Middlesex University London

Middlesex University Research Repository

An open access repository of

Middlesex University research

http://eprints.mdx.ac.uk

Panagakis, Yannis and Kotropoulos, Constantine (2013) Music classification by low-rank semantic mappings. EURASIP Journal on Audio, Speech, and Music Processing, 2013 (1). p. 13. ISSN 1687-4714

Published version (with publisher's formatting)

This version is available at: http://eprints.mdx.ac.uk/23760/

Copyright:

Middlesex University Research Repository makes the University's research available electronically.

Copyright and moral rights to this work are retained by the author and/or other copyright owners unless otherwise stated. The work is supplied on the understanding that any use for commercial gain is strictly forbidden. A copy may be downloaded for personal, non-commercial, research or study without prior permission and without charge.

Works, including theses and research projects, may not be reproduced in any format or medium, or extensive quotations taken from them, or their content changed in any way, without first obtaining permission in writing from the copyright holder(s). They may not be sold or exploited commercially in any format or medium without the prior written permission of the copyright holder(s).

Full bibliographic details must be given when referring to, or quoting from full items including the author's name, the title of the work, publication details where relevant (place, publisher, date), pagination, and for theses or dissertations the awarding institution, the degree type awarded, and the date of the award.

If you believe that any material held in the repository infringes copyright law, please contact the Repository Team at Middlesex University via the following email address:

eprints@mdx.ac.uk

The item will be removed from the repository while any claim is being investigated.

See also repository copyright: re-use policy: http://eprints.mdx.ac.uk/policies.html#copy

Panagakis and Kotropoulos *EURASIP Journal on Audio, Speech and Music Processing* 2013, **2013**:13
http://asmp.eurasipjournals.com/content/2013/1/13

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Music classification by low-rank semantic mappings

Yannis Panagakis* and Constantine Kotropoulos

Abstract

A challenging open question in music classification is which music representation (i.e., audio features) and which machine learning algorithm is appropriate for a specific music classification task. To address this challenge, given a number of audio feature vectors for each training music recording that capture the different aspects of music (i.e., timbre, harmony, etc.), the goal is to find a set of linear mappings from several feature spaces to the semantic space spanned by the class indicator vectors. These mappings should reveal the common latent variables, which characterize a given set of classes and simultaneously define a multi-class linear classifier that classifies the extracted latent common features. Such a set of mappings is obtained, building on the notion of the maximum margin matrix factorization, by minimizing a weighted sum of nuclear norms. Since the nuclear norm imposes rank constraints to the learnt mappings, the proposed method is referred to as *low-rank semantic mappings* (LRSMs). The performance of the LRSMs in music genre, mood, and multi-label classification is assessed by conducting extensive experiments on seven manually annotated benchmark datasets. The reported experimental results demonstrate the superiority of the LRSMs over the classifiers that are compared to. Furthermore, the best reported classification results are comparable with or slightly superior to those obtained by the state-of-the-art task-specific music classification methods.

Keywords: Music classification; Music genre; Music mood; Nuclear norm minimization; Auditory representations

1 Introduction

Retail and online music stores usually index their collections by artist or album name. However, people often need to search for music by content. For example, a search facility is offered by emerging music-oriented recommendation services, such as *last.fm* (http://www.last.fm/) and *Pandora* (http://www.pandora.com/), where social tags are employed as semantic descriptors of the music content. Social tags are text-based labels, provided by either human experts or amateur users to categorize music with respect to genre, mood, and other semantic tags. The major drawbacks of this approach for the semantic annotation of music content are (1) a newly added music recording must be tagged manually, before it can be retrieved [1], which is a time-consuming and expensive process and (2) unpopular music recordings may not be tagged at all [2]. Consequently, an accurate content-based automatic classification of music should be exploited to mitigate the just mentioned drawbacks, allowing the deployment of robust music browsing and recommendation engines.

A considerable volume of research in content-based music classification has been conducted so far. The interested reader may refer to [2-5] for a comprehensive survey. Most music classification methods focus on music categorization with respect to genre, mood, or multiple semantic tags. They consist mainly of two stages, namely a music representation stage and a machine learning one. In the first stage, the various aspects of music (i.e., the timbral, the harmonic, the rhythmic content, etc.) are captured by extracting either low- or mid-level features from the audio signal. Such features include timbral texture features, rhythmic features, pitch content, or their combinations, yielding a *bag-of-features* (BOF) representation [1,2,6-18]. Furthermore, spectral, cepstral, and auditory modulation-based features have been recently employed either in BOF approaches or as autonomous music representations in order to capture both the timbral and the temporal struc-

*Correspondence: panagakis@aiia.csd.auth.gr
Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 54124, Greece

Springer

© 2013 Panagakis and Kotropoulos; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ture of music [19-22]. At the machine learning stage, music genre and mood classification are treated as single-label multi-class classification problems. To this end, support vector machines (SVMs) [23], nearest-neighbor (NN) classifiers, Gaussian mixture model-based ones [3], and classifiers relying on sparse and low-rank representations [24] have been employed to classify the audio features into genre or mood classes. On the contrary, automatic music tagging (or autotagging) is considered as a multi-label, multi-class classification problem. A variety of algorithms have been exploited in order to associate the tags with the audio features. For instance, music tag prediction may be treated as a set of binary classification problems, where standard classifiers, such as the SVMs [12,14] or ada-boost [25], can be applied. Furthermore, probabilistic autotagging systems have been proposed, attempting to infer the correlations or joint probabilities between the tags and the audio features [1,9,26].

Despite the existence of many well-performing music classification methods, it is still unclear which music representation (i.e., audio features) and which machine learning algorithm is appropriate for a specific music classification task. A possible explanation for the aforementioned open question is that the classes (e.g., genre, mood, or other semantic classes) in music classification problems are related to and built on some common unknown latent variables, which are different in each problem. For instance, many different songs, although they share instrumentation (i.e., have similar timbral characteristics), convey different emotions and belong to different genres. Furthermore, cover songs, which have the same harmonic content with the originals, may differ in the instrumentation and possibly evoke a different mood, so they are classified into different genres. Therefore, the challenge is to reveal the common latent features based on given music representations, such as timbral, auditory, etc., and to simultaneously learn the models that are appropriate for each specific classification task.

In this paper, a novel, robust, general-purpose music classification method is proposed to address the aforementioned challenge. It is suitable for both single-label (i.e., genre or mood classification) and multi-label (i.e., music tagging) multi-class classification problems, providing a systematic way to handle multiple audio features capturing the different aspects of music. In particular, given a number of audio feature vectors for each training music recording, the goal is to find a set of linear mappings from the feature spaces to the semantic space defined by the class indicator vectors. Furthermore, these mappings should reveal the common latent variables, which characterize a given set of classes and simultaneously define a multi-class linear classifier that classifies the extracted latent common features. Such a model can be derived by building on the notion of the maximum margin matrix factorization [27]. That is, in the training phase, the set of mappings is found by minimizing a weighted sum of nuclear norms. To this end, an algorithm that resorts to the alternating direction augmented Lagrange multiplier method [28] is derived. In the test phase, the class indicator vector for labeling any test music recording is obtained by multiplying each mapping matrix with the corresponding feature vector and by summing all the resulting vectors next. Since the nuclear norm imposes rank constraints to the learnt mappings, the proposed classification method is referred to as *low-rank semantic mappings* (LRSMs).

The motivation behind the LRSMs arises from the fact that uncovering hidden shared variables among the classes facilitates the learning process [29]. To this end, various formulations for common latent variable extraction have been proposed for multi-task learning [30], multi-class classification [31], collaborative prediction [32], and multi-label classification [33]. The LRSMs differ significantly from the aforementioned methods [29-31,33] in that the extracted common latent variables come from many different (vector) feature spaces.

The performance of the LRSMs in music genre, mood, and multi-label classification is assessed by conducting experiments on seven manually annotated benchmark datasets. Both the standard evaluation protocols for each dataset and a small sample size setting are employed. The auditory cortical representations [34,35], the mel-frequency cepstral coefficients [36], and the chroma features [37] were used for music representation. In the single-label case (i.e., genre or mood classification), the LRSMs are compared against three well-known classifiers, namely the sparse representation-based classifier (SRC) [38], the linear SVMs, and the NN classifier with a cosine distance metric. Multi-label extensions of the aforementioned classifiers, namely the multi-label sparse representation-based classifier (MLSRC)[39], the Rank-SVMs [40], and the multi-label k-nearest neighbor (MLkNN) [41], as well as the parallel factor analysis 2 (PARAFAC2)-based autotagging method [42] are compared with the LRSMs in music tagging. The reported experimental results demonstrate the superiority of the LRSMs over the classifiers that are compared to. Moreover, the best classification results disclosed are comparable with or slightly superior to those obtained by the state-of-the-art music classification systems.

To summarize, the contributions of the paper are as follows:

- A novel method for music classification (i.e., the LRSMs) is proposed that is able to extract the common latent variables that are shared among all

the classes and simultaneously learn the models that are appropriate for each specific classification task.

- An efficient algorithm for the LRSMs is derived by resorting to the alternating direction augmented Lagrange multiplier method, which is suitable for large-scale data.
- The LRSMs provide a systematic way to handle multiple audio features for music classification.
- Extensive experiments on seven datasets demonstrate the effectiveness of the LRSMs in music genre, mood, and multi-label classification when the mel-frequency cepstral coefficients (MFCCs), the chroma, and the auditory cortical representations are employed for music representation.

