Exploring Optimal Transport in Jazz Music Analysis Application to the Real Book

Jean Dufraiche [1] & Valérie Garès [2] & Madison Giacofci [3]& Nicolas Klutchnikoff [4]

[1] *INSA Rennes, France; jean.dufraiche@insa-rennes.fr*
[2] *Univ Rennes, INSA, IRMAR - UMR 6625, France; valerie.gares@insa-rennes.fr*
[3] *Univ Rennes, IRMAR - UMR 6625, France; joyce.giacofci@univ-rennes2.fr*
[4] *Univ Rennes, IRMAR - UMR 6625, France; nicolas.klutchnikoff@univ-rennes2.fr*

Résumé. Cet article se penche sur l'analyse mathématique du Real Book, un célèbre corpus de musique de jazz. Pour simplifier le problème, nous supposons que chaque pièce musicale est exprimée comme une séquence d'accords. Notre approche introduit une représentation des accords basée sur leurs emprunts aux différents modes pythagoriciens, une dimension qui semble négligée dans la littérature existante. Cette représentation innovante permet d'établir des dissimilarités entre les accords, ce qui constitue la base de la comparaison des pièces. Plus précisément, deux pièces de Real Book peuvent être représentées par la distribution empirique de leurs accords. La dissimilarité entre ces pièces est alors déterminée par le coût de transport optimal entre leurs distributions respectives. Ce calcul repose sur le coût entre accords défini précédemment.

Mots-clés. Analyse musicale par ordinateur, transport optimal, classification non supervisée

Abstract. This article delves into the mathematical analysis of the Real Book, a renowned collection of jazz music. To simplify the problem, we assume that each musical piece is expressed as a sequence of chords. Our approach introduces a chord representation based on their borrowings from various Pythagorean modes, a dimension that seems to be neglected in the existing literature. This innovative representation allows us to establish dissimilarities between chords, which form the basis for comparing songs. Specifically, two Real Book pieces can be represented by the empirical distribution of their chords, with dissimilarity determined as the optimal transport cost between them. This calculation is based on the previously defined chord costs.

Keywords. Music Information Retrieval, optimal transport, clustering

1 Introduction

When studying Western music with the use of symbolic music data (score, Midi, MusicXML, etc.), a central aspect that one would like to describe is the harmony. For this reason, numerous representations and distances of chords have been defined in the field of music information retrieval (see Tymoczko, 2006; Rocher, Robine, and Hanna, 2010, and references therein).

These modelings have been used for various applications, such as recommandation algorithms (Aucouturier and Pachet, 2002) or to improve the performance of chord label recognition algorithms on signal-based data (Mauch, Noland, and Dixon, 2009), and it is assessed that it could help define goodness-of-fit measures when producing chord recognition on audio signal (Oudre, Grenier, and Févotte, 2011).

The link between chords and modes is usually not considered when we define such geometric representations; therefore, in our work, we propose a representation and a distance that depend on the potential belonging of a chord to a family of modes. This representation and distance might be of great interest for machine learning problems for which we need a distance or dissimilarity on the chords being the explanatory variables. We could think about clustering in unsupervised situations or k-nearest neighbors in supervised situations.

We applied this representation on a chord progressions dataset (de Berardinis, Meroño-Peñuela, Poltronieri, and Presutti, 2023) to find similarity in music extracts.

The outline is the following: having a corpus of songs for which we know their sequences of chords, we compute empirical distributions of chords for each of these songs. Then, using the Wasserstein metric on these empirical distributions based on a cost that we define on chords as in Givens and Shortt (1984), we obtain a pairwise distance between two songs.

2 Representation of chords and modes

Pitch and note. We rely on the MTS (Midi) format. In this standard, each frequency ν is mapped to a real number, called the **pitch**, according to the following function:

$$p(\nu) := 69 + 12\log_2(\nu/440), \quad \forall \nu > 0.$$

Although p is a real-valued function, it is constructed to take integer values for the pitches used in the theory of Western music. In the following, we therefore assume that any pitch belongs to $\mathbb{N}$.

