importance of perceptual and syntactical chord similarities requires further systematic investigation.

Once again, we emphasize that our methods can work with any type of surface representations – handmade inversion-free roman numerals, Roman numerals with inversions, or even sonorities extracted automatically from MIDI files. As we saw in the results section, we have obtained similar results using a variety of repertoires, representations, and annotation methods. Here, too, the different assumptions made during the preparation of corpora calls for future systematic investigation: we hope that our methodology could serve as a unified analytical layer which allows for systematic testing of assumptions made durning initial corpus generation.

We close by returning to the empirical grounding of functional language. For centuries, theorists and pedagogues have been producing theories of 'harmonic function' without attempting to ground these theories in either psychological experiments or in statistical regularities of musical corpora. This naturally raises a question as to the viability of the foundations of these theories. Our results suggest that tonal function is indeed learnable from statistical features of the musical stimulus, and moreover that it is importantly involved in *prediction*, or the formation of musical expectations. Furthermore, *local context* (the statistical relation between adjacent chords) is often sufficient to completely recover the standard functional categories. This is in line with psychological experiments that show enhanced perceptual sensitivity to local harmonic cues (for a review see Tillmann and Bigand (2004)).

At the very minimum, functional language has a place in providing a simplified description of the patterns found in actual music. A further step would be to determine the perceptual relevance of this idea: for instance, we might consider whether listeners are sensitive to functions emerging from our formalism in the contexts of unfamiliar, artificial musical languages; and if so, whether our techniques can help model this process (see for example Loui (2012), Loui and Wessel (2007) and Loui, Wessel and Kam (2010), demonstrating that listeners can acquire specific preferences for music generated by an artificial harmonic grammar).

## Acknowledgements

The authors would like to thank Martin Rohrmeier, Eytan Agmon, Tom Gurion, Tom Yuval, and Dror Menashe for their assistance with the project. Finally, the first author would specifically like to thank Carmel Raz for her extraordinary help in the research phase of the project.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## References

Agmon, E. (1995). Functional harmony revisited: a prototype-theoretic approach. *Music Theory Spectrum, 17*(2), 196–214.

Ariza, C., & Cuthbert, M. (2010). music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In: Proceedings of the International Symposium on Music Information Retrieval, pp. 637–42.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics, 18*(4), 467–479.

Chechik, G., Globerson, A., Tishby, N., & Weiss, Y. (2005). Information bottleneck for Gaussian variables. *Journal of Machine Learning Research, 6*(1), 165–188.

Cohn, R. (1998). Introduction to Neo-Riemannian theory: A survey and a historical perspective. *Journal of Music Theory, 42*(2), 167–180.

Conklin, D. (2010). Discovery of distinctive patterns in music. *Intelligent Data Analysis, 14*(5), 547–554.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. New York, NY: Wiley.

De Clercq, T., & Temperley, D. (2011). A corpus analysis of rock harmony. *Popular Music, 30*, 47–70.

Feller, W. (1950). *An introduction to probability theory and its applications*. New York, NY: Wiley.

Friedman, A., & Goldberger, J. (2013). Information theoretic pairwise clustering. In E. Hancock & M. Pelillo (Eds.), *SIMBAD 2013 (Lecture Notes in Computer Science 7953* (pp. 106–119). Berlin, Heidelberg: Springer-Verlag.

Friedman, N., Mosenzon, O., Slonim, N., & Tishby, N. (2001). Multivariate information bottleneck. *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI)*. Cambridge, MA: MIT Press.

Gasparini, F. (1715). *L'Armonico Pratico al Cimbalo*. Venice.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer, 27*(2), 83–85.

Hecht, R.M., Noor, E., & Tishby, N. (2009, Sept). *Speaker recognition by gaussian information bottleneck*. Paper presented at Tenth Annual Conference of the International Speech Communication Association (INTERSPEECH 2009); Brighton, UK,

Heinichen, J.D. (1728). *Der Generalbass in der Komposition*. Dresden.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Kirenberg, J.P. (1774). Die Kunst des reinen Satzes in der Musik. Berlin und Königsberg: G. J. Decker und G. L. Hartung.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680.

Kostka, S. M., & Payne, D. (1984). *Tonal harmony: With an introduction to twentieth-century music* (1st ed.). New York: A.A. Knopf.

Lerdahl, F., & Jackendoff, R. S. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Loui, P. (2012). Learning and liking of melody and harmony: Further studies in artificial grammar learning. *Topics in Cognitive Science, 4*(4), 554–567.

Loui, P., & Wessel, D. (2007). Harmonic expectation and affect in Western music: Effects of attention and training. *Perception & psychophysics, 69*(7), 1084–1092.

Loui, P., Wessel, D.L., & Kam, C.L.H. (2010). Humans rapidly learn grammatical structure in a new musical scale. *Music perception, 27*(5), 377–388.

Mavromatis, P. (2009). Minimum description length modelling of musical structure. *Journal of Mathematics and Music, 3*(3), 117–136. doi:10.1080/17459730903313122.

Mavromatis, P. (2012). Exploring the rhythm of the Palestrina style: A case study in probabilistic grammar induction. *Journal of Music Theory, 56*(2), 169–223.

Meeús, N. (2000). Toward a post-Schoenbergian grammar of tonal and pre-tonal harmonic progressions. *Music Theory Online, 6*(1).