The paper is organized as follows: In Section 2, basic notation conventions are introduced. The audio feature extraction process is briefly described in Section 3. In Section 4, the LRSMs are detailed. Datasets and experimental results are presented in Section 5. Conclusions are drawn in Section 6.

2 Notations

Throughout the paper, matrices are denoted by uppercase boldface letters (e.g., $\mathbf{X}, \mathbf{L}$), vectors are denoted by lowercase boldface letters (e.g., $\mathbf{x}$), and scalars appear as either uppercase or lowercase letters (e.g., N, K, i, μ, ϵ). $\mathbf{I}$ denotes the identity matrix of compatible dimensions. The ith column of $\mathbf{X}$ is denoted as $\mathbf{x}_i$. The set of real numbers is denoted by $\mathbb{R}$, while the set of nonnegative real numbers is denoted by $\mathbb{R}_+$.

A variety of norms on real-valued vectors and matrices will be used. For example, $\|\mathbf{x}\|_0$ is ℓ_0 quasi-norm counting the number of nonzero entries in $\mathbf{x}$. The matrix ℓ_1 norm is denoted by $\|\mathbf{X}\|_1 = \sum_i \sum_j |x_{ij}|$. $\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j x_{ij}^2} = \sqrt{\mathrm{tr}(\mathbf{X}^T\mathbf{X})}$ is the Frobenius norm, where tr(.) denotes the trace of a square matrix. The nuclear norm of $\mathbf{X}$ (i.e., the sum of singular values of a matrix) is denoted by $\|\mathbf{X}\|_*$. The ℓ_∞ norm of $\mathbf{X}$, denoted by $\|\mathbf{X}\|_\infty$, is defined as the element of $\mathbf{X}$ with the maximum absolute value.

3 Audio feature extraction

Each music recording is represented by three song-level feature vectors, namely the auditory cortical representations [34,35], the MFCCs [36], and the chroma features [37]. Although much more elaborated music representations have been proposed in the literature, the just mentioned features perform quite well in practice [14,22-24]. Most importantly, song-level representations are suitable for large-scale music classification problems since the space complexity for audio processing and analysis is reduced and the database overflow is prevented [3].

3.1 Auditory cortical representations

The auditory cortex plays a crucial role in the hearing process since auditory sensations turn into perception and cognition only when they are processed by the cortical area. Therefore, one should focus on how audio information is encoded in the human primary auditory cortex in order to represent music signals in a psychophysiologically consistent manner [43]. The mechanical and neural processing in the early and central stages of the auditory system can be modeled as a two-stage process. At the first stage, which models the cochlea, the audio signal is converted into an auditory representation by employing the constant-Q transform (CQT).The CQT is a time-frequency representation, where the frequency bins are geometrically spaced and the Q-factors (i.e., the ratios of the center frequencies to the bandwidths) of all bins are equal [44]. The neurons in the primary auditory cortex are organized according to their selectivity in different spectral and temporal stimuli [43]. To this end, in the second stage, the spectral and temporal modulation content of the CQT is estimated by two-dimensional (2D) multi-resolution wavelet analysis, ranging from slow to fast temporal rates and from narrow to broad spectral scales. The analysis yields a four-dimensional (4D) representation of time, frequency, rate, and scale that captures the slow spectral and temporal modulation content of audio that is referred to as *auditory cortical representation* [34]. Details on the mathematical formulation of the auditory cortical representations can be found in [34,35].

In this paper, the CQT is computed efficiently by employing the fast implementation scheme proposed in [44]. The audio signal is analyzed by employing 128 constant-Q filters covering eight octaves from 44.9 Hz to 11 KHz (i.e., 16 filters per octave). The magnitude of the CQT is compressed by raising each element of the CQT matrix to the power of 0.1. At the second stage, the 2D multi-resolution wavelet analysis is implemented via a bank of 2D Gaussian filters with *scales* $\in \{0.25, 0.5, 1, 2, 4, 8\}$ (cycles/octave) and (both positive and negative) *rates* $\in \{\pm 2, \pm 4, \pm 8, \pm 16, \pm 32\}$ (Hz). The choice of the just mentioned parameters is based on psychophysiological evidence [34]. For each music recording, the extracted 4D cortical representation is time-averaged, and the 3D rate-scale-frequency cortical representation is obtained. The overall procedure is depicted in Figure 1. Accordingly, each music recording can be represented by a vector $\mathbf{x} \in \mathbb{R}_+^{7,680}$ by stacking the elements of the 3D cortical representation into a vector. The dimension of the vectorized cortical representation comes from the product of 128 frequency channels, 6 scales, and 10 rates. An

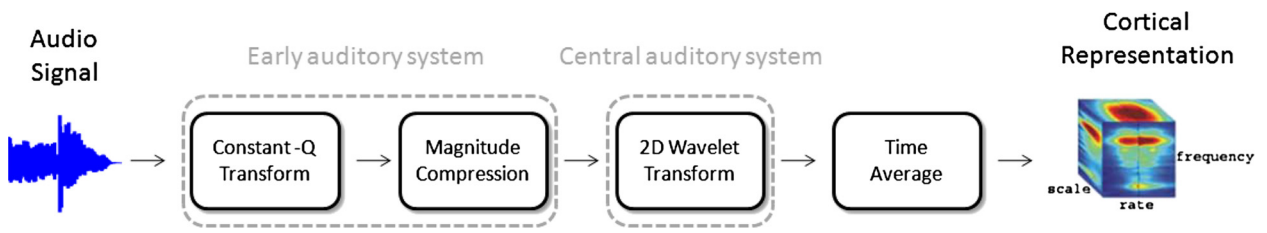

Figure 1 Flow chart of auditory cortical representation extraction.

ensemble of music recordings is represented by the data matrix $\mathbf{X} \in \mathbb{R}_{+}^{7,680 \times S}$, where S is the number of the available recordings in each dataset. Finally, the entries of $\mathbf{X}$ are post-processed as follows: Each row of $\mathbf{X}$ is normalized to the range $[0, 1]$ by subtracting from each entry the row minimum and then by dividing it with the range (i.e., the difference between the row maximum and the row minimum).

3.2 Mel-frequency cepstral coefficients

The MFCCs encode the timbral properties of the music signal by encoding the rough shape of the log-power spectrum on the mel-frequency scale [36]. They exhibit the desirable property that a numerical change in the MFCC coefficients corresponds to a perceptual change. In this paper, MFCC extraction employs frames of 92.9-ms duration with a hop size of 46.45 ms and a 42 band-pass filter bank. The filters are uniformly spaced on the mel-frequency scale. The correlation between the frequency bands is reduced by applying the discrete cosine transform along the log-energies of the bands yielding a sequence of 20-dimensional MFCC vectors. By averaging the MFCCs along the time axis, each music recording is represented by a 20-dimensional MFCC vector.

3.3 Chroma features

The chroma features [37] are adept in characterizing 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. They are calculated by employing 92.9 ms frames with a hop size of 23.22 ms as follows: First, the salience of different fundamental frequencies in the range 80 to 640 Hz is calculated. The linear frequency scale is transformed into a musical one by selecting the maximum salience value in each frequency range corresponding to one semitone. Finally, the octave equivalence classes are summed over the whole pitch range to yield a sequence of 12-dimensional chroma vectors.

The chroma as well as the MFCCs, extracted from an ensemble of music recordings, is post-processed as described in subsection 3.1.

4 Classification by low-rank semantic mappings

Let each music recording be represented by R types of feature vectors of size d_r, $\mathbf{x}^{(r)} \in \mathbb{R}^{d_r}$, $r = 1, 2, \ldots, R$. Consequently, an ensemble of N training music recordings is represented by the set $\left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(R)}\right\}$, where $\mathbf{X}^{(r)} = \left[\mathbf{x}_1^{(r)}, \mathbf{x}_2^{(r)}, \ldots, \mathbf{x}_N^{(r)}\right] \in \mathbb{R}^{d_r \times N}$, $r = 1, 2, \ldots, R$. The class labels of the N training samples are represented as indicator vectors forming the matrix $\mathbf{L} \in \{0, 1\}^{K \times N}$, where K denotes the number of classes. Clearly, $l_{kn} = 1$ if the nth training sample belongs to the kth class. In a multi-label setting, more than one non-zero elements may appear in the class indicator vector $\mathbf{l}_n \in \{0, 1\}^K$.

These R different feature vectors characterize different aspects of music (i.e., timbre, rhythm, harmony, etc.), having different properties, and thus, they live in different (vector) feature spaces. Since different feature vectors have different intrinsic discriminative power, an intuitive idea is to combine them in order to improve the classification performance. However, in practice, most of the machine learning algorithms can handle only a single type of feature vectors and thus cannot be naturally applied to multiple features. A straightforward strategy to handle multiple features is to concatenate all the feature vectors into a single feature vector. However, the resulting feature space is rather *ad hoc* and lacks physical interpretation. It is more reasonable to assume that multiple feature vectors live in a union of feature spaces, which is what the proposed method actually does in a principled way. Leveraging information contained in multiple features can dramatically improve the learning performance as indicated by the recent results in multi-view learning [30,45].