If two frequencies ν_1 and ν_2 are such that $\log_2(\nu_1/\nu_2) \in \mathbb{Z}$, then $p(\nu_1) \equiv p(\nu_2) \mod 12$. This corresponds to the situation where the two sounds are highly consonant, because they are separated by several octaves. The **pitch class** of p, denoted by $[p]$ or simply p if the context allows, is then defined as its class in $\mathbb{Z}_{12} \cong [\![0, 11]\!]$. In this paper, we use the term **note** as a synonym for pitch class, although it is generally used to designate a set of properties such as pitch and duration associated with performance elements. However, there is no ambiguity in our context. The following table shows the correspondence between notes in $\mathbb{Z}_{12}$ and their names in English and French.

Note	0	1	2	3	4	5	6	7	8	9	10	11
English	C	C♯/D♭	D	D♯/E♭	E	F	F♯/G♭	G	G♯/A♭	A	A♯/B♭	B
French	Do	Do♯/Ré♭	Ré	Ré♯/Mi♭	Mi	Fa	Fa♯/Sol♭	Sol	Sol♯/La♭	La	La♯/Si♭	Si

Chord. When multiple sounds are played simultaneously (e.g. several keys on the piano), the corresponding musical element is called a chord that is characterized, in this work, by the set of the notes it contains. In harmony analysis, though, a specific note of a chord, called the **root**, acts as a foundational anchor, providing a reference point for the other notes within the chord. This root serves as a musical cornerstone that influences the perception and interpretation of the entire chordal structure. The significance of the root lies in its ability to establish a sense of stability and grounding. In the remainder of this paper, we make the distinction between an **unrooted chord**, which is simply a set of notes, and the more subtle notion of **chord**, which includes the notion of root.

More precisely, an unrooted chord is a collection $\mathcal{C} \subset \mathbb{Z}_{12}$ that is identified with its *one-hot encoding*, namely, the binary vector $b = (b_0, \dots, b_{11})$ defined by $b_i = 1$ if $i \in \mathcal{C}$ and $b_i = 0$ otherwise (see Fujishima, 1999). A rooted chord, or more simply a chord, is a pair consisting of a set $\mathcal{C} \subset \mathbb{Z}_{12}$ and a particular note $r \in \mathcal{C}$, called the root.

For example, the chords *Cmajor* and *Aminor* respectively have roots C and A and the terms *major* and *minor* entirely define the other notes contained in *Cmajor* and *Aminor*. We then use a similar notation to Harte, Sandler, Abdallah, and Gómez (2005) that uses two features: the **root** $r \in \mathbb{Z}_{12}$ and the **kind** $k \in \mathcal{K} := \{0,1\}^{11}$, defined below.

For given unrooted chord $\mathcal{C}$ and root $r \in \mathcal{C}$, let us describe how the **kind** k is deduced from the one-hot encoding $b = (b_0, \dots, b_{11})$. Let us consider the example of the unrooted chord $\mathcal{C} = \{0, 4, 7, 9\}$ that has the following one-hot encoding:

$$b = (\underbrace{1}_{C}, 0, 0, 0, \underbrace{1}_{E}, 0, 0, \underbrace{1}_{G}, 0, \underbrace{1}_{A}, 0, 0).$$

If the root of this chord is chosen to be C ($r = 0$), then the other notes of $\mathcal{C}$ can be deduced from the positions of the non-zero elements in b, see equation above. Now, if the root of the chord is different, for example A ($r = 9$), then by applying a circular shift on the vector b, we can obtain a new vector such that we can deduce the other notes using the same process:

$$(\underbrace{1}_{C}, 0, 0, 0, \underbrace{1}_{E}, 0, 0, \underbrace{1}_{G}, 0, \underbrace{\mathbf{1}}_{\mathbf{A}}, \mathbf{0}, \mathbf{0}) \mapsto (\underbrace{\mathbf{1}}_{\mathbf{A}}, \mathbf{0}, \mathbf{0}, \underbrace{1}_{C}, 0, 0, 0, \underbrace{1}_{E}, 0, 0, \underbrace{1}_{G}, 0)$$

The vector b is shifted so that b_0 corresponds to the root of the chord. It allows one to determine the other notes of $\mathcal{C}$ with the computations $[0] = [9+3]$, $[4] = [9+7]$, $[7] = [9+10]$.

Notice that we did the same process twice, but it was trivial when $r = 0$. Then, we define the kind k as the 11 last components of the circular shifted vector that put the root on the first component.