Niedt, F.E. (1706). *Musikalische Handleitung*. Hamburg.

Pearce, M.T., & Wiggins, G.A. (2006). Expectation in melody: The influence of context and learning. *Music Perception, 23*(5), 377–405.

Pereira, F., Tishby, N., & Lee, L. (1993, June). *Distributional clustering of English words*. Paper presented at the 31st Annual Meeting on Association for Computational Linguistics; Columbus, OH USA.

Praetorius, M. (1615). *Syntagma musicum III* (3 Vols). Wittenberg: Johannes Richter.

Quinn, I. (2005, Nov). *Harmonic function without primary triads*. Paper presented at the Annual Meeting of Society for Music Theory, Boston/Cambridge, MA, USA.

Quinn, I., & Mavromatis, P. (2011). Voice-leading prototypes and harmonic function in two chorale corpora. In C. Agon, M. Andreatta, G. Assayag, E. Amiot, J. Bresson, & J. Mandereau (Eds.), *Mathematics and computation in music*, Vol. 6726 (pp. 230–240). Berlin: Springer.

Raphael, C., & Stoddard, J. (2004). Functional harmonic analysis using probabilistic models. *Computer Music Journal, 28*(3), 45–52.

Rohrmeier, M. (2005). *Towards modelling movement in music: Analysing properties and dynamic aspects of pc set sequences in Bach's chorales* (MPhil dissertation), University of Cambridge, UK. (Published as Darwin College Research Report 04.)

Rohrmeier, M., & Cross, I. (2008, Aug.). *Statistical properties of harmony in Bach's chorales*. Paper presented at the the 10th International Conference on Music Perception and Cognition, Sapporo, Japan,

Sadai, Y., Davis, J., & Shlesinger, M. (1980). *Harmony in its systemic and phenomenological aspects*. Jerusalem, Israel: Yanetz.

Schenker, H. (1925). *Das Meisterwerk in der Musik*, Vol. 1. New York, NY: G. Olms.

Schenker, H. (1935/1979). *Der frei Satz*. Vienna: Universal Edition. [Published in English as Free Composition (trans. and ed. E. Oster). London: Longman.]

Schneiderman, E., Slonim, N., Tishby, N.,de Ruyter van Steveninck, R.R., & Bialek, W., (2002). *Analyzing neural codes using the information bottleneck method (Technical Report)*. Jerusalem, Israel: The Hebrew University.

Schoenberg, A. (1969). *Structural functions of harmony*. New York, NY: WW Norton & Company.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review, 5*(1), 3–55.

Slonim, N., Friedman, N., & Tishby, N. (2006). Multivariate information bottleneck. *Neural Computation, 18*(8), 1739–1789.

Slonim, N., & Tishby, N. (2000). *Document clustering using word clusters via the information bottleneck method*. In *SIGIR '00 Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 208-215) New York, NY: ACM.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Temperley, D. (2009). A statistical analysis of tonal harmony. http://theory.esm.rochester.edu/temperley/kp-stats/index.html

Temperley, D., & VanHandel, L. (2013). Introduction to the special issues on corpus methods. *Music Perception: An Interdisciplinary Journal, 31*(1), 1–3.

Tillmann, B., & Bigand, E. (2004). The relative importance of local and global structures in music perception. *The Journal of Aesthetics and Art Criticism, 62*(2), 211–222.

Tishby, N., Pereira, F., & Bialek, W. (1999, Sept.). *The information bottleneck method*. Paper presented at the 37th Annual Allerton Conference on Communication, Control and Computing; Allerton, IL, USA.

Tishby, N., & Polani, D. (2011). Information theory of decisions and actions. In V. Cutsuridis, A. Hussain, & J. G. Taylor (Eds.), *Perception-action cycle* (pp. 601–636). Berlin: Springer.

Tymoczko, D. (2003). Progressions fondamentales, functions, degrés: Une grammaire de l'harmonie tonale élémentaire. *Musurgia, 10*(3–4), 35–64.

Tymoczko, D. (2006). The geometry of musical chords. *Science, 313*(5783), 72–74.

Tymoczko, D. (2011). *A geometry of music: Harmony and counterpoint in the extended common practice*. New York, NY: Oxford University Press.

White, C.W., & Quinn, I. (2014). *The Yale-classical archives corpus*. In International Conference for Music Perception and Cognition: Seoul, Korea. http://ycac.yale.edu

## Appendix A. Details of implementation of the Tishby et al. (1999) algorithm

Tishby et al. (1999) introduced a type of distributional clustering method (Pereira et al. (1993)). The algorithm works with the complexity measure of Table 1, case A, $I(F; X) = I(C_n; F_n)$, and any of the accuracy measures in Table 3, cases A–E, but not with the functional predictive metric in Table 3, case F. For the accuracy measures in Table 3, cases A–E the accuracy is described by $I(Y'; F)$, where $Y'$ is some combination of one or more adjacent chords. Note that in this case that the algorithm does not require as input the actual sequences of chord tokens $y_n$, but only the probability of each sequence of chord-token and context. For the first-order cases A and B in Table 3, this can be given by a histogram of consecutive chords derived from the corpus. For the cases of C–D in Table 3 it requires computing histograms of all sequences of length three or four. As mentioned in part 1, the central question in our approach is 'for *any given amount* of information loss, what categorization allows us to make the most accurate predictions?' Formally:

**Problem 1.** *For all possible $p(F|X)$ find one that maximizes $I_a(F) = I(F; Y')$ given that $I_c(F) = I(F; X) \leq I_C^0$.*

The constant $I_c^0$ is a bound on the complexity which the chosen solution $I(F; X)$ will not exceed. To find the optimal curve we scan a large sample of possible $I_c^0$s. This optimization problem can be solved using the well-known trick of Lagrange multipliers (Boyd & Vandenberghe, 2004). This trick relies on the fact that all solutions of problem 1 are given as solutions to the following problem:

**Problem 2.** *For all possible $p(F|X)$ find the one that minimizes $\mathcal{L}'(F) \equiv I(F; X) - \beta I(F; Y')$.*

Intuitively, the positive constant $\beta$ determines the relative importance of the conflicting objectives $I(F; X)$ and $I(F; Y')$. This definition is almost identical to the combined score $\mathcal{L}(F)$ introduced in Equation 17 of Section 1.9.

The only difference is that we minimize the Lagrangian of Problem 2 and in Section 1.9 we maximized the combined score. Tishby et al. (1999) is an iterative algorithm for finding the optimal theory for a given $\beta$. Here the inputs are $p(X)$, and $p(Y'|X)$ and the output is an optimal theory $p(F|X)$. We can interpret $\beta$ as a constant representing the theorist's relative weighting of simplicity and accuracy; in practice we scan all possible $\beta$s.

When the accuracy metric is a vector (for example $Y' = (C_n, C_{n+1})$), the sum over $y'$ scans all possible vector values (If $Y' = (C_n, C_{n+1})$; this requires considering all pairs $y' = (c_n, c_{n+1})$, where $c_n, c_{n+1} \in \mathcal{C}$). This computation is exponential in the length of $Y'$ as a vector, which is why we mostly consider very short $Y'$ in this article.

Algorithm 1 converges to a local minimum (the objective of Problem 2), but there is no guarantee that it will converge as well to the global minimum. However, if we want to find the optimal curve, we can follow the 'reverse deterministic annealing' procedure discussed in Slonim, Friedman and Tishby (2006) to obtain the optimal curve, which is a global minimum for all $\beta$.

**Algorithm 1** the Information bottleneck algorithm (Tishby et al. 1999)

```
***Input:*** $p(y|x)$, $p(x)$, $\beta$
***Output:*** $p(f|x)$
***Initialization:*** *randomize* $p_0\,(f|x)$
***Pseudo code:*** *iterate the following equations:*
    1. $p_t(f) = \sum_x p(x) p_t(f|x)$
    2. $p_t\,(x|f) = p_t\,(f|x)\,p(x)/p_t(f)$
    3. $p_t\,(y|f) = \sum_x p\,(y|x)\,p_t\,(x|f)$
    4. $Z_t(x,\beta) = \sum_f p_t\,(f)\exp\left(-\beta\sum_y p(y|x)\log\frac{p(y|x)}{p_t(y|f)}\right)$
    5. $p_{t+1}\,(f|x) = \frac{p_t(f)}{Z_t(x,\beta)}\exp\left(-\beta\sum_y p(y|x)\log\frac{p(y|x)}{p_t(y|f)}\right)$
```

Another version of the algorithm allows us to limit the number of categories:

**Problem 3.** *From all possible $p(F|X)$ such that $|\mathcal{F}| = k$ find one that minimizes: $I(F;X) - \beta I(F;Y')$.*

As discussed in Tishby et al. (1999), we go through all the steps of Algorithm 1, but instead of using $p(F|X)$ as a general matrix of size $|C|$ by $|C|$ (where $|C|$ is the number of surface tokens) we use a matrix of size $k$ by $|C|$. This always gives $p(F|X)$ with $|\mathcal{F}| = k$, and the resulting algorithm converges again to a local minimum. Here there is no simple way to guarantee convergence to the global minimum. Therefore, in practice we use simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) with the optimization criterion of Problem 3 $I(F;X) - \beta I(F;Y)$ as a score function.

In theory, simulated annealing is not guaranteed to converge to global minima. But when the problem is small and the algorithm is given enough iterations a global optimum can be obtained. This can be verified by running the algorithm multiple times with different initial conditions. For small problems with a few hundred surface tokens it converges in a few seconds to the global optimum. In this article, we determined that no further solutions would be found even if we ran the algorithm for several hours. Note that for continuous multivariate Gaussian distributions, Chechik, Globerson, Tishby and Weiss (2005) provide an analytic solution to the information bottleneck problem (the continuous version of algorithm 3) that finds the global maximum.

Importantly, this algorithm can be also used to find deterministic categorizations. Solving our problem with very large $\beta$ produces deterministic mappings (i.e. ones in which $p(F = f|X = x)$ is effectively 1 or 0 for each $(f, x)$ pairs); this is because the exponent in pseudo-code line 5 (Algorithm 1, above) tends towards either 0 (for cases where $x$ becomes a member of category $f$) or a large negative number (for cases where $x$ is not a member of category $f$), hence $p(F = f|X = x)$ tends towards either 1 or 0. Therefore when we have large $\beta$ we can effectively compute the optimal solution for the complexity metric given by the number of categories (as in Table 2, case A). To obtain deterministic categorization to $k$ clusters such as those reported in the results section, we choose large $\beta$ ($\beta > 100$), and verify that we have achieved the global maximum by multiple runs with different initial conditions.