Given a set of (possibly few) training samples along with the associated class indicator vectors, the goal is to learn R mappings $\mathbf{M}^{(r)} \in \mathbb{R}^{K \times d_r}$ from the feature spaces $\mathbb{R}^{d_r}$, $r = 1, 2, \ldots, R$, to the label space $\{0, 1\}^K$, having a generalization ability and appropriately utilizing the cross-feature information, so that

$$\mathbf{L} = \sum_{r=1}^{R} \mathbf{M}^{(r)} \mathbf{X}^{(r)}. \tag{1}$$

As discussed in Section 1, the mappings $\mathbf{M} \in \mathbb{R}^{K\times d_r}, r = 1, 2, \ldots, R$, should be able to (1) reveal the common latent variables across the classes and (2) predict simultaneously the class memberships based on these latent variables. To do this, we seek for $\mathbf{C}^{(r)} \in \mathbb{R}^{K\times p_r}$ and $\mathbf{F} \in \mathbb{R}^{p_r\times d_r}$, such that $\mathbf{M}^{(r)} = \mathbf{C}^{(r)}\mathbf{F}^{(r)} \in \mathbb{R}^{K\times d_r}$, $r = 1, 2, \ldots, R$. In this formulation, the rows of $\mathbf{F}^{(r)}$ reveal the p_r latent features (variables), and the rows of $\mathbf{C}^{(r)}$ are the weights predicting the classes. Clearly, the number of p_r common latent variables and the matrices $\mathbf{C}^{(r)}$, $\mathbf{F}^{(r)}$ are unknown and need to be jointly estimated.

Since the dimensionality of the R latent feature spaces (i.e., p_r) is unknown, inspired by maximum margin matrix factorization [27], we can allow the unknown matrices $\mathbf{C}^{(r)}$ to have an unbounded number of columns and $\mathbf{F}^{(r)}$, $r = 1, 2, \ldots, R$ to have an unbounded number of rows. Here, the matrices $\mathbf{C}^{(r)}$ and $\mathbf{F}^{(r)}$ are required to be low-norm. This constraint is mandatory because otherwise the resulting linear transform induced by applying first $\mathbf{F}^{(r)}$ and then $\mathbf{C}^{(r)}$ would degenerate to a single transform. Accordingly, the unknown matrices are obtained by solving the following minimization problem:

$$\begin{aligned}\underset{\{\mathbf{C}^{(r)},\mathbf{F}^{(r)}|_{r=1}^{R}\}}{\operatorname{argmin}} &\sum_{r=1}^{R}\frac{\lambda_r}{2}\left(\|\mathbf{C}^{(r)}\|_F^2 + \|\mathbf{F}^{(r)}\|_F^2\right)\\ &+\frac{1}{2}\|\mathbf{L} - \sum_{r=1}^{R}\mathbf{C}^{(r)}\mathbf{F}^{(r)}\mathbf{X}^{(r)}\|_F^2,\end{aligned} \tag{2}$$

where λ_r, $r = 1, 2, \ldots, R$, are regularization parameters and the least squares loss function $\frac{1}{2}\|\mathbf{L} - \sum_{r=1}^{R}\mathbf{C}^{(r)}\mathbf{F}^{(r)}\mathbf{X}^{(r)}\|_F^2$ measures the labeling approximation error. It is worth mentioning that the least squares loss function is comparable to other loss functions, such as the hinge loss employed in SVMs [46], since it has been proved to be (universally) Fisher consistent [47]. This property along with the fact that it leads into the formulation of a tractable optimization problem motivated us to adopt the least squares loss here. By Lemma 1 in [27], it is known that

$$\lambda_r\|\mathbf{M}^{(r)}\|_* = \underset{\mathbf{M}^{(r)}=\mathbf{F}^{(r)}\mathbf{C}^{(r)}}{\operatorname{argmin}}\frac{\lambda_r}{2}\left(\|\mathbf{C}^{(r)}\|_F^2 + \|\mathbf{F}^{(r)}\|_F^2\right). \tag{3}$$

Thus, based on (3), the optimization problem (2) can be rewritten as

$$\underset{\{\mathbf{M}^{(r)}|_{r=1}^{R}\}}{\operatorname{argmin}}\sum_{r=1}^{R}\lambda_r\|\mathbf{M}^{(r)}\|_* + \frac{1}{2}\|\mathbf{L} - \sum_{r=1}^{R}\mathbf{M}^{(r)}\mathbf{X}^{(r)}\|_F^2. \tag{4}$$

Therefore, the mappings $\mathbf{M}^{(r)}$, $r = 1, 2, \ldots R$, are obtained by minimizing the weighted sum of their nuclear norms and the labeling approximation error, that is, the nuclear norm-regularized least squares labeling approximation error. Since the nuclear norm is the convex envelope of the rank function [48], the derived mappings between the feature spaces and the semantic space spanned by the class indicator matrix $\mathbf{L}$ are low-rank as well. This justifies why the solution of (4) yields low-rank semantic mappings (LRSMs). The LRSMs are strongly related and share the same motivations with the methods in [31] and [32], which have been proposed for multi-class classification and prediction, respectively. In both methods, the nuclear norm-regularized loss is minimized in order to infer relationships between the label vectors and feature vectors. The two key differences between the methods in [31] and [32] and the LRSMs are (1) the LRSMs are able to adequately handle multiple features, drawn from different feature spaces, and (2) the least squares loss function is employed instead of hinge loss, resulting into formulation (4) which can be efficiently solved for large-scale data.

Problem (4) is solved as follows: By introducing the auxiliary variables $\mathbf{W}^{(r)}$, $r = 1, 2, \ldots, R$, (4) is equivalent to

$$\begin{aligned}\underset{\{\mathbf{M}^{(r)},\mathbf{W}^{(r)}|_{r=1}^{R}\}}{\operatorname{argmin}} &\sum_{r=1}^{R}\lambda_r\|\mathbf{W}^{(r)}\|_* + \frac{1}{2}\|\mathbf{L} - \sum_{r=1}^{R}\mathbf{M}^{(r)}\mathbf{X}^{(r)}\|_F^2\\ &\text{s.t. } \mathbf{M}^{(r)} = \mathbf{W}^{(r)}, r = 1, 2, \ldots, R,\end{aligned} \tag{5}$$

which can be solved by employing the *alternating direction augmented Lagrange multiplier* (ADALM) method, which is a simple, but powerful, algorithm that is well suited to large-scale optimization problems [28,49]. That is, by minimizing the augmented Lagrange function [28],

$$\begin{aligned}&\mathcal{L}\left(\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \ldots \mathbf{W}^{(R)}, \mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \ldots, \mathbf{M}^{(R)}, \Xi^{(1)},\right.\\ &\left.\Xi^{(2)}, \ldots, \Xi^{(R)}\right) = \sum_{r=1}^{R}\lambda_r\|\mathbf{W}^{(r)}\|_* + \frac{1}{2}\|\mathbf{L} - \sum_{r=1}^{R}\mathbf{M}^{(r)}\mathbf{X}^{(r)}\|_F^2\\ &+\sum_{r=1}^{R}\operatorname{tr}\left(\Xi^{(r)T}\left(\mathbf{M}^{(r)} - \mathbf{W}^{(r)}\right)\right) + \frac{\zeta}{2}\sum_{r=1}^{R}\|\mathbf{M}^{(r)} - \mathbf{W}^{(r)}\|_F^2,\end{aligned} \tag{6}$$

where $\Xi^{(r)}$, $r = 1, 2, \ldots, R$, are the Lagrange multipliers and $\zeta > 0$ is a penalty parameter. By applying the ADALM, (6) is minimized with respect to each variable in an alternating fashion, and finally, the Lagrange multipliers are updated at each iteration. If only $\mathbf{W}^{(1)}$ is varying and all the other variables are kept fixed, we simplify (6) writing $\mathcal{L}(\mathbf{W}^{(1)})$ instead of $\mathcal{L}\big(\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \ldots, \mathbf{W}^{(R)}, \mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \ldots, \mathbf{M}^{(R)}, \Xi^{(1)}, \Xi^{(2)}, \ldots, \Xi^{(R)}\big)$. Let t denote the iteration index. Given $\mathbf{W}_{[t]}^{(r)}, \mathbf{M}_{[t]}^{(r)}$, $r = 1, 2, \ldots R$, and $\zeta_{[t]}$, the

iterative scheme of ADALM for (6) reads as follows:

$$\begin{aligned}\mathbf{W}_{[t+1]}^{(r)} &= \underset{\mathbf{W}_{[t]}^{(r)}}{\operatorname{argmin}} \mathcal{L}\left(\mathbf{W}_{[t]}^{(r)}\right)\\ &= \underset{\mathbf{W}_{[t]}^{(r)}}{\operatorname{argmin}} \lambda_r \|\mathbf{W}_{[t]}^{(r)}\|_* + \operatorname{tr}\left(\Xi_{[t]}^{(r)T}\left(\mathbf{M}_{[t]}^{(r)} - \mathbf{W}_{[t]}^{(r)}\right)\right)\\ &+ \frac{\zeta_{[t]}}{2}\|\mathbf{M}_{[t]}^{(r)} - \mathbf{W}_{[t]}^{(r)}\|_F^2\\ &= \underset{\mathbf{W}_{[t]}^{(r)}}{\operatorname{argmin}} \frac{\lambda_r}{\zeta_{[t]}}\|\mathbf{W}_{[t]}^{(r)}\|_* + \frac{1}{2}\|\mathbf{W}_{[t]}^{(r)} - \left(\mathbf{M}_{[t]}^{(r)} + \frac{\Xi_{[t]}^{(r)}}{\zeta_{[t]}}\right)\|_F^2.\end{aligned} \tag{7}$$

$$\begin{aligned}\mathbf{M}_{[t+1]}^{(r)} &= \underset{\mathbf{M}_{[t]}^{(r)}}{\operatorname{argmin}} \mathcal{L}\left(\mathbf{M}_{[t]}^{(r)}\right)\\ &= \underset{\mathbf{M}_{[t]}^{(r)}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{L} - \sum_{r=1}^{R}\mathbf{M}_{[t]}^{(r)}\mathbf{X}^{(r)}\|_F^2\\ &+ \operatorname{tr}\left(\Xi_{[t]}^{(r)T}\left(\mathbf{M}_{[t]}^{(r)} - \mathbf{W}_{[t+1]}^{(r)}\right)\right)\\ &+ \frac{\zeta_{[t]}}{2}\|\mathbf{M}_{[t]}^{(r)} - \mathbf{W}_{[t+1]}^{(r)}\|_F^2.\end{aligned} \tag{8}$$

$$\Xi_{[t+1]}^{(r)} = \Xi_{[t]}^{(r)} + \zeta_{[t]}\left(\mathbf{M}_{[t+1]}^{(r)} - \mathbf{W}_{[t+1]}^{(r)}\right), \quad r = 1, 2, \ldots, R. \tag{9}$$

