Elastic Net Subspace Clustering Applied to Pop/Rock Music Structure Analysis

Yannis Panagakis[1] and Constantine Kotropoulos

Department of Informatics
Aristotle University of Thessaloniki
email: {panagakis,costas}@aiia.csd.auth.gr

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]Corresponding author: Yannis Panagakis, Dept. Informatics, Aristotle University of Thessaloniki, Box 451 Thessaloniki, GR-54124, GREECE,
Tel. +30-697-40-21-752, Fax. +30-231-099-8453
emails: yannisp@csd.auth.gr; panagakis@aiia.csd.auth.gr

Preprint submitted to Pattern Recognition Letters *October 3, 2013*

Keywords: Elastic Net, Subspace Clustering, Sparse Representation, Music Structure Analysis, Auditory Representations

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; 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; Paulus et al., 2010) (and the references therein).

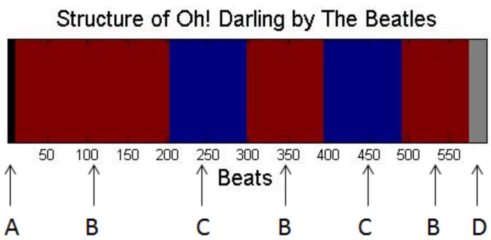

Figure 1: Structural description of Oh! Darling by The Beatles. The song contains 7 segments from 4 different section-types namely, A,B,C, and D or intro (black segment), verse (red segment), bridge (blue segment), and outro (gray segment) in musical terms.

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 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 $\mathbf{X} = \mathbf{X}\,\mathbf{Z}$, 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

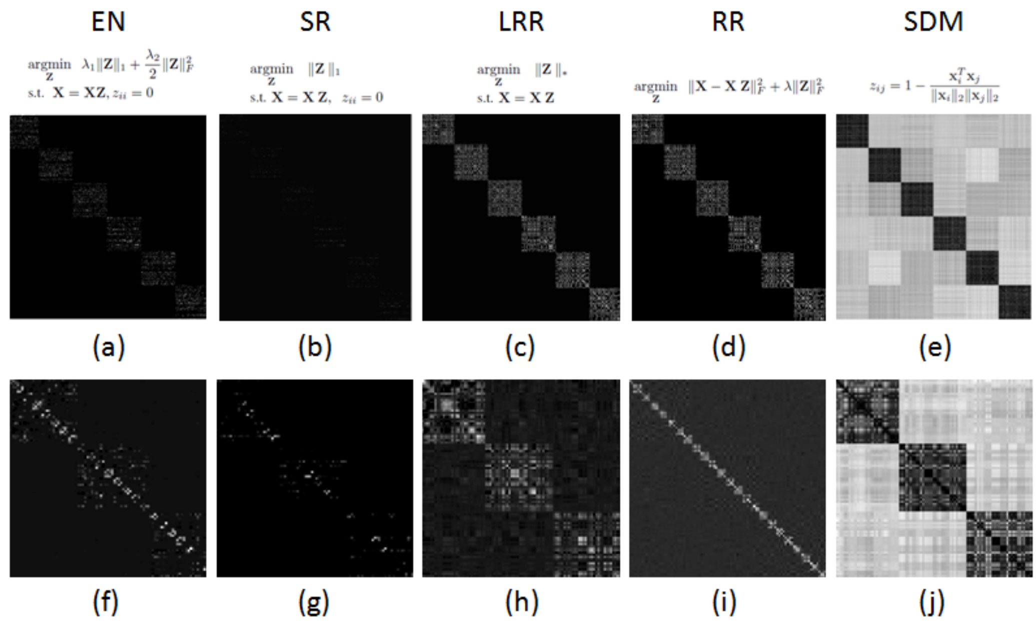

Figure 2: For illustrative purposes, 6 linear pairwise independent subspaces are constructed whose basis $\{\mathbf{U}_i\}_{i=1}^{6}$ are computed by $\mathbf{U}_{i+1} = \mathbf{R}_i\mathbf{U}_i$, $i = 1, 2, \dots, 5$. $\mathbf{U}_1 \in \mathbb{R}^{100\times 10}$ is a column orthonormal random matrix and $\mathbf{R}_i \in \mathbb{R}^{100\times 100}$ is a random rotation matrix. Consequently, the data matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_6] \in \mathbb{R}^{100\times 600}$ is drawn from a union 6 independent subspaces, where $\mathbf{X}_i = \mathbf{U}_i\mathbf{M}_i \in \mathbb{R}^{100\times 100}$, $i = 1, 2, \dots, 6$. $\mathbf{M}_i \in \mathbb{R}^{10\times 100}$, $i = 1, 2, \dots, 6$, 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: $\hat{\mathbf{X}}_s = [\bar{\mathbf{X}}_s^1, \bar{\mathbf{X}}_s^2, \dots \bar{\mathbf{X}}_s^8] \in \mathbb{R}^{100\times 64}$, $s = 1, 2, 3$, where $\bar{\mathbf{X}}_1^k = [\mathbf{x}_{1k} + \alpha_1\mathbf{x}_{2k}, \mathbf{x}_{1k} + \alpha_2\mathbf{x}_{2k}, \dots, \mathbf{x}_{1k} + \alpha_8\mathbf{x}_{2k}] \in \mathbb{R}^{100\times 8}$, $\bar{\mathbf{X}}_2^k = [\mathbf{x}_{3k} + \alpha_1\mathbf{x}_{4k}, \mathbf{x}_{3k} + \alpha_2\mathbf{x}_{4k}, \dots, \mathbf{x}_{3k} + \alpha_8\mathbf{x}_{4k}] \in \mathbb{R}^{100\times 8}$ and $\bar{\mathbf{X}}_3^k = [\mathbf{x}_{5k} + \alpha_1\mathbf{x}_{6k}, \mathbf{x}_{5k} + \alpha_2\mathbf{x}_{6k}, \dots, \mathbf{x}_{5k} + \alpha_8\mathbf{x}_{6k}] \in \mathbb{R}^{100\times 8}$, 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 $\hat{\mathbf{X}}_s$.

in Fig. 2 (f). In that sense, the EN-based similarity measure of feature vector sequences (represented as matrix columns) is obtained by solving a convex optimization problem, which involves the minimization of the *matrix EN regularizer* (i.e., the sum of matrix ℓ_1-norm and squared

Frobenius-norm).