### Functional predictive clustering (Table 3, case F)

If we apply functional predictive clustering (Table 3, case F), we cannot use the Information Bottleneck algorithm. However, we can apply simulated annealing directly on the Lagrangian:

$$\mathcal{L}'(F) = I(F_n; C_n) - \beta I(F_n; F_{n+1}).$$

This works either with deterministic (Definition 1) or probabilistic categories (Definition 2). This is in fact a variant of the algorithm proposed by Friedman and Goldberger (2013) for the deterministic case. However, that this algorithm usually converges much more slowly than in the Information Bottleneck case (in our simulations about 100 times slower) and is also more sensitive to the problem of obtaining a local rather than global minimum. That said, for pre-existing theories accuracy and complexity are easily computable in the functional predictive case; we simply apply the formulae in Tables 2 and 3.

## Appendix B. Additional tables of optimal deterministic categories

Table B1. Comparison of optimal functional labels of datasets 1–6 from three published works: Huron (2006), Temperley (2009) and Tymoczko (2011a).

| Corpus | | Deterministic Optimal Categories | |
|---|---|---|---|
| 1 | Tymoczko (2011a), figure 7.1.6 page 230: a selection of major-mode Bach chorales (7 surface tokens) | 2 | 1: I,IV,II,VI,III<br>2: V, VII |
| | | 3 | 1: I,VI, III<br>2: IV, II<br>3: V, VII |
| | | 4 | 1: I, III<br>2: II<br>3: IV, VI<br>4: V, VII |
| 2 | Tymoczko (2011a), figure 7.1.6 page 230: a selection of minor-mode Bach chorales (7 surface tokens) | 2 | 1: I, II ,III, IV, VI<br>2: V, VII |
| | | 3 | 1: I, VI, III<br>2: II, IV<br>3: V, VII |
| | | 4 | 1: I, VI, III<br>2: II<br>3: IV<br>4: V, VII |
| 3 | Tymoczko (2011a), figure 7.1.6 page 230: 19 major-mode Mozart Piano sonatas (7 surface tokens) | 2 | 1: I, II, VI<br>2: V, IV, VII, III |
| | | 3 | 1: I, II, VI<br>2: IV, VII<br>3: V, III |
| | | 4 | 1: I, II<br>2: III, V, VII<br>3: IV<br>4: VI |
| 4 | Tymoczko (2011a), figure 7.1.6 page 230: 19 minor-mode Mozart piano sonatas (7 surface tokens) | 2 | 1: I, II, VI, III<br>2: V, VII, IV |
| | | 3 | 1: I, II, VI, III<br>2: VII, IV<br>3: V |
| | | 4 | 1: I, II<br>**2: VII, IV**<br>3: V<br>4: VI, III |
| 5 | Huron (2006), table 13.2 page 251: major-mode Baroque music (7 surface tokens) | 2 | 1: I, IV, II, VI, III<br>2: V, VII |
| | | 3 | 1: I, III<br>2: IV, II, VI, VII<br>3: V |
| | | 4 | 1: I, III<br>2: II, VI<br>3: IV, VII<br>4: V |
| 6 | Temperley (2009) aggregate statistics (6/4 reanalyzed): major-mode excerpts from Stefan Kostka and Dorothy Payne's theory textbook Tonal Harmony (12 surface tokens) | 2 | 1: I, II, VI, bVI, III, #IV, bII, bIII, bVII<br>2: V, IV, VII |
| | | 3 | 1: I, II, III, #IV, bIII, bVII<br>2: IV, VI, bVI, bII<br>3: V, VII |
| | | 4 | 1: I, III<br>**2: II, #IV, bIII, bVII**<br>3: IV, VI, bVI, bII<br>4: V, VII |
| | | 5 | 1: I, III<br>**2: II, #IV, bIII**<br>3: IV, bVII<br>4: V, VII<br>5: VI, bVI, bII |

Table B2. Comparison of optimal functional labels of datasets 7–11 from two published works: Rohrmeier and Cross (2008) and de Clercq and Temperley (2011).