The solution of (7) is obtained in closed form via the singular value thresholding operator defined for any matrix $\mathbf{Q}$ as [50]: $\mathcal{D}_\tau[\mathbf{Q}] = \mathbf{U}\mathcal{S}_\tau\mathbf{V}^T$ with $\mathbf{Q} = \mathbf{U}\Sigma\mathbf{V}^T$ being the singular value decomposition and $\mathcal{S}_\tau[q] = \operatorname{sgn}(q)\max(|q| - \tau, 0)$ being the shrinkage operator [51]. The shrinkage operator can be extended to matrices by applying it element-wise. Consequently, $\mathbf{W}_{[t+1]}^{(r)} = \mathcal{D}_{\frac{\lambda_r}{\zeta_{[t]}}}\left[\mathbf{M}_{[t]}^{(r)} + \frac{\Xi_{[t]}^{(r)}}{\zeta_{[t]}}\right]$. Problem (8) is an unconstrained least squares problem, which admits a unique closed-form solution, as is indicated in Algorithm 1 summarizing the ADALM method for the minimization of (5). The convergence of Algorithm 1 is just a special case of that of the generic ADALM [28,49].

The set of the low-rank semantic matrices $\left\{\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \ldots, \mathbf{M}^{(R)}\right\}$, obtained by Algorithm 1, captures the semantic relationships between the label space and the R audio feature spaces. In music classification, the semantic relationships are expected to propagate from the R feature spaces to the label vector space. Therefore, a test music recording can be labeled as follows: Let $\hat{\mathbf{x}}^{(r)} \in \mathbb{R}^{d_r}$, $r = 1, 2, \ldots, R$, be a set of feature vectors extracted from the test music recording and $\mathbf{l} \in \{0, 1\}^K$ be the class indicator vector of this recording. First, the intermediate class indicator vector $\hat{\mathbf{l}} \in \mathbb{R}^K$ is obtained by

$$\hat{\mathbf{l}} = \sum_{r=1}^{R}\mathbf{M}^{(r)}\,\mathbf{x}^{(r)}. \tag{10}$$

Algorithm 1 Solving (5) by the ADALM method

Input: Training data $\left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(R)}\right\}$, the class indicator matrix $\mathbf{L}$, and the regularization parameters λ_r, $r = 1, 2, \ldots, R$.

Output: The set of the low-rank semantic matrices $\left\{\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \ldots, \mathbf{M}^{(R)}\right\}$.

1: Initialize: Set $\mathbf{M}_{[0]}^{(r)}$, $\mathbf{W}_{[0]}^{(r)}$, and $\Xi_{[0]}^{(r)}$, $r = 1, 2, \ldots, R$, to zero matrices of compatible dimensions, $\zeta_{[0]} = 10^{-6}$, $t = 0$, $\rho = 1.1$, $\epsilon = 10^{-8}$.
2: **while** not converged **do**
3: **for** $r = 1 \to R$ **do**
4: Fix the other variables, and update $\mathbf{W}_{[t+1]}^{(r)}$ by
5: $\mathbf{W}_{[t+1]}^{(r)} \leftarrow \mathcal{D}_{\frac{\lambda_r}{\zeta_{[t]}}}\left[\mathbf{M}_{[t]}^{(r)} + \frac{\Xi_{[t]}^{(r)}}{\zeta_{[t]}}\right]$.
6: Fix the other variables, and update $\mathbf{M}_{[t+1]}^{(r)}$ by
7: $\mathbf{T} \leftarrow \sum_{r'=1, r' \neq r}^{R}\mathbf{M}_{[t]}^{(r')}\mathbf{X}^{(r')}$.
8: $\mathbf{M}_{[t+1]}^{(r)} \leftarrow \left(\mathbf{L}\mathbf{X}^{(r)T} - \mathbf{T}\mathbf{X}^{(r)T} + \zeta_{[t]}\mathbf{W}_{[t+1]}^{(r)} - \Xi_{[t]}^{(r)T}\right)$
9: $\left(\mathbf{X}^{(r)}\mathbf{X}^{(r)T} + \zeta_{[t]}\mathbf{I}\right)^{-1}$.
10: Update the Lagrange multipliers by
11: $\Xi_{[t+1]}^{(r)} \leftarrow \Xi_{[t]}^{(r)} + \zeta_{[t]}\left(\mathbf{M}_{[t+1]}^{(r)} - \mathbf{W}_{[t+1]}^{(r)}\right)$.
12:
13: **end for**
14: Update $\zeta_{[t+1]}$ by $\zeta_{[t+1]} \leftarrow \min(\rho \cdot \zeta_{[t]}, 10^6)$.
15:
16: **for** $r = 1 \to R$ **do**
17: Check convergence conditions
18: $\|\mathbf{M}_{[t+1]}^{(r)} - \mathbf{W}_{[t+1]}^{(r)}\|_\infty < \epsilon$.
19: **end for**
20: $t \leftarrow t + 1$.
21: **end while**

The (final) class indicator vector (i.e., $\mathbf{l}$) has $\|\mathbf{l}\|_0 = \nu < K$, containing 1 in the positions, which are associated with the ν largest values in $\hat{\mathbf{l}}$. Clearly, for single-label multi-class classification, $\nu = 1$.

4.1 Computational complexity

The dominant cost for each iteration in Algorithm 1 is the computation of the singular value thresholding operator (i.e., step 4), that is, the calculation of the singular vectors of $\mathbf{M}_{[t]}^{(r)} + \frac{\Xi_{[t]}^{(r)}}{\zeta_{[t]}}$ whose corresponding singular values are larger than the threshold $\frac{\lambda_r}{\zeta_{[t]}}$. Thus, the complexity of each iteration is $O(R \cdot d \cdot N^2)$.

Since the computational cost of the LRSMs depends highly on the dimensionality of feature spaces, dimensionality reduction methods can be applied. For computational tractability, dimensionality reduction via random projections is considered. Let the true low dimensionality

of the data be denoted by z. Following [52], a random projection matrix, drawn from a normal zero-mean distribution, provides with high probability a *stable embedding* [53] with the dimensionality of the projection d'_r selected as the minimum value such that $d'_r > 2z\log(7,680/d'_r)$. Roughly speaking, a stable embedding approximately preserves the Euclidean distances between all vectors in the original space in the feature space of reduced dimensions. In this paper, we propose to estimate z by robust principal component analysis [51] on the high-dimensional training data (e.g., $\mathbf{X}^{(r)}$). That is, the principal component pursuit is solved:

$$\underset{\mathbf{L}^{(r)},\mathbf{S}^{(r)}}{\operatorname{argmin}} \|\Gamma^{(r)}\|_* + \lambda\|\Delta^{(r)}\|_1 \;\text{ s.t. }\; \mathbf{X}^{(r)} = \Gamma^{(r)} + \Delta^{(r)}. \tag{11}$$

Then, z is the rank of the outlier-free data matrix $\Gamma^{(r)}$ [51] and corresponds to the number of its non-zero singular values.

5 Experimental evaluation

5.1 Datasets and evaluation procedure

The performance of the LRSMs in music genre, mood, and multi-label music classification is assessed by conducting experiments on seven manually annotated benchmark datasets for which the audio files are publicly available. In particular, the GTZAN [17], ISMIR, Homburg [54], Unique [16], and 1517-Artists [16] datasets are employed for music genre classification, the MTV dataset [15] for music mood classification, and the CAL500 dataset [1] for music tagging. Brief descriptions of these datasets are provided next.

The *GTZAN* (http://marsyas.info/download/data_sets) consists of 10 genre classes, namely blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each genre class contains 100 excerpts of 30-s duration.

The *ISMIR* (http://ismir2004.ismir.net/ISMIR_Contest.html) comes from the ISMIR 2004 Genre classification contest and contains 1,458 full music recordings distributed over six genre classes as follows: classical (640), electronic (229), jazz-blues (52), metal-punk (90), rock-pop (203), and world (244), where the number within parentheses refers to the number of recordings which belong to each genre class. Therefore, 43.9% of the music recordings belong to the classical genre.

The *Homburg* (http://www-ai.cs.uni-dortmund.de/audio.html) contains 1,886 music excerpts of 10-s length by 1,463 different artists. These excerpts are unequally distributed over nine genres, namely alternative, blues, electronic, folk-country, funk/soul/RnB, jazz, pop, rap/hip-hop, and rock. The largest class is the rap/hip-hop genre containing 26.72% of the music excerpts, while the funk/soul/RnB is the smallest one containing 2.49% of the music excerpts.