Let $b \in \{0,1\}^{12}$ be the one-hot encoding of a chord, and let $r \in \mathbb{Z}_{12}$ be the root of b. We denote the kind:

$$k = (k_1, \dots, k_{11}) \in \mathcal{K},$$

where $k_i = 1$ if and only if there exists a note p in the chord $\mathcal{C}$ such that $[r + i] = [p]$. Then a chord can be defined by the elements (r, b) or (r, k). In the following, we choose the second solution and define a chord as a pair (r, k) consisting of a fundamental note $r \in \mathbb{Z}_{12}$ and a kind $k \in \mathcal{K}$.

Mode. When a musical segment uses only a subset of notes, say $\mathcal{P}$, in its compositional choices, whether it contains chords (simultaneous) or melodies (consecutive), the elements of $\mathcal{P}$ are often perceived in relation to a fundamental note. This provides a representation similar to that of chords, in the form of a pair $(r, \ell) \in \mathbb{Z}_{12} \times \mathcal{K}$. In this context, we refer to ℓ as a **mode** and to (r, ℓ) as a **rooted mode**. By definition, $\ell_i = 1$ if and only if there exists a note $p \in \mathcal{P}$ such that $r + i \equiv p \mod 12$.

The two most famous modes are undoubtedly *major* and *minor* which correspond to $\ell = (0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1)$ and $\ell = (0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0)$, respectively. We can also mention Debussy's unitonal mode, which corresponds to $\ell = (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)$, as well as Messiaen's limited transposition modes.

At this point, it should be noted that musicians do not necessarily attach the same importance to all the pitches of a mode. To take this property into account, we can introduce a weight $w = (w_1, \ldots, w_{11})$ to reflect the importance of the different pitches present in a mode. Although the choice of a weight may seem subjective, it is possible to use the style of a musical corpus to give a reasonable definition in such a context.

Pythagorean modes. Pythagorean modes are undoubtedly the seven most widely used modes in Western music. They can be visually represented using the white keys on a piano. For example, constructing the Dorian mode involves playing a scale starting from D and corresponds to $\ell = (0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0)$. In jazz, musicians often employ modal borrowings as a technique to infuse color and diversity into their performances. Pythagorean modes serve as a structured foundation for generating tension and resolution, allowing musicians to create complex harmonic progressions.

In this context, it is well-known that specific pitches hold greater significance than others. This applies to the tonic triad—composed of the third and fifth notes of the mode—and the two notes within the mode separated by a tritone, spanning 6 semitones. With this in mind, we propose the following weights associated with each Pythagorean mode.

Name	ℓ	$10 \cdot w$
Lydian	$(0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1)$	$(0, 1, 0, 2, 0, 3, 2, 0, 1, 0, 1)$
Ionian	$(0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1)$	$(0, 1, 0, 2, 2, 0, 2, 0, 1, 0, 2)$
Mixolydian	$(0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0)$	$(0, 1, 0, 3, 1, 0, 2, 0, 1, 2, 0)$
Dorian	$(0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0)$	$(0, 1, 3, 0, 1, 0, 2, 0, 2, 1, 0)$
Aeolian	$(0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0)$	$(0, 2, 2, 0, 1, 0, 2, 2, 0, 1, 0)$
Phrygian	$(1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0)$	$(2, 0, 2, 0, 1, 0, 3, 1, 0, 1, 0)$
Locrian	$(1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0)$	$(1, 0, 2, 0, 1, 4, 0, 1, 0, 1, 0)$

Elaborating on the notion of modes and their associated weights, we define a **musical system** as the couple $(\mathbb{M}, \mathbb{W})$ consisting of both a family of M modes $\mathbb{M} = \{\ell^1, \ldots, \ell^M\}$ and a family of M weights $\mathbb{W} = \{w^1, \ldots, w^M\}$. Musical system will allow us, in the next section, to represent all the chords and define a distance between them.

3 Cost between chords

A strong link exists between chords and modes. For a given mode ℓ, chords can be created by selecting both a root r and a kind k such that $k_i \leq \ell_i,\ \forall i \in [\![1, 11]\!]$. In contrast, for a given chord (r, k), there are several possibilities of rooted modes (r, ℓ) from which this chord is extracted. The cost we define in the following takes full advantage of this link. To our knowledge, this is the first time modes have been used in this way.