| Corpus | | Deterministic Optimal Categories | |
|---|---|---|---|
| 7 | de Clercq & Temperley (2011), tables 2-3 pages 60-61: 100 major-mode Rock songs (12 surface tokens) | 2 | 1: I, III<br>2: IV, V, bVII, VI, bVI, II, bIII, bII, VII, #IV |
| | | 3 | 1: I<br>2: IV, bVI, bII, VII<br>3: V, bVII, VI, II, bIII, III, #IV |
| | | 4 | 1: I<br>2: IV, bVI ,bII, VII<br>3: V, bVII, bIII, #IV<br>4: VI, II, III |
| | | 5 | 1: I<br>2: IV, bVI, bII, VII<br>3: V, III<br>4: VI, II<br>5: bVII, bIII, #IV |
| | | 6 | 1: I<br>2: III<br>3: IV, bVI, bII, VII<br>4: V<br>5: VI, II<br>6: bVII, bIII, #IV |
| | | 7 | 1: I<br>2: III<br>3: bVI, II, VII<br>4: IV, bII<br>5: V<br>6: VI<br>7: bVII, bIII, #IV |
| 8 | Reduction of data taken from de Clercq & Temperley (2011), tables 2-3 pages 60-61: 100 major/minor mode Rock songs (7 surface tokens) | 2 | 1: I<br>2: IV, V, VI, VII, III, II |
| | | 3 | 1: I<br>2: IV<br>3: V, VI, VII, III, II |
| | | 4 | 1: I<br>2: IV<br>3: V, VII<br>4: VI, III, II |
| 9 | Sequences of two adjacent chords taken from de Clercq & Temperley's dataset (2011): 100 major/minor mode Rock songs (101 surface tokens) (*) only common bigrams are displayed. | 2* | 1: II->I, IV->I, V->I, VI->I, I->II, VI->II, I->IV, V->IV, VI->IV, I->V, II->V, IV->V, VI->V, I->VI, V->VI<br>2: bVI->I,bVII->I,I->bIII,bVII->IV,I->bVI, bVII->bVI,I->bVII,bVI->bVII |
| 10 | Rohrmeier & Cross (2008) table 1-2 pages 8-9: Simultaneous transposed pitch classes extracted from MIDI renditions of major-mode Bach chorales (15 surface tokens) | 2 | 1: DGB, DFGB, DFB<br>2: CEG, CFA, DFA, DF#A, EGB, CDF#A, EG#B, CDFA, CEGA, |
| | | 3 | 1: DF#A, CDF#A, CDFA, CDG<br>**2: EG#B**<br>3: CEG, DGB, CFA, DFA, DFGB, EGB, CEGA, DFB, C#EA |
| | | 4 | 1: DF#A, CDF#A, CDFA, CDG<br>2: DGB, DFGB, DFB<br>3: EG#B<br>4: CEG,CFA,CEA, DFA,EGB,CEGA, C#EA |
| | | 7 | 1: CEG, CFA<br>2: DFA<br>3: EG#B, CEA<br>4: DGB, DFGB, DFB<br>5: CEGA, C#EA<br>6: DF#A, CDF#A, CDFA, CDG<br>7: CEA,EGB |
| 11 | Rohrmeier & Cross (2008) table 1-2 pages 8-9: Simultaneous transposed pitch classes extracted from MIDI renditions of minor-mode Bach chorales (15 surface tokens) | 2 | 1: CEbG, DGB, CFAb, CEG, DFGB, CDFAb, DFAb, DF#A<br>2: EbGBb, DFBb, CEbAb, DGBb, CFA, CEbFA, DFAbBb |
| | | 4 | 1: EbGBb, DFBb, CEbAb, CFA, CEbFA, DFAbBb<br>2: CEbG, CFAb, CDFAb, DFAb<br>3: CEG<br>4: DGB, DFGB |
| | | 5 | 1: CEbG, CFAb, CEG, CDFAb<br>2: DGB, DFGB<br>3: EbGBb, DFBb, CEbAb, DGBb, DFAbBb<br>4: CFA, CEbFA<br>5: DGBb, DF#A |

Table B3. Comparison of optimal functional labels in datasets 12–14 from unpublished datasets by Dmitri Tymoczko.