The *1517-Artists* (http://www.seyerlehner.info/index.php?p=1_3_Download) consists of 3,180 full-length music recordings from 1,517 different artists, downloaded free from download.com. The 190 most popular songs, according to the number of total listenings, were selected for each of the 19 genres, i.e., alternative/punk, blues, children's, classical, comedy/spoken, country, easy listening/vocal, electronic, folk, hip-hop, jazz, latin, new age, RnB/soul, reggae, religious, rock/pop, soundtracks, and world. In this dataset, the music recordings are distributed almost uniformly over the genre classes.

The *Unique* (http://www.seyerlehner.info/index.php?p=1_3_Download) consist of 3,115 music excerpts of popular and well-known songs, distributed over 14 genres, namely blues, classic, country, dance, electronica, hip-hop, jazz, reggae, rock, schlager (i.e., music hits), soul/RnB, folk, world, and spoken. Each excerpt has 30-s duration. The class distribution is skewed. That is, the smallest class (i.e., spoken music) accounts for 0.83%, and the largest class (i.e., classic) for 24.59% of the available music excerpts.

The *MTV* (http://www.openaudio.eu/) contains 195 full-music recordings with a total duration of 14.2 h from the MTV Europe Most Wanted Top Ten of 20 years (1981 to 2000), covering a wide variety of popular music genres. The ground truth was obtained by five annotators (Rater A to Rater E, four males and one female), who were asked to make a forced binary decision according to the two dimensions in Thayer's mood plane [55] (i.e., assigning either $+1$ or -1 for arousal and valence, respectively) according their mood perception.

The *CAL500* (http://cosmal.ucsd.edu/cal/) is a corpus of 500 recordings of Western popular music, each of which has been manually annotated by at least three human annotators, who employ a vocabulary of 174 tags. The tags used in CAL500 dataset annotation span six semantic categories, namely instrumentation, vocal characteristics, genres, emotions, acoustic quality of the song, and usage terms (e.g., 'I would like to listen this song while driving') [1].

Each music recording in the aforementioned datasets was represented by three song-level feature vectors, namely the 20-dimensional MFCCs, the 12-dimensional chroma features, and the auditory cortical representations of reduced dimensions. The dimensionality of the cortical features was reduced via random projections as described in Section 4. In particular, the dimensions of the cortical features after random projections are 1,570 for the GTZAN, 1,391 for the ISMIR, 2,261 for the Homburg, 2,842 for the 1517-Artists, 2,868 for the Unique, 518 for the MTV, and 935 for the CAL500 dataset, respectively.

Two sets of experiments in music classification were conducted. First, to be able to compare the performance of the LRSMs with that of the state-of-the-art music classification methods, standard evaluation protocols were applied to the seven datasets. In particular, following [16,17,20,22,56,57], stratified 10-fold cross-validation was applied to the GTZAN dataset. According to [15,16,54], the same protocol was also applied to the Homburg, Unique, 1517-Artists, and MTV datasets. The experiments on the ISMIR 2004 Genre dataset were conducted according to the ISMIR 2004 Audio Description Contest protocol. The protocol defines training and evaluation sets, which consist of 729 audio files each. The experiments on music tagging were conducted following the experimental procedure defined in [26]. That is, 78 tags, which have been employed to annotate at least 50 music recordings in the CAL500 dataset, were used in the experiments by applying fivefold cross-validation.

Fu et al. [3] indicated that the main challenge for future music information retrieval systems is to be able to train the music classification systems for large-scale datasets from few labeled data. This situation is very common in practice since the number of annotated music recordings per class is often limited [3]. To this end, the performance of the LRSMs in music classification given a few training music recordings is investigated in the second set of experiments. In this small-sample size setting, only 10% of the available recordings were used as the training set and the remaining 90% for the test in all, but the CAL500, datasets. The experiments were repeated 10 times. In music tagging, 20% of the recordings in the CAL500 were used as the training set and the remaining 80% for the test. This experiment was repeated five times.

The LRSMs are compared against three well-known classifiers, namely the SRC [38], the linear SVMs[a], and the NN classifier with a cosine distance metric in music genre and mood classification, by applying the aforementioned experimental procedures. In music tagging, the LRSMs are compared against the multi-label variants of the aforementioned single-label classifiers, namely the MLSRC

Table 1 Music genre classification accuracies for the GTZAN, ISMIR, Homburg, 1517-Artists, and Unique datasets

Method	Features	GTZAN	ISMIR	Homburg	1517-Artists	Unique
LRSMs	Fusion cmc	*87.00* (2.62)	*82.99*	*62.40* (3.65)	*54.91* (2.54)	72.90 (1.26)
	Fusion cm	86.80 (2.85)	82.30	62.29 (4.04)	54.74 (2.68)	72.84 (1.11)
	Cortical	85.50 (2.79)	81.62	61.71 (4.02)	54.43 (2.58)	72.35 (1.05)
	MFCCs	50.60 (5.35)	59.08	43.26 (2.30)	23.45 (1.96)	54.60 (1.87)
	Chroma	17.6 (4.03)	43.90	26.93 (1.39)	9.77 (1.13)	24.71 (1.87)
SRC	Fusion cmc	84.40 (2.27)	82.85	59.64 (3.24)	53.08 (2.83)	72.61 (1.18)
	Fusion cm	84.40 (2.71)	80.50	58.10 (4.15)	50.78 (2.41)	71.97 (1.85)
	Cortical	84.10 (3.04)	79.97	57.52 (3.98)	50.72 (2.61)	67.48 (1.14)
	MFCCs	63.60 (5.01)	70.50	38.10 (2.74)	30.12 (1.87)	56.59 (1.06)
	Chroma	36.80 (5.67)	47.73	26.61 (2.58)	17.01 (1.31)	31.20 (2.94)
SVMs	Fusion cmc	86.80 (2.82)	*82.99*	62.61 (3.22)	53.30 (3.19)	*75.15* (1.48)
	Fusion cm	86.40 (2.98)	73.93	61.07 (3.32)	53.08 (3.38)	73.54 (1.87)
	Cortical	86.00 (2.83)	73.79	60.92 (2.83)	53.71 (3.18)	68.89 (2.22)
	MFCCs	54.90 (3.14)	52.67	43.95 (2.05)	26.16 (2.96)	53.22 (1.06)
	Chroma	16.90 (4.02)	48.42	34.99 (1.96)	12.16 (2.27)	39.87 (2.67)
NN	Fusion cmc	81.40 (3.20)	78.64	50.26 (4.21)	44.87 (2.21)	64.68 (2.31)
	Fusion cm	81.10 (3.31)	79.02	50.21 (3.48)	44.90 (2.43)	64.68 (2.31)
	Cortical	80.70 (3.26)	79.69	49.78 (2.98)	44.84 (2.55)	64.43 (2.57)
	MFCCs	57.60 (5.05)	67.76	29.79 (3.13)	26.57 (1.84)	48.82 (2.17)
	Chroma	34.10 (4.67)	42.24	23.64 (1.93)	14.40 (1.80)	25.32 (2.96)
		[20] 90.60	[20] 86.83	[16] 61.20	[16] 41.10	[16] 72.00
		[22] 84.30	[10] 83.50	[60] 57.81	[61] 35.00	
		[56] 82.50	[22] 83.15	[61] 55.30		
		[16] 82.00	[62] 82.30	[54] 53.23		
		[57] 77.20				

The numbers within the parentheses indicate the standard deviations obtained by 10-fold cross-validation. The best results are indicated in italics.