- The matrix EN is obtained by a novel algorithm, whose convergence is guaranteed and suits well for large scale optimization problems, since it is based on Linearized Alternating Directions Method (Lin et al., 2011).

- Based on the matrix EN induced similarity measure, music structure analysis can be performed by applying the normalized cuts algorithm (NCuts) (Shi and Malik, 2000) to the EN-based affinity matrix of audio feature vector sequences. The above procedure is referred to as elastic subspace clustering (ENSC). By conducting extensive experiments on the manually annotated Beatles benchmark dataset (cf. Section 4.1), the descriptive power of the EN-based similarity measure is demonstrated over common reconstruction- and distance-based similarity measures with respect to several evaluation criteria. The best results reported here match those obtained by the state-of-the-art music structure analysis methods (Kaiser and Sikora, 2010; Levy and Sandler, 2008; Paulus and Klapuri, 2009), which have also been evaluated in the same dataset following the same experimental protocol.

2. Audio feature extraction

The variations between different music segment-types are captured by extracting three audio features from each recording. In particular, the mel-frequency cepstral coefficients (MFCCs), the chroma features (Ryynanen and Klapuri, 2008), and the auditory temporal modulations (ATMs) (Panagakis et al., 2010) are employed in order to form sequences of *beat-synchronous* feature vectors using the beat tracking algorithm described in (Ellis, 2007). That is, the feature vectors between two consecutive beats are averaged to yield a single feature vector per beat. Beat-synchronous feature vectors undergo a normalization in order to have zero mean and unit ℓ_2-norm.

The MFCCs encode the timbral properties of the music signal. They are calculated by employing frames of duration 92.9 ms with a hop size of 46.45 ms and a 42-band filter bank as in (Paulus and Klapuri, 2009). The zeroth order coefficient is discarded yielding a sequence of 12-dimensional MFCC vectors.

The chroma features characterize the harmonic content of the music signal by projecting the entire spectrum onto 12 bins representing the 12 distinct semitones (or chroma) of a musical octave. Frames of 92.9 ms with a hop size of 23.22 ms were employed for their calculation, resulting into a sequence of 12-dimensional chroma vectors.

The ATMs are obtained by modeling the path of human auditory processing as a two-stage process. In the first stage, which models the early auditory system, the auditory spectrogram is obtained. The early auditory system is modeled by Lyons' passive ear model (Lyon, 1982) employing 96 frequency channels ranging from 62 Hz to 11 kHz. The auditory spectrogram is then downsampled along the time axis in order to obtain 10 feature vectors between two successive beats. The underlying temporal modulations of the music signal are derived by applying a biorthogonal wavelet filter along each temporal row of the auditory spectrogram, having previously subtracted its mean, for 8 discrete rates $\{2, 4, 8, 16, 32, 64, 128, 256\}$ Hz ranging from slow to fast temporal rates. By doing so, the entire auditory spectrogram is modeled by a three-dimensional representation of frequency, rate, and time, which is then unfolded along the time-mode in order to obtain a sequence of $96 \times 8 = 728$-dimensional ATM features.

3. Elastic Net subspace clustering for music structural segmentation

As argued in Section 1, a critical issue in music structure analysis is to robustly measure the similarity between the feature vectors, revealing the hidden subspaces. That is, the feature vectors of a music section need to be similar with respect to a subset of attributes (captured by subspaces) only, a property ignored whenever the Euclidean or other related distance measure is employed (Cheng et al., 2012). To accomplish this, a novel reconstruction-based similarity measure, which is based on the matrix EN regularization, is proposed to exploit properly the correlations between the beat-synchronous feature vectors within time windows having duration of a few beats.

3.1. Elastic Net induced similarity measure for clean data

Let a given music recording of K section-types (i.e., intro, verse, chorus, bridge, etc.) be represented by a sequence of N beat-synchronous audio feature vectors of size d, i.e., $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\dots|\mathbf{x}_N] \in \mathbb{R}^{d\times N}$. Two reasonable assumptions for $\mathbf{X}$ are as follows: 1) If the feature vectors belong to a music section, they will lie into the same union of subspaces. That is, the columns of $\mathbf{X}$ are drawn from a union of K unions of independent linear subspaces of unknown dimensions. 2) Groups of a few contiguous dictionary atoms (i.e., columns of $\mathbf{X}$) are quite similar and thus are expected to be highly correlated.

Based on the aforementioned assumptions, one would like to learn the representation matrix $\mathbf{Z} \in \mathbb{R}^{N\times N}$, such that $\mathbf{X} = \mathbf{XZ}$, with $z_{ij} = 0$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ lie on different unions of subspaces and nonzero z_{ij} otherwise. Such a representation matrix $\mathbf{Z}$ measures the similarity between all the features by unveiling the hidden subspace structure and it is obtained by solving:

$$\underset{\mathbf{Z}}{\operatorname{argmin}} \ \ \lambda_1\|\mathbf{Z}\|_1 + \frac{\lambda_2}{2}\|\mathbf{Z}\|_F^2 \ \text{ s.t. } \ \mathbf{X} = \mathbf{XZ}, z_{ii} = 0. \tag{1}$$

In (1), the matrix ℓ_1-norm is defined as $\|\mathbf{Z}\|_1 = \sum_i \sum_j |z_{ij}|$ and $\|\mathbf{Z}\|_F = \sqrt{\sum_i \sum_j z_{ij}^2}$ denotes the Frobenius norm. It is observed that (1) is a combination of the matrix ℓ_1-norm and squared Frobenius norm. Accordingly, it is actually an extension of the vector EN regularizer (Zou and Hastie, 2005) to matrices. The solution of (1), which is referred to as EN representation matrix, admits nonzero entries for within-subspace affinities and zero entries for between-subspace affinities. This fact is proved in Theorem 1, which is a consequence of Lemma 1 (Bhatia and Kittaneh, 1990).

Lemma 1. Let the parametric norm $\|.\|_\lambda = \lambda_1\|.\|_1 + \frac{\lambda_2}{2}\|.\|_F^2$, with $\lambda_1, \lambda_2 > 0$. For any four matrices $\mathbf{B}, \mathbf{C}, \mathbf{D}$, and $\mathbf{F}$ of compatible dimensions,

$$\left\| \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{F} \end{bmatrix} \right\|_\lambda \geq \left\| \begin{bmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{F} \end{bmatrix} \right\|_\lambda = \|\mathbf{B}\|_\lambda + \|\mathbf{F}\|_\lambda. \tag{2}$$

Theorem 1. Assume the columns of $\mathbf{X}$ are drawn from a union of K linear independent subspaces of unknown dimensions. Without loss of generality, $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2|\ldots|\mathbf{X}_K] \in \mathbb{R}^{d\times N}$, where the columns of $\mathbf{X}_k \in \mathbb{R}^{d\times N_k}$, $k = 1, 2, \ldots, K$ correspond to the N_k feature vectors originating from the

kth subspace. The minimizer of (1) is block-diagonal.