Representation of kinds in a musical system. Let $\mathbb{M} = \{\ell^1, \dots, \ell^M\}$ and $\mathbb{W} = \{w^1, \dots, w^M\}$ be a musical system and let $k \in \mathcal{K}$ be a kind of chord. We define the representation of k in this musical system as the vector $x_{\mathbb{M},\mathbb{W}}(k) \in \mathbb{R}^M$ whose j-th coordinate is the w_j-weighted l_1–norm of $k - \ell^j$, that is:

$$x_{\mathbb{M},\mathbb{W}}(k) = \begin{pmatrix} \|k - \ell^1\|_1 \\ \vdots \\ \|k - \ell^M\|_M \end{pmatrix} \quad \text{where} \quad \|k - \ell^j\|_j = \sum_{i=1}^{11} w_i^j |k_i - \ell_i^j|.$$

If the representation $x_{\mathbb{M},\mathbb{W}}(\cdot)$ is injective (which is the case for the Pythagorean musical system but usually depends on the choice of $\mathbb{M}$ and $\mathbb{W}$), then the function $d : \mathcal{K} \times \mathcal{K} \to \mathbb{R}^+$ defined by

$$d_{\mathbb{M},\mathbb{W}}(k, k') = \|x_{\mathbb{M},\mathbb{W}}(k) - x_{\mathbb{M},\mathbb{W}}(k')\| = \left(\sum_{j=1}^{M} \left(\|k - \ell^j\|_j - \|k' - \ell^j\|_j \right)^2 \right)^{1/2}$$

is a distance. Otherwise, d is a pseudo-metric.

Example in the Pythagorean musical system. Figure 1 shows four examples of representations of very commonly used kinds. Without going to much into the music theory details, we know that the major kind is the tonic triad of the Ionian, Lydian and Mixolydian modes, the minor kind is the tonic triad of the Dorian, Phrygian, and Aeolian modes, and the diminished kind is the tonic triad of the Locrian mode. In these three cases, we can see that the representation does show smaller values for the corresponding modes. In particular, the mode of minimum cost always contains the represented kind. The dominant kind is a major triad with an additional note that is only in the Mixolydian mode, this diagram shows that the belonging of the dominant kind to the Mixolydian mode is more discriminant than it was with the major kind.

Cost on chords with the same roots. If two chords (r, k) and (r, k') share the same root, the cost between them is defined by

$$c_{\mathbb{M},\mathbb{W}}\big((r, k), (r, k')\big) = d_{\mathbb{M},\mathbb{W}}(k, k').$$

Figure 2 gives an example (in the Pythagorean musical system) of clustering using this cost, which motivates the fact that it shows substitutions of chord kinds that are very frequent in the jazz genre.