Table 2 Music mood classification accuracies for the MTV dataset

Method	Features	Rater A		Rater B		Rater C		Rater D		Rater E		Overall	
		Arousal	**Valence**	**Arousal**	**Valence**	**Arousal**	**Valence**	**Arousal**	**Valence**	**Arousal**	**Valence**	**Arousal**	**Valence**
LRSMs	Fusion cmc	76.42 (10.79)	66.02 (10.45)	81.55 (7.73)	63.57 (8.42)	70.21 (10.48)	57.28 (12.01)	66.71 (5.72)	69.21 (7.97)	70.26 (8.92)	61.55 (4.72)	*73.03* (1.97)	*63.53* (2.61)
	Fusion cm	76.42 (10.79)	66.02 (10.45)	81.55 (7.73)	63.57 (8.41)	70.21 (10.48)	57.28 (12.01)	66.71 (5.72)	69.21 (7.97)	70.26 (8.92)	61.55 (4.72)	*73.03* (1.97)	*63.53* (2.61)
	Cortical	76.42 (10.79)	66.52 (10.77)	81.52 (7.27)	63.57 (8.41)	69.68 (10.12)	57.28 (12.01)	66.18 (4.61)	68.15 (9.09)	70.26 (8.92)	61.55 (4.72)	72.72 (2.34)	63.42 (2.63)
	MFCCs	69.76 (8.51)	62.15 (7.39)	68.15 (8.84)	66.07 (10.65)	64.21 (13.52)	60.92 (11.43)	64.57 (6.38)	57.52 (16.75)	63.65 (6.09)	54.84 (13.72)	66.07 (2.82)	60.30 (3.13)
	Chroma	59.55 (11.61)	62.15 (7.39)	58.39 (8.45)	66.07 (10.65)	53.84 (16.36)	60.92 (11.43)	57.55 (12.58)	55.47 (15.96)	59.02 (6.87)	55.34 (14.22)	57.67 (3.51)	59.99 (3.14)
SRC	Fusion cmc	77.39 (12.01)	59.97 (7.74)	79.92 (7.67)	64.00 (10.62)	70.28 (10.40)	58.92 (12.11)	64.15 (7.39)	65.02 (12.59)	63.00 (10.25)	56.36 (8.89)	70.95 (1.86)	60.85 (1.96)
	Fusion cm	74.78 (11.37)	59.97 (8.17)	77.28 (9.49)	63.00 (8.26)	69.28 (10.88)	62.50 (12.70)	65.65 (7.22)	62.89 (13.73)	62.00 (10.81)	54.23 (9.72)	69.08 (1.59)	60.52 (2.43)
	Cortical	75.36 (9.91)	60.52 (6.35)	77.81 (8.37)	65.02 (10.26)	68.73 (11.63)	58.94 (10.14)	62.10 (6.03)	65.55 (12.14)	63.53 (11.39)	57.86 (7.62)	69.51 (2.16)	61.58 (2.18)
	MFCCs	53.81 (11.11)	59.55 (10.24)	60.97 (7.68)	62.47 (12.71)	64.15 (13.71)	57.94 (9.30)	53.86 (7.80)	58.36 (8.60)	63.13 (6.32)	46.13 (13.43)	59.18 (2.86)	56.84 (2.00)
	Chroma	54.63 (14.75)	53.31 (7.50)	53.73 (11.93)	61.65 (11.65)	50.26 (9.08)	48.60 (9.71)	55.97 (9.81)	56.00 (9.22)	53.91 (9.22)	48.55 (12.05)	53.70 (2.28)	53.62 (1.76)
SVMs	Fusion cmc	76.39 (12.71)	61.52 (10.01)	78.97 (7.04)	66.07 (10.65)	68.23 (9.90)	56.28 (13.23)	61.05 (8.19)	65.02 (9.55)	69.84 (12.41)	58.42 (13.71)	70.90 (2.37)	61.46 (1.81)
	Fusion cm	73.36 (12.59)	61.52 (10.01)	78.47 (6.74)	66.07 (10.65)	67.73 (9.71)	56.26 (13.54)	61.05 (8.19)	65.02 (9.55)	70.34 (13.17)	56.86 (12.52)	70.59 (2.62)	61.15 (1.61)
	Cortical	75.36 (12.59)	61.00 (10.27)	78.97 (7.04)	66.07 (10.65)	68.26 (9.25)	56.26 (13.54)	61.07 (7.33)	64.02 (8.24)	69.84 (14.09)	56.86 (14.14)	70.70 (2.99)	60.84 (2.31)
	MFCCs	57.07 (15.30)	62.15 (7.39)	57.47 (13.08)	66.07 (10.65)	56.42 (10.93)	60.92 (11.42)	57.63 (10.93)	55.47 (15.96)	46.57 (11.54)	55.34 (14.22)	55.03 (2.41)	59.99 (3.14)
	Chroma	57.07 (15.30)	62.15 (7.39)	57.47 (13.08)	66.07 (10.65)	56.42 (10.93)	60.92 (11.42)	57.63 (10.93)	55.47 (15.96)	46.57 (11.54)	55.34 (14.22)	55.03 (2.41)	59.99 (3.14)
NN	Fusion cmc	65.63 (5.86)	51.28 (8.03)	66.57 (11.73)	57.31 (7.92)	58.55 (11.15)	57.84 (9.27)	61.55 (6.43)	54.73 (13.50)	63.52 (9.98)	52.81 (6.72)	63.16 (2.57)	54.80 (2.49)
	Fusion cm	66.68 (6.35)	52.34 (8.11)	68.15 (8.74)	60.39 (9.84)	59.57 (10.87)	57.36 (9.20)	62.05 (6.90)	55.26 (9.72)	61 (10.03)	52.23 (8.07)	63.49 (1.84)	55.52 (0.81)

Table 2 Music mood classification accuracies for the MTV dataset *continued*

	Cortical	64.60 (7.07)	53.42 (9.29)	69.71 (10.16)	62.47 (7.69)	60.60 (9.96)	58.47 (7.39)	61.00 (7.03)	56.68 (12.02)	62.60 (10.30)	50.21 (7.97)	63.60 (1.60)	56.25 (1.82)
	MFCCs	58.97 (9.24)	57.31 (10.20)	59.39 (6.70)	55.86 (11.66)	57.52 (15.61)	54.31 (8.52)	52.97 (14.82)	57.84 (11.71)	62.47 (10.14)	47.63 (11.52)	58.26 (3.60)	54.59 (1.30)
	Chroma	56.65 (15.01)	49.21 (10.93)	54.78 (14.22)	57.02 (11.39)	47.73 (9.08)	49.73 (5.72)	54.92 (8.53)	55.42 (9.81)	53.39 (9.55)	46.60 (11.32)	53.50 (2.68)	51.6 (2.26)
[15]		71.80	62.10	71.30	70.80	74.40	63.10	66.70	68.70	69.90	60.50	71.80	60.50

The numbers within the parentheses indicate the standard deviations obtained by 10-fold cross-validation. The best results are indicated in italics.

[39], the Rank-SVMs [40], the MLkNN [41], as well as the well-performing PARAFAC2-based autotagging method [42]. The number of neighbors used in the MLkNN was set to 15. The sparse coefficients in the SRC and MLSRC are estimated by the LASSO[b] [58].

The performance in music genre and mood classification is assessed by reporting the classification accuracy. Three metrics, namely the mean per-tag precision, the mean per-tag recall, and the F_1 score, are used in order to assess the performance of autotagging. These metrics are defined as follows [1]: Per-tag precision is defined as the fraction of music recordings annotated by any method with label w that are actually labeled with tag w. Per-tag recall is defined as the fraction of music recordings actually labeled with tag w that the method annotates with label w. The F_1 score is the harmonic mean of precision and recall. That is, $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$ yields a scalar measure of overall annotation performance. If a tag is never selected for annotation, then following [1,26], the corresponding precision (that otherwise would be undefined) is set to the tag prior to the training set, which equals the performance of a random classifier. In the music tagging experiments, the length of the class indicator vector returned by the LRSMs as well as the MLSRC, the Rank-SVMs, the MLkNN, and the PARAFAC2-based autotagging method was set to 10 as in [1,26]. That is, each test music recording is annotated with 10 tags. The parameters in the LRSMs have been estimated by

Table 3 Music tagging performance on the CAL500 dataset by applying fivefold cross-validation

Method	Features	Precision	Recall	F_1 score
LRSMs	Fusion cmc	*0.504*	0.202	0.289
	Fusion cm	*0.504*	0.202	0.289
	Cortical	0.500	0.203	0.288
	MFCCs	0.379	0.167	0.232
	Chroma	0.277	0.130	0.174
MLSRC	Fusion cmc	0.458	0.184	0.263
	Fusion cm	0.466	0.186	0.266
	Cortical	0.476	0.186	0.267
	MFCCs	0.384	0.162	0.228
	Chroma	0.355	0.151	0.212
Rank-SVMs	Fusion cmc	0.469	*0.209*	0.289
	Fusion cm	0.477	*0.209*	*0.291*
	Cortical	0.472	0.208	0.288
	MFCCs	0.313	0.140	0.194
	Chroma	0.287	0.128	0.177
MLkNN	Fusion cmc	0.404	0.173	0.243
	Fusion cm	0.398	0.173	0.241
	Cortical	0.404	0.173	0.242
	MFCCs	0.338	0.156	0.213
	Chroma	0.305	0.138	0.190
PARAFAC2	Fusion cmc	0.483	0.196	0.279
	Fusion cm	0.480	0.197	0.279
	Cortical	0.472	0.208	0.288
	MFCCs	0.304	0.136	0.188
	Chroma	0.289	0.130	0.179
[63]		0.480	0.260	0.340
[26]		0.490	0.230	0.260
[9] as evaluated in [26]		0.410	0.240	0.250
[12] as evaluated in [26]		0.380	0.240	0.250
[64] as evaluated in [26]		0.370	0.170	0.200

The best results are indicated in italics.

Table 4 Music classification results on various datasets obtained by employing a few labeled music recordings

Method	Features	Music Genre					Music Mood: MTV	
		GTZAN	ISMIR	Homburg	1517-Artists	Unique	Arousal	Valence
LRSMs	Fusion cmc	*72.02* (1.97)	*74.05* (1.69)	*57.02* (0.44)	*41.03* (1.45)	68.68 (0.52)	*62.76* (1.57)	56.28 (1.18)
	Fusion cm	71.73 (1.79)	73.28 (1.47)	*57.02* (0.44)	*41.03* (1.45)	68.55 (0.66)	*62.76* (1.57)	57.74 (1.42)
SRC	Fusion cmc	69.91 (1.57)	72.36 (1.69)	53.63 (0.62)	35.92 (0.85)	67.83 (1.11)	60.29 (1.68)	58.19 (2.00)
	Fusion cm	68.33 (2.21)	72.33 (1.30)	53.48 (0.83)	35.82 (0.89)	67.70 (1.19)	60.46 (1.40)	*58.25* (2.33)
SVMs	Fusion cmc	70.84 (1.41)	72.55 (1.43)	55.39 (0.86)	37.83 (1.10)	*69.85* (0.68)	55.24 (0.88)	57.81 (1.17)
	Fusion cm	70.33 (1.53)	64.78 (0.86)	55.43 (0.78)	37.60 (1.00)	69.46 (0.66)	55.12 (0.87)	57.88 (1.79)
NN	Fusion cmc	63.01 (3.18)	71.07 (1.34)	45.02 (1.19)	29.95 (1.52)	59.49 (1.49)	59.14 (1.77)	55.43 (1.21)
	Fusion cm	61.63 (3.11)	70.89 (1.10)	44.84 (1.34)	29.76 (1.42)	59.41 (1.47)	59.10 (1.47)	55.42 (1.59)

The best results are indicated in italics.

employing the method in [59]. That is, for each training set, a validation set (disjoint from the test set) was randomly selected and used next for tuning the parameters (i.e., λ_r, $r = 1, 2, \ldots, R$).