The proof of Theorem 1 follows similar lines to that included in (Panagakis and Kotropoulos, 2012a).

3.2. Elastic Net induced similarity measure for noisy data

In practice, the assumption $\mathbf{X} = \mathbf{X}\mathbf{Z}$ does not hold exactly, because the data are *approximately* drawn from unions of subspaces. This fact introduces certain deviations from the ideal modeling assumptions. The latter can be treated collectively as additive *noise* contaminating the ideal model i.e., $\mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}$. To account for the noise, a distortion term is inserted into (1) and a robust solution is sought for the following convex optimization problem:

$$\operatorname*{argmin}_{\mathbf{Z},\mathbf{E}} \quad \lambda_1\|\mathbf{Z}\|_1 + \frac{\lambda_2}{2}\|\mathbf{Z}\|_F^2 + \lambda_3\|\mathbf{E}\|_1 \quad \text{s.t. } \mathbf{X} = \mathbf{X}\,\mathbf{Z} + \mathbf{E}, z_{ii} = 0, \tag{3}$$

where $\lambda_3 > 0$ is a regularization parameter.

To efficiently solve (3), the Linearized Alternating Directions Method (LADM) (Lin et al., 2011) is employed, which is suitable for large scale optimization problems. By applying the LADM, one seeks to minimize the (partial) augmented Lagrangian function:

$$\begin{aligned}\operatorname*{argmin}_{\mathbf{Z},\mathbf{E}} \ & \mathcal{L}(\mathbf{Z},\mathbf{E},\boldsymbol{\Xi}) = \lambda_1\|\mathbf{Z}\|_1 + \frac{\lambda_2}{2}\|\mathbf{Z}\|_F^2 + \lambda_3\|\mathbf{E}\|_1 \\ & + \operatorname{tr}\left(\boldsymbol{\Xi}^T(\mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E})\right) + \frac{\mu}{2}\|\mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E}\|_F^2, \text{ s.t. } z_{ii} = 0,\end{aligned} \tag{4}$$

where $\boldsymbol{\Xi}$ gathers the Lagrange multipliers for the equality constraints in (3) and $\mu > 0$ is a penalty parameter. Let t denotes the iteration index and σ

be the largest singular value of $\mathbf{X}$. Then, (4) is minimized with respect to each variable in an alternating fashion as outlined in Algorithm 1.

Following (Lin et al., 2011), since (5) does not admit a closed-form solution, the smooth term in (4) is linearly approximated and a simple closed-form solution is obtained. Its derivation can be found in the Appendix. The approximate solution of (5) employs the shrinkage operator $\mathcal{S}_\tau[q] = \text{sgn}(q)\text{max}(|q| - \tau, 0)$ (Candes et al., 2011), which can be extended to matrices by applying it element-wise. Similarly, a closed-form solution in (8) is obtained by applying the shrinkage operator (9). The diagonal elements of $\mathbf{Z}_{[t+1]}$ are set to zero in (7) in order to fulfil the constraint in (4).

To set the internal parameters of the Algorithm 1, i.e., $\theta = \eta\sigma^2$ and ρ which are independent from the data $\mathbf{X}$, 10 data matrices have been constructed, as in Fig 2. By fixing the data dependent parameters $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$, the parameters ρ and θ set to those values, which yield the fastest drop of the mean approximation error (obtained by executing Algorithm 1 10 times) as depicted in Fig. 3. By inspecting Fig. 3, ρ was set to 1.9 and $\eta = 1.02$. Regarding the parameters related to the stoping conditions of Algorithm, $\epsilon_1 = 10^{-4}$ and $\epsilon_2 = 10^{-5}$ are typical choices e.g., (Lin et al., 2011).

The *convergence* of Algorithm 1 is guaranteed, since only two variables (i.e., $\mathbf{Z}, \mathbf{E}$) are involved in the optimization problem (Bertsekas, 1996; Lin et al., 2011). Moreover, since Algorithm 1 is an alternating directions method, its *converge rate* is $\mathcal{O}(1/t)$ (He and Yuan, 2012).

Algorithm 1 Solving (4) by the LADM method.

Input: Data matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$ and the parameters λ_1, λ_2, λ_3.
Output: Matrix $\mathbf{Z} \in \mathbb{R}^{N\times N}$ and matrix $\mathbf{E} \in \mathbb{R}^{d\times N}$.

1: Initialize: $\mathbf{Z}_{[0]} = \mathbf{0}, \mathbf{E}_{[0]} = \mathbf{0}$, $\boldsymbol{\Xi}_{[0]} = \mathbf{0}$, $\mu_{[0]} = 10^{-6}$, $\rho = 1.9$, $\theta = 1.02\sigma^2$ $\epsilon_1 = 10^{-4}$, and $\epsilon_2 = 10^{-5}$.
2: **while** not converged **do**
3: Fix $\mathbf{E}_{[t]}$, and update $\mathbf{Z}_{[t+1]}$ by

$$\mathbf{Z}_{[t+1]} = \operatorname*{argmin}_{\mathbf{Z}_{[t]}} \mathcal{L}(\mathbf{Z}_{[t]}, \mathbf{E}_{[t]}, \boldsymbol{\Xi}_{[t]}) \tag{5}$$

$$\approx \mathcal{S}_{\frac{\lambda_1}{\theta\mu_{[t]}}}\left[\mathbf{Z}_{[t]} + \frac{1}{\theta}\left(\mathbf{X}^T(\mathbf{X} - \mathbf{X}\mathbf{Z}_{[t]} - \mathbf{E}_{[t]} + \frac{1}{\mu_{[t]}}\boldsymbol{\Xi}_{[t]}) - \frac{\lambda_2}{\mu_{[t]}}\mathbf{Z}_{[t]}\right)\right] \tag{6}$$

$$z_{ii[t+1]} = 0. \tag{7}$$

4: Fix $\mathbf{Z}_{[t+1]}$ and update $\mathbf{E}_{[t]}$ by

$$\mathbf{E}_{[t+1]} = \operatorname*{argmin}_{\mathbf{E}_{[t]}} \mathcal{L}(\mathbf{Z}_{[t+1]}, \mathbf{E}_{[t]}, \boldsymbol{\Xi}_{[t]}) \tag{8}$$

$$= \mathcal{S}_{\frac{\lambda_3}{\mu_{[t]}}}\left[\mathbf{X} - \mathbf{X}\mathbf{Z}_{[t+1]} + \frac{1}{\mu_{[t]}}\boldsymbol{\Xi}_{[t]}\right] \tag{9}$$

