Pattern Recognition Letters 38 (2014) 46–53

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Elastic Net subspace clustering applied to pop/rock music structure analysis

Yannis Panagakis *, Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 18 January 2013
Available online 12 November 2013

Communicated by Y. Liu

Keywords:
Elastic Net
Subspace clustering
Sparse representation
Music structure analysis
Auditory representations

ABSTRACT

A novel homogeneity-based method for music structure analysis is proposed. The heart of the method is a similarity measure, derived from first principles, that is based on the matrix Elastic Net (EN) regularization and deals efficiently with highly correlated audio feature vectors. In particular, beat-synchronous mel-frequency cepstral coefficients, chroma features, and auditory temporal modulations model the audio signal. The EN induced similarity measure is employed to construct an affinity matrix, yielding a novel subspace clustering method referred to as Elastic Net subspace clustering (ENSC). The performance of the ENSC in structure analysis is assessed by conducting extensive experiments on the Beatles dataset. The experimental findings demonstrate the descriptive power of the EN-based affinity matrix over the affinity matrices employed in subspace clustering methods, attaining the state-of-the-art performance reported for the Beatles dataset.

1. Introduction

The musical form refers to the structural description of a music piece at the time scale of sections. That is, a music piece is described in terms of shorter, possibly repeated sections, which are often labeled according to their musical function in the piece. In Western pop/rock music and other related genres, common section labels are intro, verse, chorus, bridge, etc. (Paulus et al., 2010).

Automatic music structure analysis aims at describing a music piece in terms of sections by analyzing the audio signal. It employs low-level feature sequences extracted from the audio signal in order to model the timbral, melodic, and rhythmic content over time (Paulus et al., 2010). The underlying hypothesis is that, the structure is induced by the repetition of similar audio content (Dannenberg and Goto, 2008). Repetition implies that, there is some notion of similarity among the audio features, which can be exploited to segment the music into sections. That is, contiguous regions of similar music can be grouped together into segments and the resulting segments can be clustered together, defining the music sections. Technically, the segmentation of audio feature sequences into structural parts (i.e., the music sections) is achieved by employing methods detecting either homogeneity/novelty or repetition in a recurrence plot or a self-distance matrix (SDM) of audio features (Chen and Ming, 2011; Kaiser and Sikora, 2010; Levy and Sandler, 2008; Maddage, 2006; Paulus and Klapuri, 2009; Paulus et al., 2010; Weiss and Bello, 2010). Apart from a few exceptions e.g., Maddage (2006) and Paulus and Klapuri (2009), the majority of the aforementioned methods represent the music structure in terms of tag sequences, instead of assigning musically meaningful labels to the sections. For instance, the sequence of tags describing the structure of Oh! Darling by The Beatles is *ABCBCBD* as depicted in Fig. 1. Such a representation of the music structure is sufficient for music information retrieval applications (Dannenberg and Goto, 2008). For a comprehensive review on automatic music structure analysis, the interested reader is referred to Dannenberg and Goto (2008) and Paulus et al. (2010) (and the references therein).

Here, we focus on the structure analysis of pop/rock music. In these genres, a music section is often characterized by some sort of inherent homogeneity. That is, the instrumentation, tempo, or harmonic content is similar within the section (Paulus et al., 2010). Since the content of a music signal is modeled by appropriate audio feature vectors, a conventional way to reveal the desired within-section similarities is to construct an SDM containing the pairwise distances between all feature vectors and then to cluster the similar feature vectors into the same music section (Dannenberg and Goto, 2008; Paulus et al., 2010). However, similarity measures, such as the Euclidean distance, the inner product, the cosine distance, and the normalized correlation, which are often used to construct the SDM for music structure analysis, ignore the subspace structure of the music sections (Cheng et al., 2012). Such subspace structures are known to be valuable for feature vector similarity

* Corresponding author. Address: Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki GR-54124, Greece. Tel.: +30 697 402 1752; fax: +30 231 099 8453.