5.2 Experimental results

In Tables 1, 2, and 3, the experimental results in music genre, mood, and multi-label classification are summarized, respectively. These results have been obtained by applying the standard protocol defined for each dataset. In Tables 4 and 5, music classification results are reported, when a small training set is employed. Each classifier is applied to the auditory cortical representations (cortical features) of reduced dimensions, the 20-dimensional MFCCs, the 12-dimensional chroma features, the linear combination of cortical features and MFCCs (fusion cm, i.e., $R = 2$), and the linear combination of all the aforementioned features (fusion cmc, i.e., $R = 3$). Apart from the proposed LRSMs, the other competitive classifiers handle the fusion of multiple audio features in an *ad hoc* manner. That is, an augmented feature vector is constructed by stacking the cortical features on the top of the 20-dimensional MFCCs and the 12-dimensional chroma features. In the last rows of Tables 1, 2, and 3, the figures of merit for the top performing music classification methods are included for comparison purposes.

By inspecting Table 1, the best music genre classification accuracy has been obtained by the LRSMs in four out five datasets, when all the features have been exploited for music representation. Comparable performance has been achieved by the combination of cortical features and the MFCCs. This is not the case for the Unique dataset, where the SVMs achieve the best classification accuracy when employing the fusion of the cortical features, the MFCCs, and the chroma features. Furthermore, the LRSMs outperform all the classifiers being compared to when they are applied to cortical features. The MFCCs are classified more accurately by the SRC or the SVMs than the LRSMs. This is because the MFCCs and the chroma features have a low dimensionality and the LRSMs are not able to extract the appropriate common latent features the genre classes are built on. The best classification accuracy obtained by

Table 5 Music classification results on the CAL500 dataset obtained by employing a few labeled music recordings

Method	Features	Music Tagging: CAL500		
		Precision	Recall	F_1 score
LRSMs	Fusion cmc	*0.480*	*0.191*	*0.273*
	Fusion cm	*0.480*	*0.191*	*0.273*
MLSRC	Fusion cmc	0.467	0.178	0.257
	Fusion cm	0.467	0.178	0.257
Rank-SVMs	Fusion cmc	0.433	0.181	0.255
	Fusion cm	0.435	0.182	0.257
MLkNN	Fusion cmc	0.331	0.151	0.208
	Fusion cm	0.331	0.149	0.205
PARAFAC2	Fusion cmc	0.460	0.187	0.266
	Fusion cm	0.462	0.188	0.267

The best results are indicated in italics.

the LRSMs on all datasets ranks high compared to that obtained by the majority of music genre classification techniques, as listed in last rows of Table 1. In particular, for the Homburg, 1517-Artists, and Unique datasets, the best accuracy achieved by the LRSMs outperforms that obtained by the state-of-the-art music classification methods. Regarding to the GTZAN and ISMIR datasets, it is worth mentioning that the results reported in [20] have been obtained by applying feature aggregation on the combination of four elaborated audio features.

Schuller et al. argued that the two dimensions in Thayer's mood model, namely the arousal and the valence, are independent of each other [15]. Therefore, mood classification can be reasonably done independently in each dimension, as presented in Table 2. That is, each classifier makes binary decisions between excitation and calmness on the arousal scale as well as negativity and positivity in the valence dimension, respectively. Both overall and per-rater music mood classification accuracies are reported. The overall accuracies are the mean accuracies over all raters for all songs in the dataset. The LRSMs outperform the classifiers that are compared to when the cortical features and their fusion with the MFCCs and the chroma features are employed for music representation, yielding higher classification accuracies than those reported in the row entry NONLYR in Tables twelve and thirteen [15] when only audio features are employed. It is seen that the inclusion of the chroma features does not alter the measured figures of merit. Accordingly, the chroma features could be omitted without any performance deterioration. It is worth mentioning that substantial improvements in the classification accuracy are reported when audio features are combined with lyric features [15]. The overall accuracy achieved by the LRSMs in valence and arousal is considered satisfactory, considering the inherent ambiguity in the mood assignments and the realistic nature of the MTV dataset.

The results reported in Table 3 indicate that in music tagging, the LRSMs outperform the MLSRC, the MLkNN, and the PARAFAC2 with respect to per-tag precision, per-tag recall, and F_1 score for all the music representations employed. The Rank-SVMs yield the best tagging performance with respect to the F_1 score and the recall. The cortical features seem to be more appropriate for music annotation than the MFCCs, no matter which annotation method is employed. Although the LRSMs achieve top performance against the state-of-the-art methods with respect to per-tag precision, the reported recall is much smaller compared to that published for the majority of music tagging methods (last five rows in Table 3). This result is due to the song-level features employed here, which fail to capture the temporal information with some tags (e.g., instrumentation). In contrast, the well-performing autotagging method with respect to recall, which is reported in Table 3, employs sequences of audio features for music representation.

In Tables 4 and 5, music classification results, by applying a small-sample size setting, are summarized. These results have been obtained by employing either the fusion of the cortical features, the MFCCs, and the chroma features or the fusion of the former two audio representations. Clearly, the LRSMs outperform all the classifiers they are compared to in most music classification tasks. The only exceptions are the prediction of valence on the MTV dataset, where the best classification accuracy is achieved by the SRC, and the music genre classification accuracy on the Unique dataset, where the top performance is achieved by the SVMs. Given the relatively small number of training music recordings, the results in Tables 4 and 5 are quite acceptable, indicating that the LRSMs are an appealing method for music classification in real-world conditions.

6 Conclusions

The LRSMs have been proposed as a general-purpose music classification method. Given a number of music representations, the LRSMs are able to extract the appropriate features for each specific music classification task, yielding higher performance than the methods they are compared to. Furthermore, the best classification results obtained by the LRSMs either meet or slightly outperform those obtained by the state-of-the-art methods for music genre, mood, and multi-label music classification. The superiority of the auditory cortical representations has been demonstrated over the conventional MFCCs and chroma features in the three music classification tasks studied as well. Finally, the LRSMs yield high music classification performance when a small number of training recordings is employed. This result highlights the potential of the proposed method for practical music information retrieval systems.

Endnotes

[a] The LIBSVM was used in the experiments (http://www.csie.ntu.edu.tw/~cjlin/libsvm/).

[b] The SPGL1 Matlab solver was used in the implementation of the SRC and the MLSRC (http://www.cs.ubc.ca/~mpf/spgl1/).

Abbreviations

ADALM: Alternating direction augmented Lagrange multiplier; BOF: Bag-of-features; CQT: Constant-Q transform; LRSMs: Low-rank semantic mappings; MFCCs: Mel-frequency cepstral coefficients; MLkNN: Multi-label k-nearest neighbor; MLSRC: Multi-label sparse representation-based classifier; NN: Nearest neighbor; PARAFAC2: Parallel factor analysis 2; SVMs: Support vector machines; 2D: Two-dimensional; 4D: Four-dimensional.

Competing interests

Both authors declare that they have no competing interests.

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.