5: Update the Lagrange multiplier by
$\boldsymbol{\Xi}_{[t+1]} = \boldsymbol{\Xi}_{[t]} + \mu_{[t]}(\mathbf{X} - \mathbf{X}\mathbf{Z}_{[t+1]} - \mathbf{E}_{[t+1]})$.
6: Update $\mu_{[t+1]}$ by $\mu_{[t+1]} \leftarrow \min(\rho \cdot \mu_{[t]}, 10^{10})$.
7: Check convergence conditions
$\|\mathbf{X} - \mathbf{X}\mathbf{Z}_{[t]} - \mathbf{E}_{[t]}\|_F / \|\mathbf{X}\|_F \leq \epsilon_1$
and $\max\left(\|\mathbf{E}_{[t]} - \mathbf{E}_{[t-1]}\|_F/\|\mathbf{X}\|_F, \|\mathbf{Z}_{[t]} - \mathbf{Z}_{[t-1]}\|_F/\|\mathbf{X}\|_F\right) \leq \epsilon_2$.
8: $t \leftarrow t + 1$.
9: **end while**

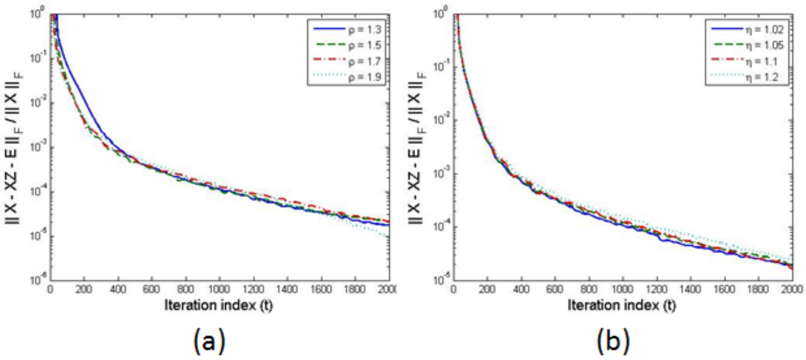

Figure 3: (a) Mean approximation error as a function of the iteration index (t) for fixed $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$, $\eta = 1.02$, and $\rho \in \{1.3, 1.5, 1.7, 1.9\}$. (b) Mean approximation error as a function of the iteration index (t) for fixed $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$, $\rho = 1.9$, and $\eta \in \{1.02, 1.05, 1.1, 1.2\}$.

3.3. Segmentation based on the Elastic Net induced similarity measure

Having found $\mathbf{Z}$ by applying the LADM, the column space of the EN representation matrix $\mathbf{Z}$ is useful for subspace segmentation. Let $\mathbf{Z} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be the singular value decomposition of $\mathbf{Z}$ and $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{U}^T = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T$. Then, an EN-based nonnegative symmetric affinity matrix $\mathbf{W} \in \mathbb{R}_+^{N \times N}$ has elements (Liu et al., 2013):

$$w_{ij} = m_{ij}^2. \tag{10}$$

The EN-based affinity matrix, is further post-processed by applying a 2D Gabor filter with angle $\pi/4$ in order to enhance any diagonal structures in it. The segmentation of the columns of $\mathbf{X}$ into K section-types is performed by applying the NCuts (Shi and Malik, 2000) to the post-processed EN-based affinity matrix.

3.4. Estimation of the number of section-types

A challenging problem in music structure analysis is the automatic estimation of the number of different section-types in the music piece. If the affinity matrix $\mathbf{W}$ has exactly nonzero within-subspace affinities and zero between-subspace affinities then the number of section-types $\bar{K}$ (number of subspaces in general) could be found be counting the zero singular values of the Laplacian matrix derived by $\mathbf{W}$. However in practice, the affinity matrix $\mathbf{W}$ has almost zero between-subspace affinities and thus one could estimate the number of section-types $\bar{K}$ by counting the number of singular values which are smaller than a threshold. That is, the number of section-types $\bar{K}$ is estimated by employing a soft-thresholding approach (Liu et al., 2013):

$$\bar{K} = N - \mathrm{int}(\sum_{i=1}^{N} f_\tau(\sigma_i)), \quad \tau \in (0, 1), \tag{11}$$

where $\mathrm{int}(\cdot)$ returns the nearest integer of a real number, $\{\sigma_i\}_{i=1}^{N}$ denotes the set of the singular values of the Laplacian matrix derived by the corresponding affinity matrix, and $f_\tau(\cdot)$ is the soft-thresholding operator defined as $f_\tau(\sigma) = 1$ if $\sigma \geq \tau$ and $\log_2(1 + \frac{\sigma^2}{\tau^2})$, otherwise.

4. Experimental evaluation

4.1. Dataset, evaluation procedure, and evaluation metrics

Beatles dataset[1]: The dataset consists of 180 songs by The Beatles. The songs were annotated by the musicologist Alan W. Pollack. Segmentation

[1] http://www.dtic.upf.edu/ perfe/annotations/sections/license.html

time stamps were inserted at Universitat Pompeu Fabra. Some minor corrections to annotations were made at Tampere University of Technology (TUT)[2]. Each music recording contains on average 10 sections from 5 unique section-types (Weiss and Bello, 2010).