E-mail addresses: yannisp@csd.auth.gr, panagakis@aiia.csd.auth.gr (Y. Panagakis), costas@aiia.csd.auth.gr (C. Kotropoulos).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.021

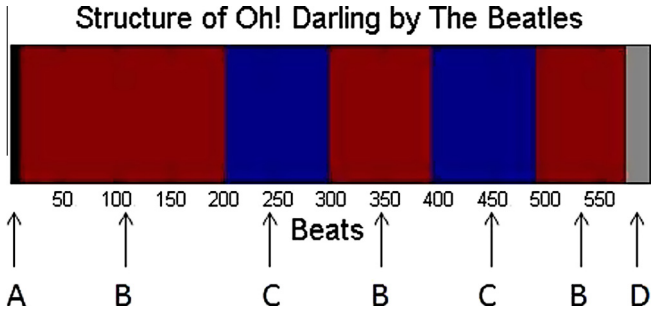

Fig. 1. Structural description of Oh! Darling by The Beatles. The song contains 7 segments from 4 different section-types namely, A,B,C, and D or intro (black segment), verse (red segment), bridge (blue segment), and outro (gray segment) in musical terms. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

measures in many clustering and classification problems (Cheng et al., 2012; Vidal, 2011; Liu et al., 2013). Moreover, the aforementioned similarity measures are extremely fragile in the presence of outliers (Vidal, 2011), hindering a reliable segmentation.

To exploit the hidden subspace structure and to increase robustness, reconstruction-based (as opposed to distance-based) similarity measures, such as the *sparse* (SR) (Vidal, 2011), the *low-rank* (LRR) (Liu et al., 2013), and the *ridge regression representation* (RR) (Panagakis and Kotropoulos, 2012b) of audio features are employed. The aforementioned representations measure the similarities among the feature vectors by decomposing each feature vector as s a linear combination of all other feature vectors seeking a sparse representation, a low-rank representation, or a representation minimizing the least squares error. That is, they minimize a proper norm of the representation matrix $\mathbf{Z}$, requiring $\mathbf{X} = \mathbf{XZ}$, where $\mathbf{X}$ is the data matrix, by solving a convex optimization problem indicated on the top of Fig. 2. If the data live in unions of independent subspaces (Vidal, 2011; Liu et al., 2013) any of the aforementioned three representations reveals the hidden subspace structure, since it exhibits nonzero within-subspace affinities and zero between-subspace affinities as illustrated in Fig. 2(a)–(e).

However, due to the homogeneity within the music sections, it is expected groups of contiguous audio feature vectors to be *highly correlated*. In this case, the SR, the LRR, and the RR can not reveal accurately the hidden subspace structure of audio feature vectors, hindering their reliable segmentation into music sections. Indeed, the SR does not discriminate between correlated feature vectors adequately (Tan et al., 2011). The low-rank constraint in the LRR does not take into account explicitly the relationships between contiguous audio feature vectors, since the nuclear norm applies sparsity constraints on the spectrum (i.e., the singular values) of the representation matrix and the RR does not perform feature vector selection by shrinking together the coefficients of the correlated feature vectors. The degraded performance of the aforementioned representations in handling highly correlated feature vectors is demonstrated in Fig. 2(g)–(j).

In this paper, to alleviate the inability of the SR, the LRR, and the RR-based similarity measures to cope with correlated feature vector sequences, as those emerging in music structure analysis, a novel reconstruction-based similarity measure, namely the *matrix Elastic Net* induced similarity measure of audio features is proposed. The contributions of the paper are:

- The matrix Elastic Net induced similarity measure is derived from first principles by extending the Elastic Net (EN) (i.e., the sum of ℓ_1-norm and squared ℓ_2-norm) regularized regression in compressive sensing (Zou and Hastie, 2005) to the more general setting of matrix subspace recovery (Liu et al., 2013). The main motivation behind this, is that the EN is not only able to cope with data drawn from independent subspaces shown in 2(a), but can also handle efficiently highly correlated feature vector sequences as analyzed in Tan et al. (2011) and depicted