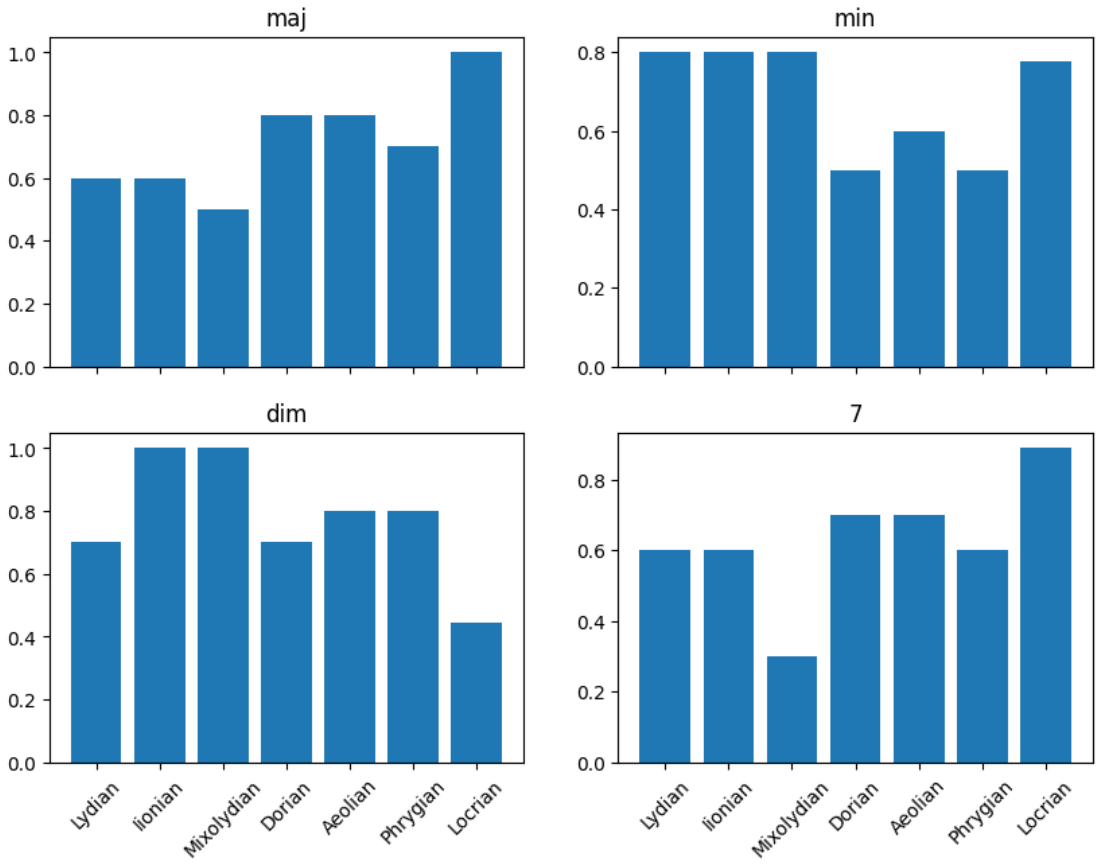

Figure 1: Representations of $x_{\mathbb{M},\mathbb{W}}(\cdot)$ for several kinds of chord: major, minor, diminished and dominant (i.e. 7) kinds.

Cost on chords with different roots. If the chords do not share the same roots, we consider that we should penalize the fact that playing one chord with the root of the other leads to a great difference between the original kind representation $x_{\mathbb{M},\mathbb{W}}(\cdot)$ and the new one. For example, the chords *Cmaj6* and *Amin7* have exactly the same notes $\{C, E, G, A\}$ but by choosing the roots C and A we obtain two different kinds, *maj6* and *min7*. The cost we propose will give a value of 0 for these two chords, as changing the root of *Cmaj6* to A produces exactly *Amin7*.

To define this cost, let (r^a, k^a) and (r^b, k^b) be two chords. Denote by $\mathcal{C}^a$ the unrooted chord associated with (r^a, k^a), that is, the notes that form this chord. We also consider the unrooted chord $\mathcal{C}^{ba} = \mathcal{C}^a \cup \{r^b\} \subset \mathbb{Z}_{12}$ and define (r^b, k^{ba}) as the chord $\mathcal{C}^{ba}$ with root r^b. The chord (r^a, k^{ab}) is defined in a similar way. These constructions allows us to define:

$$c_{\mathbb{M},\mathbb{W}}\big((r^a, k^a), (r^b, k^b)\big) = \Big[c_{\mathbb{M},\mathbb{W}}\big((r^a, k^a), (r^a, k^{ab})\big) + c_{\mathbb{M},\mathbb{W}}\big((r^b, k^b), (r^b, k^{ba})\big)\Big]/2.$$

Notice that $c_{\mathbb{M},\mathbb{W}}$ is a dissimilarity (which is not a distance in the general case) since for all