The audio signal is modeled using three beat-synchronous feature vector sequences described in Section 2. Structure segmentation is obtained by determining the affinity matrices employed in the reconstruction-based subspace clustering methods. To this end, the proposed EN induced similarity measure is compared against the similarity measures induced by the sparse, low-rank, and ridge regression. The corresponding affinity matrices are constructed as follows: The EN-based affinity matrix is given by (10), the SR-based affinity matrix is obtained element-wise as $w_{ij} = 0.5(|z_{ij}| + |z_{ji}|)$ (Vidal, 2011). The LRR- and the RR-based affinity matrices are obtained by applying the procedures proposed in (Liu et al., 2013) and (Panagakis and Kotropoulos, 2012b), respectively, to derive $\mathbf{Z}$ and finally employing (10). Next, all affinity matrices are enhanced by Gabor filtering, and finally the NCuts algorithm is applied to all post-processed affinity matrices. The procedure described above leads to the ENSC, the sparse subspace clustering (SSC), the low-rank subspace clustering (LRRSC), and the ridge-regression subspace clustering (RRSC) applied to beat-synchronous feature vector sequences. For the conventional distance-based similarity measures, we replace the affinity matrices employed in subspace clustering by the SDM constructed using the cosine distance of the beat-synchronous feature vectors. Next, the NCuts is applied to the similarly post-processed SDM.

[2]http://www.cs.tut.fi/sgn/arg/paulus/structure.html

Furthermore, the combination of multiple features (i.e., cross-feature information) is investigated. To this end, cross-feature affinity matrices are obtained by linearly combining the affinity matrices computed for each different feature vector sequence and employing the aforementioned similarity measures.

Two sets of experiments were conducted on the Beatles dataset. First, in order to fairly compare the proposed method with the methods in (Kaiser and Sikora, 2010; Levy and Sandler, 2008; Paulus and Klapuri, 2009), the number of section-types (i.e., clusters) K was set equal to 5. In the second experiment, the number of section-types was estimated using (11). The optimal values for λ_i, $i = 1, 2, 3$ involved in the ENSC as well as in SSC, the LRR, and the RRSC were determined by a grid search over 10 randomly selected music recordings of the dataset. The same procedure was employed to determine the parameter τ in (11).

Three different metrics are used for music segmentation evaluation. That is, the pairwise F-measure (PF), the conditional entropy-based measure for over-segmentation (S_o), and under-segmentation (S_u) (Lukashevich, 2008). In the following, the discussion refers to beat synchronous feature vectors that are called beats for brevity. They compare pairs of beats, which are assigned to the same section-type by automatic analysis methods against the reference segmentation. Let $\mathbb{F}_A$ be the set of similarly labeled pairs of beats in a recording according to the music structure analysis method and $\mathbb{F}_H$ be the set of similarly labeled pairs in the human reference segmentation. PF is defined as $PF = 2 \cdot \frac{PP \cdot PR}{PP+PR}$, where the pairwise precision, PP, and the pairwise recall, PR, are defined as: $PP = \frac{|\mathbb{F}_A \cap \mathbb{F}_H|}{|\mathbb{F}_A|}$, $PR = \frac{|\mathbb{F}_A \cap \mathbb{F}_H|}{|\mathbb{F}_H|}$ with $| \cdot |$

denoting the set cardinality. S_o and S_u are defined as follows:

$$S_o = 1 - \frac{-\sum_{i=1}^{N_H}(\frac{n_i^H}{\sum_{i=1}^{N_H}\sum_{j=1}^{N_A} n_{ij}})\sum_{j=1}^{N_A}\frac{n_{ij}}{n_i^H}\log_2\frac{n_{ij}}{n_i^H}}{\log_2 N_A}, \tag{12}$$

$$S_u = 1 - \frac{-\sum_{j=1}^{N_A}(\frac{n_j^A}{\sum_{i=1}^{N_H}\sum_{j=1}^{N_A} n_{ij}})\sum_{i=1}^{N_H}\frac{n_{ij}}{n_j^A}\log_2\frac{n_{ij}}{n_j^A}}{\log_2 N_H}, \tag{13}$$

where N_A and N_H are the number of section-types in the estimated segmentation and human reference segmentation, respectively. n_{ij} denotes the number of beats that simultaneously belong to the ith section-type in the ground-truth segmentation and to the jth section-type in the estimated one. n_i^H is the total number of beats, that belong to the ith section-type in the ground-truth segmentation and n_j^A is the total number of beats belonging to the jth section-type in the automatic segmentation. The numerator in (13) corresponds to the conditional entropy measuring the amount of ground-truth segmentation information that is missing in the estimated segmentation. In analogy, the numerator in (12) measures the amount of the spurious information. The aforementioned three metrics admit values in $[0, 1]$. They reach their maximum value, when the segmentation is perfect and approach zero, when the segmentation tends to be random. The average number of the final segments (NoS) obtained by the various segmentation methods and the average running time (ART) in CPU seconds for each method, excluding the time for feature extraction, are also reported. Although the proposed method is a segmentation method and not a boundary detection one, a few boundary retrieval results are reported for comparison with the state-of-the-art methods. To this end, the segment boundary retrieval performance is evaluated with respect to the standard precision (P), recall (R), and F-measure (F) (Manning et al., 2008). Following (Levy and Sandler, 2008; Paulus and Klapuri, 2009), a boundary in the results is considered as correct, if it is within 3 sec from the boundary in the annotation.

4.2. Experimental results

The structure segmentation performance on the Beatles dataset for a fixed number of section-types (i.e., $K = 5$) is summarized in Table 1 for individual audio feature vector sequences and in Table 2 for the combination of multiple feature vectors. Any metric gain larger than approximately 0.08 is statistically significant at 95% level of significance.

Table 1: Structure segmentation performance on the Beatles dataset with fixed $K = 5$. The numbers within parentheses indicate figures of merit, if different, after excluding the 10 music recordings used for parameter selection.