| Corpus | | Deterministic Optimal Categories | |
|---|---|---|---|
| 12 | Major-mode Bach chorales, 52 surface tokens, 9527 total chords in the corpus. | 2 | 1: I, I6, vi, IV, IV6, ii6/5, ii, V2, ii7, I6/4, ii6, iii, V6/5/V, IVmaj7, vi6, viio6/V, V7/IV, IVmaj6/5, V6/5/IV, iii6, vi7, V2/IV, ii2, vii/o4/3, V/V, V6/V, vii/o7/V, viio6/4, IVmaj2, vii/o6/5, V7/V, vi6/4, vii/o4/3/V, V2/V, Imaj7, V4/3/IV, V6/5/ii, v<br>2: V, V7, V6, viio6, V6/5, vii/o7, V4/3, V6/vi, viio, V6/5/vi, V6/4, viio7/vi, IV6/4, V/vi |
| | | 3 | 1: I, vi, V2, iii, vi6, V7/IV, V6/5/IV, V6/vi, iii6, V2/IV, V6/5/vi, vii/o4/3, viio6/4, vii/o6/5, V2/V, Imaj7, V4/3/IV,viio7/vi,V/vi, v<br>2: IV, IV6, ii6/5, ii, ii7, I6/4, ii6, V6/5/V, IVmaj7, viio6/V, IVmaj6/5, vi7, ii2, V/V, V6/V, vii/o7/V, IVmaj2, V7/V, vi6/4, vii/o4/3/V, V6/5/ii, I6<br>3: V, V7, V6, viio6, V6/5, vii/o7, V4/3, viio, V6/4, IV6/4 |
| | | 4 | 1: I. I6. vi. IV. IV6. iii. vi6. V7/IV. V6/5/IV. iii6. V2/IV. V6/5/vi. vi6/4. V2/V. Imai7. V4/3/IV. V6/5/ii. v<br>2: ii6/5. ii. ii7. I6/4. ii6. V6/5/V. IVmai7. viio6/V. IVmai6/5. vi7. ii2. V/V. V6/V. vii/o7/V. IVmai2. V7/V. vii/o4/3/V<br>3: V. V6<br>4: V7. viio6. V6/5. V2. vii/o7. V4/3. V6/vi. viio. vii/o4/3. viio6/4. vii/o6/5. V6/4. viio7/vi. IV6/4. V/vi |
| | | 5 | 1: I. I6. vi. IV. iii. vi6. V7/IV. V6/5/IV. iii6. V2/IV. V6/5/vi. Imai7. V4/3/IV. V6/5/ii. V/vi. v<br>2: ii6/5. ii7. I6/4. ii6. V6/5/V. IVmai7. viio6/V. V/V. V6/V. vii/o7/V. V7/V. vii/o4/3/V<br>3: IV6. ii. IVmai6/5. vi7. ii2. IVmai2. vi6/4. V2/V<br>4: V. V6<br>5: V7. viio6. V6/5. V2. vii/o7. V4/3. V6/vi. viio. vii/o4/3. viio6/4. vii/o6/5. V6/4. viio7/vi. IV6/4 |
| | | 6 | 1: I. I6. vi. IV. iii. vi6. V7/IV. V6/5/IV. iii6. V2/IV. Imai7. V4/3/IV. V6/5/ii. IV6/4. v<br>2: ii6/5. ii7. I6/4. ii6. V6/5/V. IVmai7. viio6/V. V/V. V6/V. vii/o7/V. V7/V. vii/o4/3/V<br>3: IV6. ii. IVmai6/5. vi7. ii2. IVmai2. vi6/4. V2/V<br>4: V<br>5: V7. V6. V6/5. vii/o7. V6/vi. viio. V6/5/vi. viio7/vi. V/vi<br>6: viio6. V2. V4/3. vii/o4/3. viio6/4. vii/o6/5. V6/4 |
| | | 7 | 1: I. vi. vi6. vi6/4. vi7. vi2. iii. iii6. iii7. iii6/5. iii4/3<br>2: I6. I6/4. ii6/5<br>3: IV. ii. .ii6. ii6/4. ii7. vi6/5. vi4/3<br>4: IV6. IV6/4. ii4/3. ii2. viiø2<br>5: V<br>**6: V6. V7. V6/5. viio. viiø7. iii6/4**<br>**7:V6/4. V4/3. V2. viio6. viio6/4. viiø6/5. viiø4/3** |
| 13 | Minor-mode Bach chorales, 69 surface tokens, 5700 total chords in the corpus. | 2 | 1: i, i6, iv6, ii/o6/5, iv, VI, i6/4, V2, iio6, v, ii/o7, III, iv7, ii, viio6/5, v6, IV2, IV, ii/o4/3, VI6, III6, III+6, VII, viio2, iv6/5, iv6/4, v2, ii6, vio6, VII6, vi/o4/3, ii6/5, vi/o6/5, iv2, i7, i2, IV7, iio6/4, ii7, ii2, III7, III+, III+6/5, VI6/4, VI7, III6/4, v6/5, VI2, vio6/4, VII7<br>2: V. V7. V6. V6/5. viio6. viio7. IV6. IV6/5. ii/o2. viio. V4/3. viio4/3. vi/o7. V6/4. iio. v7. viio6/4. vio. iv4/3: |
| | | 3 | 1: i, i6, iv6, ii/o6/5, iv, VI, i6/4, V2, iio6, v, ii/o7, III, iv7, ii, viio6/5, v6, IV2, IV, ii/o4/3, VI6, III6, III+6, VII, viio2, iv6/5, iv6/4, v2, vio6, vi/o4/3, iio, ii6/5, vi/o6/5, iv2, i7, i2, IV7, iio6/4, ii7, ii2, III+, III+6/5, VI6/4, VI7, III6/4, VI2, vio6/4, VII7<br>2: V .ii6. vio. v6/5<br>3: V7. V6. V6/5. viio6. viio7. IV6. IV6/5. ii/o2. viio. V4/3. viio4/3. vi/o7. V6/4. VII6. v7. viio6/4. III7. iv4/3 |