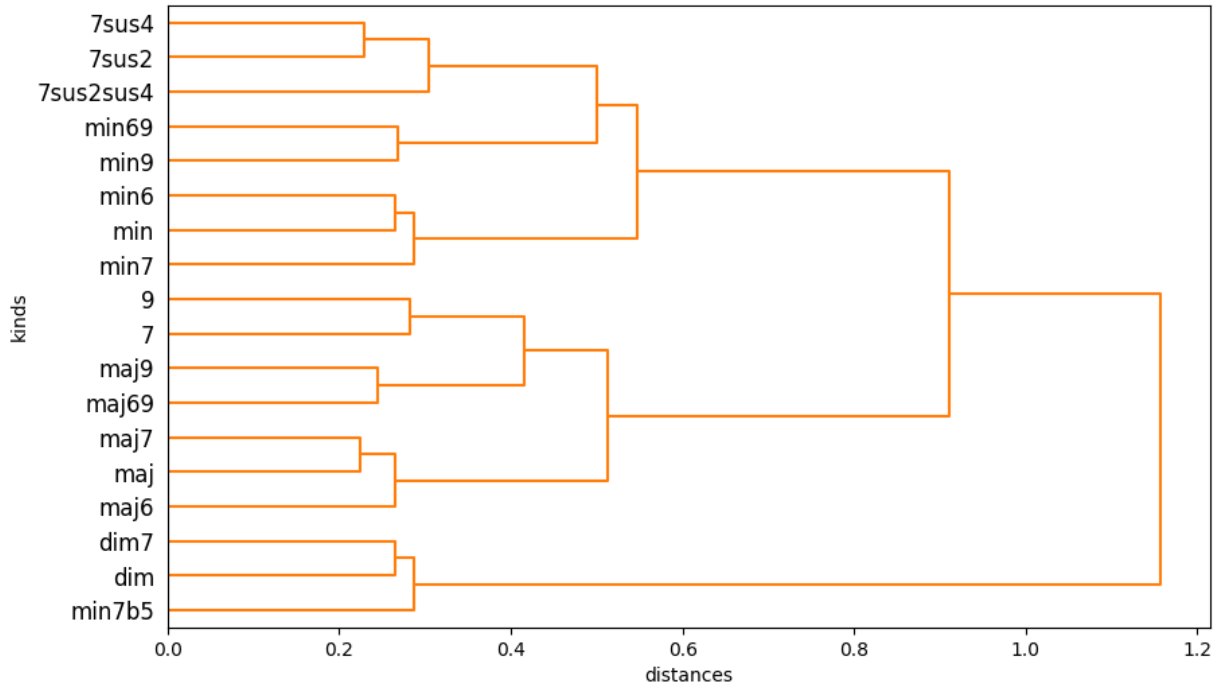

Figure 2: Hierarchical clustering performed on a small set of very often used kinds in the jazz repertoire

chords (r^a, k^a) and (r^b, k^b) in $\mathbb{Z}_{12} \times \mathcal{K}$ we have:

$$c_{\mathbb{M},\mathbb{W}}\big((r^a, k^a), (r^b, k^b)\big) = c_{\mathbb{M},\mathbb{W}}\big((r^b, k^b), (r^a, k^a)\big) \geq 0 \quad \text{and} \quad c_{\mathbb{M},\mathbb{W}}\left((r^a, k^a), (r^a, k^a)\right) = 0.$$

Figure 3 gives an example (in the Pythagorean musical system) of clustering that we produce using this dissimilarity on the chords that appear in the song *Cry Of The Wild Goose*[1]. This figure illustrates the fact that chords with a small dissimilarity are more likely to be substitutions of each other (adding G to *B♭major* produces the chord *Gmin7* and removing F from *Gmin7* produces *Gmin*).

4 Dissimilarity between songs

In our study, we assume that a song S can be represented by a sequence of chords, that is, $S = \{(r^1, k^1), \ldots, (r^{n_S}, k^{n_S})\}$. As the number of chords can vary from one song to another, it is essential to find a representation that is independent of this number to compare two songs. We thus associate a probability measure, defined on $\mathbb{Z}_{12} \times \mathcal{K}$, to each song S in the following

[1]The names of the chords are provided by the `commonName` function of the `music21` python package for computational Musicology (Cuthbert & Ariza, 2010).