Method	**Features (Parameters)**	PF			S_o			S_u			NoS	ART
		Mean	Best	Worst	Mean	Best	Worst	Mean	Best	Worst	Mean	
ENSC	MFCCs ($\lambda_1 = 0.1, \lambda_2 = 0.2, \lambda_3 = 0.1$)	0.56	0.88	0.32	**0.64**	0.80	0.50	0.51 (0.52)	0.85	0.23	18	28.4
	chroma ($\lambda_1 = 0.1, \lambda_2 = 0.1, \lambda_3 = 0.1$)	0.51 (0.50)	0.85	0.41	0.59	0.81	0.36	0.46	0.70	0.29	21	22.2
	ATMs ($\lambda_1 = 0.3, \lambda_2 = 0.1, \lambda_3 = 0.1$)	**0.62**	0.91	0.34	0.60	0.88	0.10	**0.70**	0.86	0.82	9	109.1
SSC	MFCCs (0.5)	0.51 (0.52)	0.84	0.33	0.52	0.83	0.08	0.5	0.73	0.44	36	13.5
	chroma (0.3)	0.40	0.67	0.24	0.35	0.51	0.14	0.41	0.70	0.14	54	13.2
	ATMs (0.5)	0.60	0.92	0.40	0.59	0.89	0.31	0.66 (0.67)	0.85	0.42	11	45.8
LRRSC	MFCCs (0.3)	0.44	0.79	0.31	0.40	0.72	0.15	0.47	0.80	0.28	55	141.2
	chroma (0.3)	0.39	0.53	0.26	0.30	0.41	0.12	0.39	0.72	0.19	69	133.8
	ATMs (0.9)	0.54	0.88	0.39	0.55	0.83	0.35	0.60	0.91	0.32	17	173.8
RRSC	MFCCs (0.3)	0.44	0.79	0.31	0.40	0.72	0.25	0.47	0.80	0.21	56	0.8
	chroma (0.3)	0.39	0.53	0.26	0.30	0.41	0.12	0.39	0.72	0.18	69	0.8
	ATMs (0.1)	0.57	0.91	0.35	0.62	0.87	0.43	0.59	0.92	0.25	12	0.9
NCuts on SDM	MFCCs	0.32	0.50	0.23	0.15	0.53	0.07	0.36	0.42	0.10	127	3.6
	chroma	0.32	0.48	0.22	0.15	0.34	0.07	0.36	0.62	0.08	118	3.4
	ATMs	0.41	0.63	0.26	0.32	0.58	0.12	0.49	0.62	0.16	46	3.6

For individual features, the experimental results in Table 1 indicate that: 1) the ENSC outperforms all the other methods with respect to all evaluation metrics employed. The PF and S_o gain of the ENSC against the other subspace clustering methods is statistically significant for the chroma features in the case of the SSC and for both the MFCCs and the chroma features in the case of the LRRSC and RRSC. Comparing the performance of the ENSC with that of the SDM, the reported improvements are statistically significant for all the features. 2) The SSC, the LRRSC, and the RRSC produce better segmentation results than the SDM-based structure segmentation for all evaluation metrics and features. The reported improvements in PF and S_o are statistically significant for the MFCCs and the chroma features. The same holds for all metrics in case of the ATMs. These results indicate that the SR-, the LRR-, the RR- the EN-based affinity matrices produce more reliable structure segmentation than the SDM, validating that the similarity measures employed by the subspace clustering methods are more robust than the distance-based similarity measure employed in the SDM. 3) The ATMs are more suitable for music segmentation than the MFCCs and the chroma features, when subspace clustering methods are employed. 4) The best parameters of the subspace clustering methods can be reliably determined using only 10 songs. Most importantly, the experimental findings do not alter, if these validation music recordings are excluded from the evaluation.

Table 2: Structure segmentation performance on the Beatles dataset with fixed $K = 5$ by employing cross-features affinity matrices.

Method	**Features (Parameters)**	PF			S_o			S_u			NoS
		Mean	Best	Worst	Mean	Best	Worst	Mean	Best	Worst	Mean
ENSC	MFCCs & chroma	0.55	0.87	0.43	0.62	0.80	0.34	0.52	0.85	0.35	18
	MFCCs & ATMs	**0.61**	0.87	0.37	0.64	0.88	0.49	0.63	0.80	0.36	9
	Chroma & ATMs	0.58	0.87	0.39	0.65	0.78	0.37	0.57	0.88	0.30	**10**
	MFCCs & chroma & ATMs	0.60	0.88	0.38	**0.66**	0.81	0.38	0.60	0.88	0.28	**10**
SSC	MFCCs & chroma	0.51	0.86	0.31	0.52	0.73	0.26	0.51	0.87	0.21	36
	MFCCs & ATMs	**0.61**	0.93	0.32	0.60	0.83	0.07	**0.65**	0.91	0.47	14
	Chroma & ATMs	0.57	0.89	0.34	0.58	0.83	0.32	0.63	0.90	0.29	13
	MFCCs & chroma & ATMs	0.60	0.92	0.33	0.61	0.83	0.41	0.64	0.92	0.24	13
LRRSC	MFCCs & chroma	0.43	0.71	0.32	0.37	0.67	0.23	0.46	0.70	0.20	55
	MFCCs & ATMs	0.53	0.83	0.35	0.54	0.85	0.38	0.59	0.73	0.22	18
	Chroma & ATMs	0.53	0.83	0.38	0.54	0.78	0.31	0.59	0.89	0.33	18
	MFCCs & chroma & ATMs	0.53	0.86	0.35	0.54	0.85	0.37	0.59	0.76	0.22	19
RRSC	MFCCs & chroma	0.43	0.71	0.32	0.36	0.67	0.23	0.46	0.69	0.20	56
	MFCCs & ATMs	0.56	0.88	0.35	0.62	0.84	0.43	0.58	0.89	0.25	13
	Chroma & ATMs	0.57	0.90	0.36	0.63	0.86	0.47	0.58	0.89	0.24	12
	MFCCs & chroma & ATMs	0.56	0.90	0.36	0.63	0.86	0.47	0.62	0.91	0.24	13
NCuts on SDM	MFCCs & chroma	0.34	0.54	0.23	0.19	0.45	0.12	0.38	0.61	0.11	105
	MFCCs & ATMs	0.38	0.63	0.25	0.28	0.57	0.09	0.44	0.68	0.12	79
	Chroma & ATMs	0.34	0.56	0.23	0.19	0.41	0.09	0.38	0.67	0.10	105
	MFCCs & chroma & ATMs	0.36	0.55	0.24	0.23	0.48	0.12	0.40	0.64	0.13	91

By inspecting Table 2, we can make the following remarks regarding the combination of multiple features. 1) Again, the ENSC outperforms all the subspace clustering methods that is compared to, with respect to all evaluation metrics employed. The only exception is the SSC, which outperforms the ENSC with respect to the S_o, when the MFCCs are combined with the ATMs. Moreover, in contrast to the competing subspace clustering methods, the ENSC is able to find the correct number of sections on average. 2) The subspace clustering methods achieve a better segmentation performance, which is statistically significant, than the SDM-based structure segmentation for all evaluation metrics and all feature combinations. This result combined with a similar observation made for individual feature vectors, highlights the potential of the similarity measures used in the subspace clustering methods to be employed as alternatives to SDM in (Chen and Ming, 2011; Weiss and Bello, 2010; Levy and Sandler, 2008; Paulus and Klapuri, 2009). 3) The best feature combination for each method in Table 2 includes the MFCCs and the ATMs always. If chroma features are also considered then the top S_0 is measured. The structure segmentation obtained by the combination of the MFCCs and the chroma features is not reliable, regardless the method employed. 4) Combining MFCCs and/or chroma features with ATMs yields a better segmentation than using the ATMs only with respect to the S_o and NoS in many cases.