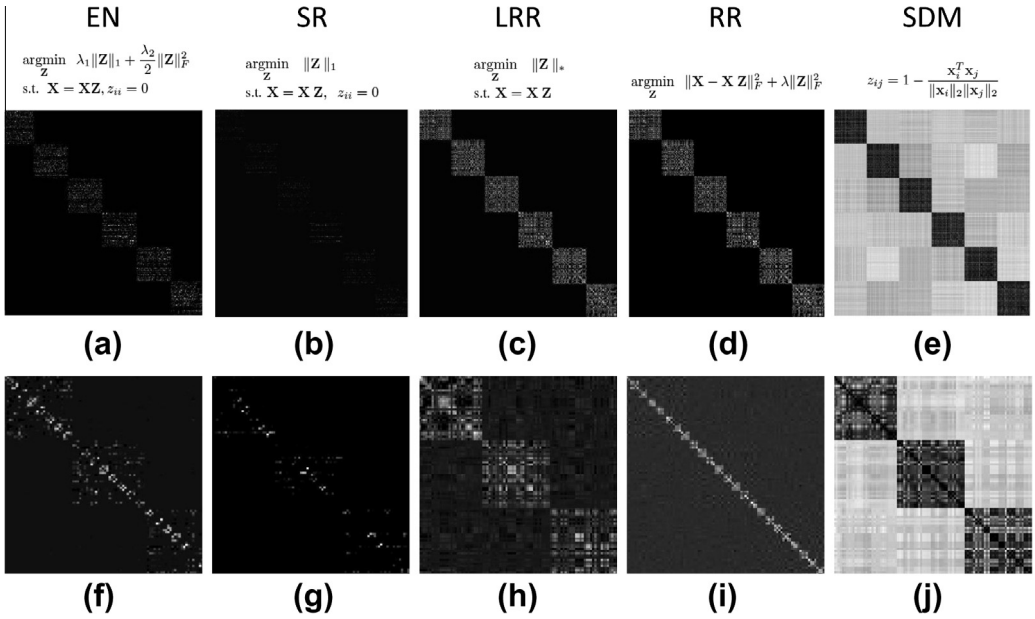

Fig. 2. For illustrative purposes, 6 linear pairwise independent subspaces are constructed whose basis $\{\mathbf{U}_i\}_{i=1}^6$ are computed by $\mathbf{U}_{i+1} = \mathbf{R}_i\mathbf{U}_i, i = 1, 2, \ldots, 5$. $\mathbf{U}_1 \in \mathbb{R}^{100\times 10}$ is a column orthonormal random matrix and $\mathbf{R}_i \in \mathbb{R}^{100\times 100}$ is a random rotation matrix. Consequently, the data matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_6] \in \mathbb{R}^{100\times 600}$ is drawn from a union 6 independent subspaces, where $\mathbf{X}_i = \mathbf{U}_i\mathbf{M}_i \in \mathbb{R}^{100\times 100}, i = 1, 2, \ldots, 6$. $\mathbf{M}_i \in \mathbb{R}^{10\times 100}, i = 1, 2, \ldots, 6$, is a random mixing matrix. Clearly the representation matrix $\mathbf{Z}$ is block-diagonal (a–d) if the the EN, the SR, the LRR, or the RR is applied onto $\mathbf{X}$. This does not hold for the SDM in (e) where non-zero between subspace affinities are observed. Next, to simulate the case of highly correlated feature vectors, the data matrix $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \hat{\mathbf{X}}_3] \in \mathbb{R}^{100\times 192}$ is constructed as follows: $\hat{\mathbf{X}}_s = [\bar{\mathbf{X}}_s^1, \bar{\mathbf{X}}_s^2, \ldots \bar{\mathbf{X}}_s^8] \in \mathbb{R}^{100\times 64}, s = 1, 2, 3$, where $\bar{\mathbf{X}}_1^k = [\mathbf{x}_{1k} + \alpha_1\mathbf{x}_{2k}, \mathbf{x}_{1k} + \alpha_2\mathbf{x}_{2k}, \ldots, \mathbf{x}_{1k} + \alpha_8\mathbf{x}_{2k}] \in \mathbb{R}^{100\times 8}$, $\bar{\mathbf{X}}_2^k = [\mathbf{x}_{3k} + \alpha_1\mathbf{x}_{4k}, \mathbf{x}_{3k} + \alpha_2\mathbf{x}_{4k}, \ldots, \mathbf{x}_{3k} + \alpha_8\mathbf{x}_{4k}] \in \mathbb{R}^{100\times 8}$ and $\bar{\mathbf{X}}_3^k = [\mathbf{x}_{5k} + \alpha_1\mathbf{x}_{6k}, \mathbf{x}_{5k} + \alpha_2\mathbf{x}_{6k}, \ldots, \mathbf{x}_{5k} + \alpha_8\mathbf{x}_{6k}] \in \mathbb{R}^{100\times 8}$, a_i are random weights, and $\mathbf{x}_{ij}$ denotes the jth column of $\mathbf{X}_i$. In other words, $\hat{\mathbf{X}}_s$ is drawn from a union of 2 subspaces containing in its columns highly correlated vectors and thus the columns of $\hat{\mathbf{X}}$ live in 3 unions of subspaces. It is clear form (f–j) that only the EN, is able to reveal the hidden subspace structure of $\hat{\mathbf{X}}_s$.