| | | 4 | 1: i. V2. v. IV6. IV6/5. III. ii/o2. viio6/5. v6. VI6. III6. III+6. VII. vi/o7. iv6/4. v2. VII6. iio. i2. ii2. III7. III+. III+6/5. VI7. III6/4. VII7<br>2: i6, iv6, ii/o6/5, iv,VI, i6/4, iio6, ii/o7, iv7, ii, IV2, IV, ii/o4/3, viio2, iv6/5, vio6, vi/o4/3, ii6/5, vi/o6/5, iv2, i7, IV7, iio6/4, ii7, VI6/4, v6/5, VI2, vio6/4<br>3: V7 .V6. V6/5. viio6. viio7. viio. V4/3. viio4/3. V6/4. v7. viio6/4. iv4/3<br>4:  V. ii6. vio |
| | | 5 | 1: i. VI. V2. v. III. viio6/5. v6. VI6. III6. III+6. v2. i7. i2. III7. III+. III+6/5. VI6/4. VI7. III6/4.  VI2. VII7<br>2: i6. iv6. ii/o6/5. iv. i6/4. iio6. ii/o7. iv7. ii/o4/3. viio2. iv6/5. vio6. ii6/5. vi/o6/5. IV7. iio6/4. ii7. v6/5<br>3: V7. V6. V6/5. viio6. viio7. viio. V4/3. viio4/3. V6/4. v7. viio6/4<br>4:  V. ii6. vio<br>5: IV6. IV6/5. ii/o2. ii. IV2. IV. VII. vi/o7. iv6/4. VII6. vi/o4/3. iio. iv2. ii2. iv4/3. vio6/4 |
| | | 6 | 1: i. V2. v. viio6/5. v6. VI6. III6. III+6. v2. i2. III+. III+6/5. VI7. III6/4<br>2: i6. iv. III. ii. IV2. IV. vio6. VII6. vi/o4/3. ii6/5. vi/o6/5. iv2. IV7. ii7. III7. vio6/4<br>3: V7. V6. V6/5. viio6. viio7. viio. V4/3. viio4/3. V6/4. v7. viio6/4<br>4: V. ii6. vio<br>5: IV6. IV6/5. ii/o2. VII. vi/o7. iv6/4. iio. ii2. iv4/3. VII7<br>6: iv6. ii/o6/5. VI. i6/4. iio6. ii/o7. iv7. ii/o4/3. viio2. iv6/5. i7. iio6/4. VI6/4. v6/5. VI2 |
| | | 7 | 1: i. v. v6. III6. III+6. i2.III+. VI7. III6/4<br>2: i6. iv. III. ii. IV2. IV.V I6. vio6. VII6. vi/o4/3. vi/o6/5. iv2. IV7. ii7. III7. vio6/4<br>3: V7. V6. V6/5. viio7. viio. v7<br>4: V. ii6. vio<br>5: IV6. IV6/5. ii/o2. VII. vi/o7. iv6/4. iio. ii2. iv4/3. VII7<br>6: iv6. ii/o6/5. VI. i6/4. iio6. ii/o7. iv7. ii/o4/3. viio2. iv6/5. ii6/5. i7. iio6/4. VI6/4. v6/5. VI2<br>7: viio6. V2. viio6/5. V4/3. viio4/3. V6/4. v2. viio6/4. III+6/5 |
| 14 | Major-mode Mozart piano sonatas, 38 surface tokens, 7826 total chords in the corpus. | 2 | 1: I. I6/4. I6. ii6. IV. vi. ii. IV6. ii6/5. vi7. viio2. ii7. vi6. ii4/3. ii6/4. ii2. vi2. vi6/4. I2. IV6/5<br>2: V7. V. V6/5. V4/3. V2. V6. IV6/4. viio. viio6. V6/4. viio4/3. viio6/4. viio7. vii/o7. vii/o4/3. iii. vii/o6/5. viio6/5 |
| | | 3 | 1: I. I6. IV. vi. ii. IV6. vi7. ii7. vi6. ii2. vi6/4. IV6/5<br>2: V7. V. V6/5. V4/3. V2. V6. IV6/4. viio. viio6. V6/4. viio4/3. viio6/4. viio7. vii/o7. vii/o4/3 .iii. vii/o6/5. viio6/5<br>3: I6/4. ii6. ii6/5. viio2. ii4/3. ii6/4. vi2. I2 |
| | | 4 | 1: I. I6. vi. ii. vi7. ii7. vi6. ii2. vi6/4. IV6/5<br>2: V7. V. V6/5. V4/3. V6. IV6/4. viio. viio6. V6/4. viio7. vii/o7. iii. vii/o6/5. viio6/5<br>3: I6/4. viio2. ii4/3. ii6/4 .vi2. I2<br>4: ii6. IV. V2. IV6. ii6/5. viio4/3. viio6/4. vii/o4/3 |
| | | 5 | 1: I. I6. vi. ii. vi7. ii7. vi6. ii2. vi6/4. IV6/5<br>2: V7. V. V6/5. V4/3. V6. IV6/4. viio. viio6. V6/4. viio7. vii/o7. iii. vii/o6/5. viio6/5<br>3: I6/4. viio2. ii4/3. ii6/4. vi2<br>4: ii6. IV. IV6. ii6/5<br>5: V2. viio4/3. viio6/4. vii/o4/3. I2 |
| | | 6 | 1: I. ii. vi7. ii7. ii2. IV6/5<br>2: V7. V. V6/5. V4/3. V6. IV6/4. viio. viio6. V6/4. viio7. vii/o7. iii. vii/o6/5. viio6/5<br>3: I6/4. viio2. ii4/3. ii6/4. vi2<br>4: ii6. IV. IV6. ii6/5<br>5: V2. viio4/3. viio6/4. vii/o4/3. I2<br>6: I6. vi. vi6. vi6/4 |
| | | 7 | 1: I. ii. vi7. ii7. ii2. IV6/5<br>2: V7. V. V6/5. V4/3. V6. IV6/4. viio. viio6. V6/4. viio7. vii/o7. iii. vii/o6/5. viio6/5<br>3: I6/4. viio2. ii6/4. vi2<br>4: ii6. IV6. ii6/5. ii4/3<br>5: V2. viio4/3. viio6/4. vii/o4/3. I2<br>6: I6. vi6. vi6/4<br>7: IV. vi |