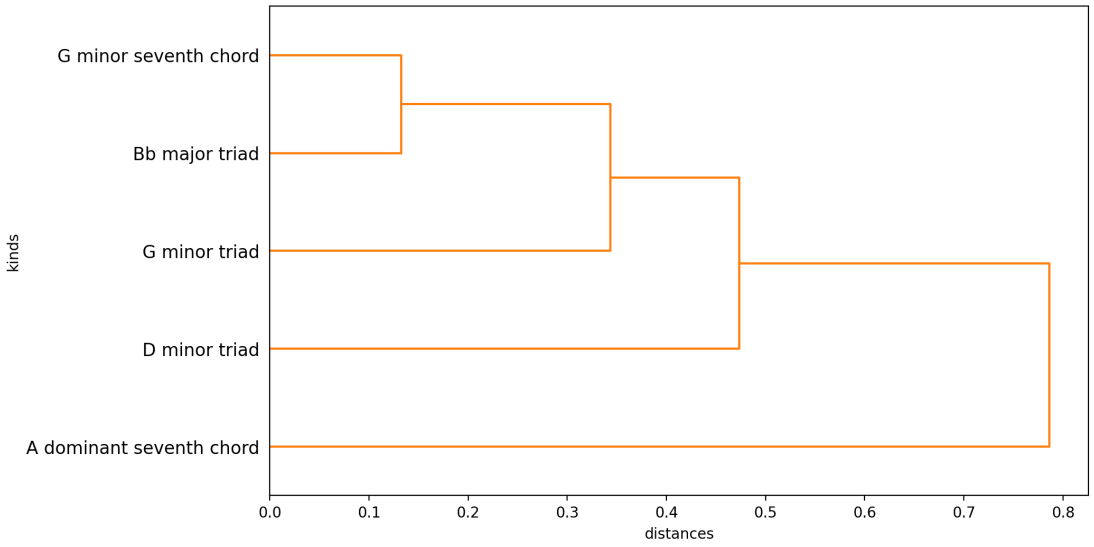

Figure 3: Hierarchical clustering performed on the chords of the song *Cry Of The Wild Goose*

way:

$$P_S(r,k) = \frac{1}{n_S}\sum_{i=1}^{n_S} 1_{[r^i=r,k^i=k]}, \quad \forall (r,k) \in \mathbb{Z}_{12} \times \mathcal{K}.$$

We then measure the proximity between two songs S_1 and S_2, as the optimal transport cost between P_{S_1} and P_{S_2}, that is:

$$\delta_{\mathbb{M},\mathbb{W}}(S_1,S_2) = \left(\inf_{\gamma\in\Gamma(P_{S_1},P_{S_2})} \int_{E^2} c^2_{\mathbb{M},\mathbb{W}}(x,y)\,\mathrm{d}\gamma(x,y)\right)^{1/2},$$

where $E = \mathbb{Z}_{12} \times \mathcal{K}$ and $\Gamma(P_{S_1}, P_{S_2})$ denotes the set of all joint probability measures γ on E^2 whose marginals are P_{S_1} and P_{S_2}. The Monge-Kantorovich problem (see Villani, 2009) mentioned above can be reformulated in a more explicit and equivalent manner as follows:

$$\delta^2_{\mathbb{M},\mathbb{W}}(S_1,S_2) = \min_{\gamma\in\Gamma(P_{S_1},P_{S_2})} \sum_{x\in E}\sum_{y\in E} c^2_{\mathbb{M},\mathbb{W}}(x,y)\gamma(x,y),$$

$$\text{subject to} \begin{cases} \sum_{y\in E} \gamma(x,y) = P_{S_1}(x), & \forall x \in E, \\ \sum_{x\in E} \gamma(x,y) = P_{S_2}(y), & \forall y \in E, \\ \gamma(x,y) \geq 0, & \forall (x,y) \in E^2. \end{cases}$$

Since $c_{\mathbb{M},\mathbb{W}}$ is a dissimilarity on chords, $\delta_{\mathbb{M},\mathbb{W}}$ is also a dissimilarity between songs. Thus, one can use specific data analysis tools, like hierarchical cluster analysis, which only require a dissimilarity matrix.

For a given song S, very few chords (r,k) satisfy $P_S(r,k) > 0$ compared to the 24.576 elements of E. This sparsity property motivates the use of a dissimilarity that is defined using Optimal

Transport to take into account the fact that some chords can easily be transported to another one with a small cost. Indeed, many chords can be considered as substitutes one to the other (e.g., *Cmaj6* and *Amin7* that have a cost of 0).

5 Experiments

To produce our experiments and to test our dissimilarity between songs, we exclusively used the chord progression dataset compiled and published by de Berardinis et al. (2023). We limited our study to a total of 2.846 songs from the dataset, specifically those from the Real Book corpus (see Mauch, Dixon, Harte, Casey, and Fields, 2007).

Figure 4 shows the empirical distribution of the chords in the song *Cry Of The Wild Goose* from the Real Book (we can see the sparsity property discussed in the last section, as only five chords are used). We also represented two other songs, *Into It* and *Got A Match*, for which we computed the dissimilarity with respect to *Cry Of The Wild Goose*.

These results are easily interpreted from a harmonic point of view. *Got A Match* and *Cry Of The Wild Goose* share chords in common (or very close to each other) which leads to a smaller dissimilarity than between *Into It* and *Cry Of The Wild Goose* whose chords are not consonant, such as *D minor* and *D♭ minor 7*. This can be seen as a drawback of the definition of $\delta_{\mathbb{M},\mathbb{W}}$ that does not take into account possible transpositions. This point, as well as taking into account the harmonic evolution of the songs, is left for future research.