Comparisons with methods in (Kaiser and Sikora, 2010; Levy and Sandler, 2008; Paulus and Klapuri, 2009): Here, the best segmentation results on the Beatles dataset are obtained by the ENSC, either when the ATMs are employed for audio representation (i.e., $PF = 0.62$, $S_o = 0.60$,

$S_u = 0.70$, $NoS = 9$), or when the MFCCs are combined with the ATMs (i.e., $PF = 0.61$, $S_o = 0.64$, $S_u = 0.63$, $NoS = 9$). These results can be fairly compared with those reported in (Kaiser and Sikora, 2010; Paulus and Klapuri, 2009) and the figures of merit of the method in (Levy and Sandler, 2008) as evaluated in (Paulus and Klapuri, 2009), since the same annotations from the TUT were employed. In particular, the method (Kaiser and Sikora, 2010) achieves $PF = 0.62$. The best results reported in (Paulus and Klapuri, 2009) on the Beatles dataset are as follows: $PF = 0.599$, $S_o = 0.604$, $S_u = 0.717$, $NoS = 10.3$. The method (Levy and Sandler, 2008) yields $PF = 0.584$, $S_o = 0.552$, $S_u = 0.683$, $NoS = 9.48$. Regarding to the segment boundary retrieval, the ENSC achieves on average $P = 0.54, R = 0.61, F = 0.55$, when the ATMs are employed and $P = 0.52, R = 0.61, F = 0.54$, when the MFCCs are combined with the ATMs. In the same task, the method (Paulus and Klapuri, 2009) yields $P = 0.52, R = 0.61, F = 0.55$. Thus, we conclude that the proposed method performs comparably with those in (Kaiser and Sikora, 2010; Paulus and Klapuri, 2009), while it outperforms the method in (Levy and Sandler, 2008).

Since either the ATMs or their combination with the MFCCs produce reliable structure segmentation, they are employed in order to automatically determine the actual number of section-types (i.e., clusters) of each music piece. The experimental findings are summarized, in Table 3. The ENSC outperforms the other methods for both individual features and combinations of multiple features with respect to all evaluation metrics but the S_0, where the RRSC yields a slightly higher value. Accordingly, it is possible to perform a robust music structure analysis in a fully automatic setting.

Table 3: Structure segmentation performance on the Beatles dataset with automatically determined K by employing (11).

Method	**Features (Parameters)**	PF			S_o			S_u			NoS
		Mean	Best	Worst	Mean	Best	Worst	Mean	Best	Worst	Mean
ENSC	ATMs	**0.59**	0.81	0.42	0.60	0.77	0.39	**0.68**	0.80	0.33	11
SSC	ATMs	0.52	0.87	0.37	0.53	0.88	0.28	0.65	0.84	0.51	8
LRRSC	ATMs	0.56	0.92	0.40	0.60	0.86	0.25	0.54	0.93	0.39	15
RRSC	ATMs	0.55	0.93	0.35	**0.61**	0.86	0.00	0.48	0.88	0.07	8
NCuts on SDM	ATMs	0.44	0.90	0.10	0.34	0.62	0.17	0.47	0.62	0.14	36
ENSC	MFCCs & ATMs	**0.58**	0.95	0.30	**0.60**	0.88	0.29	**0.69**	0.86	0.68	12
SSC	MFCCs & ATMs	0.56	0.85	0.40	0.58	0.84	0.25	0.58	0.74	0.39	17
LRRSC	MFCCs & ATMs	0.56	0.92	0.40	0.60	0.86	0.25	0.54	0.93	0.39	17
RRSC	MFCCs & ATMs	0.55	0.93	0.25	0.63	0.86	0.00	0.49	0.91	0.07	9
NCuts on SDM	MFCCs & ATMs	0.56	0.90	0.10	0.60	0.91	0.28	0.51	0.91	0.25	13

The experimental results indicate several advantages of the ENSC over the methods that is compared to in structure analysis of pop/rock music. However, the ENSC needs more computational time compared with the SSC, the RRSC, and the SDM, especially when high-dimensional features such as the ATMs are employed. The best results presented in Tables 1, 2 and 3 are obtained by analyzing songs with high between-section homogeneity such as the "Not a second time" by The Beatles. The worst results are mainly obtained in songs where the beats did not accurately estimated by the beat tracking algorithm (Ellis, 2007). The proposed approach for music structure analysis cannot be easily applied in music genres, such as free jazz, ambient, and non-Western genres music etc. where the notion of musical form does not resort to the homogeneity of the music sections.

5. Conclusions and future work

In this paper, it has been demonstrated that music structure analysis can be treated as a subspace clustering problem. A novel subspace clustering method (i.e., the ENSC) that builds on the elastic net representation of beat-synchronous audio features has been derived by solving (3) using the LADM. The experimental results on the Beatles dataset demonstrate the power of the ENSC.

In the future, the performance of the ENSC in music structure analysis can be improved with respect to the accuracy and computational effort by: 1) making the method independent of the beat tracking algorithms, 2) accelerating the convergence of Algorithm 1 by employing Nesterov-type acceleration step (Nesterov, 2004), and 3) reducing the dimensions of the ATMs using computational efficient dimensionality reduction methods, such as the random projections.

Acknowledgements

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heraclitus II. Investing in Knowledge Society through the European Social Fund.

Appendix

Solving subproblem (5):

In order to solve (5), we have to solve (4) with respect to $\mathbf{Z}$, which does not admit a closed form solution. Let $f(\mathbf{Z})$ be the smooth term in (4) i.e., $f(\mathbf{Z}) = \frac{\lambda_2}{2}\|\mathbf{Z}\|_F^2 + \operatorname{tr}\left(\mathbf{\Xi}^T(\mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E})\right) + \frac{\mu}{2}\|\mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E}\|_F^2$.