Table B4. Comparison of optimal functional labels of datasets 15–16 from unpublished datasets by Dmitri Tymoczko.

| Corpus | | Deterministic Optimal Categories | |
|---|---|---|---|
| 15 | Minor-mode Mozart piano sonatas, 40 surface tokens, 1602 total chords in the corpus. | 2 | 1: i, i6/4, i6, iv, iio6, iv6, VI, viio6/5, iio, ii/o6/5, ii, VI7, ii/o7, viio2, IV6, ii/o4/3, ii6, IV, VI6, vio, VII, ii/o2, ii6/4<br>2: V, V7, V6/5, V6, V4/3, viio4/3, V2, viio7, viio6, viio, V6/4, iv6/4, v, v6, i2, III6, viio6/4 |
| | | 3 | 1: i, i6, viio4/3, V2, VI, viio6/5, iio, VI7, ii/o7, IV6, IV, VI6, VII, ii/o2, ii6/4, III6, viio6/4<br>2: V, V7, V6/5, V6, V4/3, viio7, viio6, viio, V6/4, iv6/4, v, v6, i2<br>3: i6/4, iv, iio6, iv6, ii/o6/5, ii, viio2, ii/o4/3 ,ii6, vio |
| | | 4 | 1: i, i6, VI, iio, VI7, ii/o7, IV6, IV, vio, VII, ii6/4<br>2: V, V7, V6/5, V6, viio6, viio, v6, i2<br>3: i6/4, iv, iio6, iv6, ii/o6/5, ii, viio2, ii/o4/3, ii6, III6<br>4: V4/3, viio4/3, V2, viio7, viio6/5, V6/4, iv6/4, v, VI6, ii/o2, viio6/4 |
| | | 5 | 1: i ,i6, iio, iv6/4, VI7, ii/o7, IV6, IV, VII, ii6/4<br>2: V, V7, V6/5, V6, viio6, viio, v, v6, i2<br>3: i6/4, ii/o6/5, ii, viio2, ii/o4/3, ii6, vio<br>4: V4/3, viio4/3, V2, viio7, viio6/5, V6/4, ii/o2, III6, viio6/4<br>5: iv, iio6, iv6, VI, VI6 |
| | | 6 | 1: i, iio, ii/o7, IV6, VII, ii6/4, III6<br>2: V, V7, V6/5, V6, viio6, viio, v, v6, i2<br>3: i6/4, ii/o6/5, ii, viio2, ii/o4/3, ii6, vio<br>4: V4/3, viio4/3, V2, viio7, viio6/5, V6/4, iv6/4, ii/o2, viio6/4<br>5: iv, iio6, iv6, VI6<br>6: i6, VI, VI7, IV |
| | | 7 | 1: i, iio, ii/o7, VII, ii/o2 ,ii6/4, III6<br>2: V, V7, v6, i2<br>3: i6/4, ii/o6/5, ii, viio2, ii/o4/3, ii6,vio<br>4: viio4/3, V2, iio6/5, viio6/4, v<br>5: iv, iio6, iv6, VI6<br>6: i6, VI, VI7, IV6, IV<br>7: V6/5, V6, V4/3, viio7, viio6, viio, V6/4, iv6/4 |
| 16 | Palestrina | 2 | 1: I, IV, vi, I6, ii, IV6, iii, ii6, iii6, vi6, I6/4, viio6/V, V/V, ii6/5, ii7, bVII, v, V2, IVmai7, IVmai6/5, V6/V, v6<br>2: V, V6, viio6, V7, viio, IV6/4, V6/5, vi7, ii7/V, iii6/5 |
| | | 3 | 1: I, vi, IV6, vi6, I6/4, viio6/V, V/V, ii6/5, bVII, v, IVmai6/5, V6/V, v6<br>2: IV, I6, ii, iii, ii6, iii6, ii7, V2, IVmai7<br>3: V, V6, viio6, V7, viio, IV6/4, V6/5, vi7, ii7/V, iii6/5 |
| | | 4 | 1: I, vi, iii6, vi6, bVII, v, vi7, ii7/V, v6<br>2: IV, I6, ii, iii, ii6, ii7, V2, IVmai7<br>3: IV6, I6/4, viio6/V, V/V, ii6/5, IVmai6/5, V6/V<br>4: V, V6, viio6, V7, viio, IV6/4, V6/5, iii6/5 |
| | | 5 | 1: I, vi, vi6, v, vi7, ii7/V, v6<br>2: I6, ii7, IVmai7<br>3: IV, ii, iii, ii6, iii6, bVII, V2<br>4: IV6, I6/4, viio6/V, V/V, ii6/5, IVmai6/5, V6/V<br>5: V, V6, viio6, V7, viio, IV6/4, V6/5, iii6/5 |
| | | 6 | 1: I, vi, vi6, bVII, v, vi7, v6<br>2: ii, ii6, ii7, V2<br>3: IV6, IVmai6/5, iii6/5<br>4: IV, I6, iii, iii6, IVmai7<br>5: I6/4, viio6/V, V/V, ii6/5, ii7/V, V6/V<br>6: V, V6, viio6, V7, viio, IV6/4, V6/5 |
| | | 7 | 1: I, vi, vi6, bVII, v, vi7, v6<br>2: ii, ii6, ii7, V2<br>3: IV, I6, iii, iii6, IVmai7<br>4: IV6, IVmai6/5, iii6/5<br>5: V6, viio6, V7, viio, IV6/4, V6/5, ii7/V<br>6: I6/4, viio6/V, V/V, ii6/5, V6/V<br>7: V |