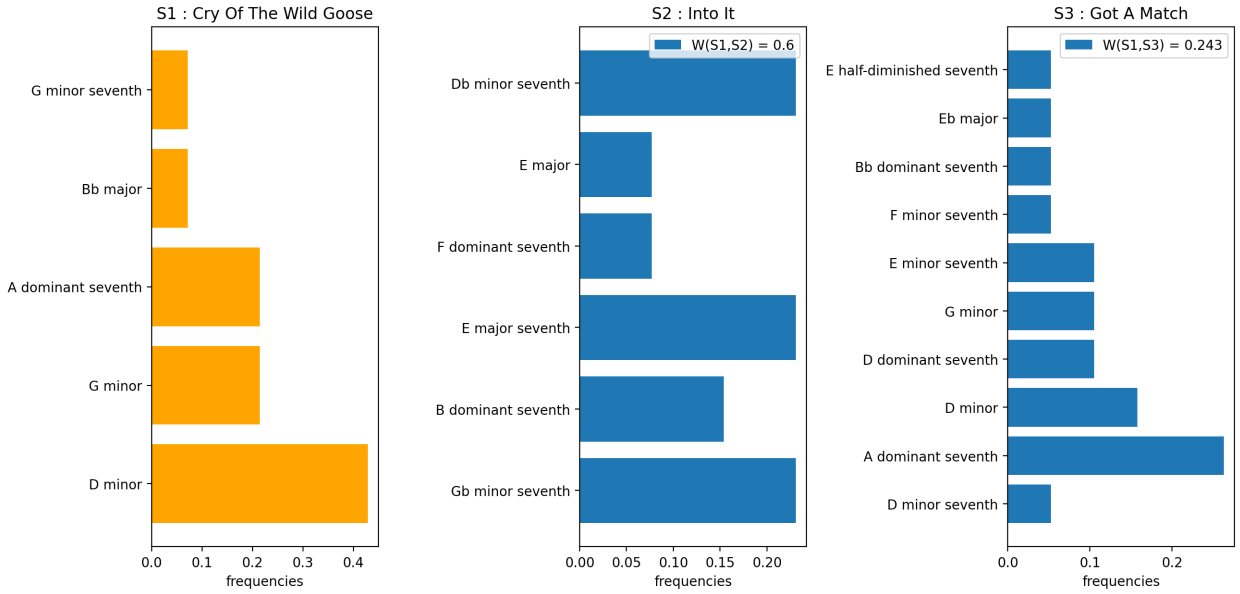

Figure 4: Histograms of the empirical distributions of chords in 3 songs of the Real Book

References

Aucouturier, J.-J. and F. Pachet (2002). Music similarity measures: What's the use? In *International Society for Music Information Retrieval Conference*.

de Berardinis, J., A. Meroño-Peñuela, A. Poltronieri, and V. Presutti (2023). Choco: a chord corpus and a data transformation workflow for musical harmony knowledge graphs. *Scientific Data 10*, 1–25.

Fujishima, T. (1999). Realtime chord recognition of musical sound: a system using common lisp music. In *International Conference on Mathematics and Computing*.

Givens, C. R. and R. M. Shortt (1984). A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal 31*, 231–240.

Harte, C., M. B. Sandler, S. A. Abdallah, and E. Gómez (2005). Symbolic representation of musical chords: A proposed syntax for text annotations. In *International Society for Music Information Retrieval Conference*.

Mauch, M., S. Dixon, C. Harte, M. A. Casey, and B. Fields (2007). Discovering chord idioms through beatles and real book songs. In *International Society for Music Information Retrieval Conference*.

Mauch, M., K. C. Noland, and S. Dixon (2009). Using musical structure to enhance automatic chord transcription. In *International Society for Music Information Retrieval Conference*.

Oudre, L., Y. Grenier, and C. Févotte (2011). Chord recognition by fitting rescaled chroma vectors to chord templates. *IEEE Transactions on Audio, Speech, and Language Processing 19*, 2222–2233.

Rocher, T., M. Robine, and P. Hanna (2010). A survey of chord distances with comparison for chord analysis. In *International Conference on Mathematics and Computing*.

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

Villani, C. (2009). *Optimal transport, old and new*, Volume 338. Springer Berlin, Heidelberg.