Following (Lin et al., 2011), $f(\mathbf{Z})$ is linearly approximated with respect to $\mathbf{Z}$ at $\mathbf{Z}_{[t]}$ as follows: $f(\mathbf{Z}) \approx f(\mathbf{Z}_{[t]}) + \operatorname{tr}\left((\mathbf{Z} - \mathbf{Z}_{[t]})^T \nabla f(\mathbf{Z}_{[t]})\right) + \frac{\mu\theta}{2}\|\mathbf{Z} - \mathbf{Z}_{[t]}\|_F^2$, where $\theta > 0$ is a proximal parameter and $\nabla f(\mathbf{Z}) = \lambda_2\mathbf{Z} - \mathbf{X}^T\mathbf{\Xi} + \mu(-\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X}\mathbf{Z} + \mathbf{X}^T\mathbf{E})$. Therefore, an approximate solution of (5) can be obtained by minimizing the partial linearized augmented Lagrangian function as follows:

$$\begin{aligned}
\mathbf{Z}_{[t+1]} &\approx \underset{\mathbf{Z}}{\operatorname{argmin}}\, \lambda_1\|\mathbf{Z}\|_1 + f(\mathbf{Z}_{[t]}) + \operatorname{tr}\left((\mathbf{Z}-\mathbf{Z}_{[t]})^T\nabla f(\mathbf{Z}_{[t]})\right) + \frac{\mu\theta}{2}\|\mathbf{Z}-\mathbf{Z}_{[t]}\|_F^2 \\
&= \underset{\mathbf{Z}}{\operatorname{argmin}}\, \lambda_1\|\mathbf{Z}\|_1 + \frac{\mu\theta}{2}\|\mathbf{Z}-(\mathbf{Z}_{[t]} - \frac{1}{\mu\theta}\nabla f(\mathbf{Z}_{[t]})\|_F^2 \\
&= \mathcal{S}_{\frac{\lambda_1}{\theta\mu}}\left[\mathbf{Z}_{[t]} + \frac{1}{\theta}\left(\mathbf{X}^T(\mathbf{X}-\mathbf{X}\mathbf{Z}_{[t]} - \mathbf{E}_{[t]} + \frac{1}{\mu}\boldsymbol{\Xi}_{[t]}) - \frac{\lambda_2}{\mu}\mathbf{Z}_{[t]}\right)\right]. \qquad (14)
\end{aligned}$$

References

Bertsekas, D.P., 1996. Constrained Optimization and Lagrange Multiplier Methods. Athena Scientific, Belmont, MA. 2nd edition.

Bhatia, R., Kittaneh, F., 1990. Norm inequalities for partitioned operators and an application. Math. Ann. 287, 719–726.

Candes, E.J., Li, X., Ma, Y., Wright, J., 2011. Robust principal component analysis? Journal of ACM 58, 1–37.

Chen, R., Ming, L., 2011. Music structural segmentation by combining harmonic and timbral information, in: Proc. 12th Int. Conf. Music Information Retrieval, Miami, USA. pp. 477–482.

Cheng, H., Liu, Z., Hou, L., Yang, J., 2012. Sparsity induced similarity measure and its applications. IEEE Trans. Circuits and Systems for Video Technology (accepted for publication).

Dannenberg, R.B., Goto, M., 2008. Music structure analysis from acoustic signals, in: Havelock, D., Kuwano, S., Vorländer, M. (Eds.), Handbook of Signal Processing in Acoustics. Springer, New York, N.Y., USA, pp. 305–331.

Ellis, D., 2007. Beat tracking by dynamic programming. J. New Music Research 36, 51–60.

He, B., Yuan, X., 2012. On the $o(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numer. Anal., 50, 700–709.

Kaiser, F., Sikora, T., 2010. Music structure discovery in popular music using non-negative matrix factorization, in: Proc. 11th Int. Conf. Music Information Retrieval, Utrecht, The Netherlands. pp. 429–434.

Levy, M., Sandler, M., 2008. Structural segmentation of musical audio by constrained clustering. IEEE Trans. Audio, Speech, and Language Processing 16, 318–326.

Lin, Z., Liu, R., Su, Z., 2011. Linearized alternating direction method with adaptive penalty for low-rank representation, in: Proc. 2011 Neural Information Processing Systems Conf., Granada, Spain. pp. 612–620.

Liu, G., Lin, Z., Yan, S., Sun, J., Ma, Y., 2013. Robust recovery of subspace structures by low-rank representation. IEEE Trans. Pattern Analysis and Machine Intelligence 35, 171–184.

Lukashevich, H., 2008. Towards quantitative measures of evaluating song segmentation, in: Proc. 9th Int. Conf. Music Inf. Retrieval, Philadelphia, PA, USA. pp. 375–380.

Lyon, R., 1982. A computational model of filtering, detection, and compression in the cochlea, in: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Paris, France. pp. 1282–1285.

Maddage, N.C., 2006. Automatic structure detection for popular music. IEEE MultiMedia 13, 65–77.

Manning, C., Raghavan, P., Schutze, H., 2008. Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA.

Nesterov, Y., 2004. Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Press, New York, NY.

Panagakis, Y., Kotropoulos, C., 2012a. Music structure analysis by ridge regression of beat-synchronous audio features, in: Proc. 13th Int. Conf. Music Information Retrieval, Porto, Portugal. pp. 271–276.

Panagakis, Y., Kotropoulos, C., 2012b. Music structure analysis by subspace modeling, in: Proc. 20th European Signal Processing Conf., Bucharest, Romania. pp. 1459–1463.

Panagakis, Y., Kotropoulos, C., Arce, G.R., 2010. Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. IEEE Trans. Audio, Speech, and Language Technology 18, 576–588.

Paulus, J., Klapuri, A., 2009. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. IEEE Trans. Audio, Speech, and Language Processing 17, 1159–1170.

Paulus, J., Müller, M., Klapuri, A., 2010. Audio-based music structure analysis, in: Proc. 11th Int. Conf. Music Information Retrieval, Utrecht, The Netherlands. pp. 625–636.

Ryynanen, M., Klapuri, A., 2008. Automatic transcription of melody, bass line, and chords in polyphonic music. Computer Music Journal 32, 72–86.

Shi, J., Malik, J., 2000. Normalized cuts and image segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence 22, 888–905.

Tan, Q.F., Georgiou, P.G., Narayanan, S., 2011. Enhanced sparse imputation techniques for a robust speech recognition front-end. IEEE Trans. Audio, Speech, and Language Processing 19, 2418 –2429.

Vidal, R., 2011. Subspace clustering. IEEE Signal Processing Magazine 28, 52–68.

Weiss, R., Bello, J., 2010. Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization, in: Proc. 11th Int. Conf. Music Information Retrieval, Utrecht, The Netherlands. pp. 123–128.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc., Series B 67, 301–320.