Received: 26 October 2012 Accepted: 22 May 2013
Published: 24 June 2013

References

1. D Turnbull, L Barrington, D Torres, G Lanckriet, Semantic annotation and retrieval of music and sound effects. IEEE Trans. Audio, Speech, Lang. Proc. **16**(2), 467–476 (2008)
2. T Bertin-Mahieux, D Eck, M Mandel, in *Machine Audition: Principles, Algorithms and Systems*, ed. by W Wang. Automatic tagging of audio: the state-of-the-art (IGI Hershey, 2010), pp. 334–352
3. Z Fu, G Lu, KM Ting, D Zhang, A survey of audio-based music classification and annotation. IEEE Trans. Multimedia. **13**(2), 303–319 (2011)
4. YE Kim, EM Schmidt, R Migneco, BG Morton, P Richardson, J Scott, JA Speck, D Turnbull, in *Proc. 11th Int. Conf. Music Information Retrieval*. Music emotion recognition: a state of the art review (Utrecht, 9–13 Aug 2010), pp. 255–266
5. N Scaringella, G Zoia, D Mlynek, Automatic genre classification of music content: a survey. IEEE Signal Proc. Mag. **23**(2), 133–141 (2006)
6. J Bergstra, M Mandel, D Eck, in *Proc. 11th Int. Conf. Music Inform. Retrieval*. Scalable genre and tag prediction with spectral covariance (Utrecht, 9–13 Aug 2010), pp. 507–512
7. L Chen, P Wright, W Nejdl, in *Proc. ACM 2nd Int. Conf. Web Search and Data Mining*. Improving music genre classification using collaborative tagging data (ACM Barcelona, 9–12 Feb 2009), pp. 84–93
8. K Chang, JSR Jang, CS Iliopoulos, in *Proc. 11th Int. Conf. Music Information Retrieval*. Music genre classification via compressive sampling (Utrecht, 9–13 Aug 2010), pp. 387–392
9. M Hoffman, D Blei, P Cook, in *Proc. 10th Int. Conf. Music Information Retrieval*. Easy as CBA: a simple probabilistic model for tagging music (Kobe, 26–30 Oct 2009), pp. 369–374
10. A Holzapfel, Y Stylianou, Musical genre classification using nonnegative matrix factorization-based features. IEEE Trans. Audio, Speech, Lang. Proc. **16**(2), 424–434 (2008)
11. H Lukashevich, J Abeber, C Dittmar, H Grossman, in *Proc. 10th Int. Conf. Music Information Retrieval*. From multi-labeling to multi-domain-laneling: a novel two-dimensional approach to music genre classification (Kobe, 26–30 Oct 2009), pp. 459–464
12. MI Mandel, DPW Ellis, in *Proc. 9th Int. Conf. Music Information Retrieval*. Multiple-instance learning for music information retrieval (Philadelphia, 14–18 Sept 2008), pp. 577–582
13. R Miotto, L Barrington, G Lanckriet, in *Proc. 11th Int. Conf. Music Information Retrieval*. Improving auto-tagging by modeling semantic co-occurrences (Utrecht, 9–13 Aug 2010), pp. 297–302
14. SR Ness, A Theocharis, G Tzanetakis, LG Martins, in *Proc. 17th ACM Int. Conf. Multimedia*. Improving automatic music tag annotation using stacked generalization of probabilistic SVM outputs (Beijing, 19–22 Oct 2009), pp. 705–708
15. B Schuller, C Hage, D Schuller, G Rigoll, "Mister D.J., Cheer Me Up!": musical and textual features for automatic mood classification. J. New Music Res. **39**, 13–34 (2010)
16. K Seyerlehner, G Widmer, T Pohle, P Knees, in *Proc. 13th Int. Conf. Digital Audio Effects*. Fusing block-level features for music similarity estimation (Graz, 6–10 Sept 2010), pp. 528–531
17. G Tzanetakis, P Cook, Musical genre classification of audio signals. IEEE Trans. Speech Audio Proc. **10**(5), 293–302 (2002)
18. C Zhen, J Xu, in *Proc. 3rd IEEE Int. Conf. Computer Science and Information Technology*. Multi-modal music genre classification approach (Chengdu, 9–11 July 2010), pp. 398–402
19. D Garcia-Garcia, J Arenas-Garcia, E Parrado-Hernandez, F Diaz-de Maria, in *Proc. IEEE 20th Int. Workshop Machine Learning for Signal Processing*. Music genre classification using the temporal structure of songs (Kittila, 29 Aug–1 Sept 2010), pp. 266–271
20. CH Lee, JL Shih, KM Yu, HS Lin, Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. IEEE Trans. Multimedia. **11**(4), 670–682 (2009)
21. A Nagathil, T Gerkmann, R Martin, in *Proc. 18th European Signal Processing Conf.* Musical genre classification based on a highly-resolved cepstral modulation spectrum (Aalborg, 23–27 Aug 2010), pp. 462–466
22. Y Panagakis, C Kotropoulos, GR Arce, Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification. IEEE Trans. Audio, Speech, Lang Tech. **18**(3), 576–588 (2010)
23. M Mandel, DPW Ellis, in *Proc. 6th Int. Conf. Music Information Retrieval*. Song-level features support vector machines for music classification (London, 11–15 Sept 2005), pp. 594–599
24. Y Panagakis, C Kotropoulos, in *Proc. 19th European Signal Processing Conf.* Automatic music mood classification via low-rank representation (Barcelona, 29 Aug–2 Sept 2011), pp. 689–693
25. T Bertin-Mahieux, D Eck, F Maillet, P Lamere, Autotagger: a model for predicting social tags from acoustic features on large music databases. J. New Music Res. **37**(2), 115–135 (2008)
26. E Coviello, A Chan, G Lanckriet, Time series models for semantic music annotation. IEEE Trans. Audio, Speech, Lang Proc. **19**(5), 1343 –1359 (2011)
27. N Srebro, JDM Rennie, T Jaakkola, in *2004 Advances in Neural Information Processing Systems*. Maximum-margin matrix factorization (Vancouver, 13–18 Dec 2004), pp. 1329–1336
28. DP Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, 2nd edn. (Athena Scientific, Belmont, 1996)
29. R Ando, R Kubota, T Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data. J. Mach. Learn. Res. **6**, 1817–1853 (2005)
30. A Torralba, KP Murphy, WT Freeman, Sharing visual features for multiclass and multiview object detection. IEEE Trans. Pattern Anal Mach. Intel. **29**(5), 854–869 (2007)
31. Y Amit, M Fink, N Srebro, S Ullman, in *Proc. 24th Int. Conf. on Machine Learning*. Uncovering shared structures in multiclass classification (Corvallis, 20–24 June 2007), pp. 17–24
32. JDM Rennie, N Srebro, in *Proc. 2005 Int. Conf. Machine Learning*. Fast maximum margin matrix factorization for collaborative prediction (Los Angeles, 15–17 Dec 2005), pp. 713–719
33. S Ji, L Tang, S Yu, J Ye, in *Proc. 2005 Int. Conf. Machine Learning*. Extracting shared subspace for multi-label classification (Corvallis, 20–24 June 2007)
34. X Yang, K Wang, SA Shamma, Auditory representations of acoustic signals. IEEE Trans. Inf. Theory. **38**(2), 824–839 (1992)
35. N Mesgarani, M Slaney, SA Shamma, Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE Trans. Audio, Speech, Lang Proc. **14**(3), 920–930 (2006)
36. B Logan, in *Proc. 1st Int. Symposium Music Information Retrieval*. Mel frequency cepstral coefficients for music modeling (Plymouth, 23–25 Oct 2000)
37. M Ryynanen, A Klapuri, Automatic transcription of melody, bass line, and chords in polyphonic music. Comput Music J. **32**(3), 72–86 (2008)
38. J Wright, A Yang, A Ganesh, S Sastry, Y Ma, Robust face recognition via sparse representation. IEEE Trans. Pattern Anal Mach. Int. **31**(2), 210–227 (2009)
39. T Sakai, H Itoh, A Imiya, in *Proc. Computer Vision-ACCV 2010 Workshops*. Multi-label classification for image annotation via sparse similarity voting (Springer Queenstown, 8–12 Nov 2010), pp. 344–353
40. A Elisseeff, JA Weston, in *14th Advances in Neural Information Processing Systems*. A kernel method for multi-labelled classification (MIT Cambridge, 2002), pp. 681–687
41. ML Zhang, ZH Zhou, ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn. **40**(7), 2038–2048 (2007)
42. Y Panagakis, C Kotropoulos, in *Proc. 2011 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*. Automatic music tagging via PARAFAC2 (Prague, 22–27 May 2011), pp. 481–484
43. R Munkong, Biing-J Hwang, Auditory perception and cognition. IEEE Signal Proc Mag. **25**(3), 98–117 (2008)
44. C Schoerkhuber, A Klapuri, in *Proc. 7th Sound and Music Computing Conf.* Constant-Q transform toolbox for music processing (Barcelona, 21–24 July 2010)
45. R Memisevic, in *Proc. 2012 Int. Conf. Machine Learning*. On multi-view feature learning (Edinburgh, 26 June)

46. GM Fung, OL Mangasarian, Multicategory proximal support vector machine classifiers. Mach Learn. **59**(1-2), 77–97 (2005)
47. H Zou, J Zhu, T Hastie, New multicategory boosting algorithms based on multicategory Fisher-consistent losses. Ann Appl. Stat. **2**(4), 1290–1306 (2008)
48. M Fazel, *Matrix rank minimization with applications*. (Department of Electrical Engineering, Stanford University, PhD thesis, 2002)
49. S Boyd, N Parikh, E Chu, B Peleato, J Eckstein, Distributed optimization statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**, 1–122 (2011)
50. JF Cai, EJ Candes, Z Shen, A singular value thresholding algorithm for matrix completion. SIAM J. Optimization. **2**(2), 569–592 (2009)
51. EJ Candes, X Li, Y Ma, J Wright, Robust principal component analysis? ACM J. **58**(3), 1–37 (2011)
52. DL Donoho, J Tanner, Counting faces of randomly projected polytopes when the projection radically lowers dimension. J. Am. Math. Soc. **22**, 1–53 (2009)
53. R Baraniuk, V Cevher, M WakinM, Low-dimensional models for dimensionality reduction and signal recovery: a geometric perspective. Proc IEEE. **98**(6), 959–971 (2010)
54. H Homburg, I Mierswa, B Moller, K Morik, M Wurst, in *Proc. 6th Int. Conf. Music Information Retrieval*. A benchmark dataset for audio classification and clustering (London, 11–15 Sept 2005), pp. 528–531
55. RE Thayer, *The Biopsychology of Mood and Arousal*. (Oxford University Press, Boston, 1989)
56. J Bergstra, N Casagrande, D Erhan, D Eck, B Kegl, Aggregate features and ADABOOST for music classification. Mach Learn. **65**(2-3), 473–484 (2006)
57. E Tsunoo, G Tzanetakis, N Ono, S Sagayama, Beyond timbral statistics: improving music classification using percussive patterns and bass lines. IEEE Trans. Audio, Speech, Lang. Proc. **19**(4), 1003–1014 (2011)
58. R Tibshirani, Regression shrinkage selection via the LASSO. J. R. Statist. Soc B. **58**, 267–288 (1996)
59. R Tibshirani, On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min. Knowl. Discov. **1**(3), 317–328 (1997)
60. K Aryafar, A Shokoufandeh, in *Proc. 1st ACM Int. Workshop Music Information Retrieval with User-Centered and Multimodal Strategies*. Music genre classification using explicit semantic analysis (Scottsdale, 28 Nov–1 Dec 2011), pp. 33–38
61. C Osendorfer, J Schluter, J Schmidhuber, P van der Smagt, in *Proc. 28th Int. Conf. Machine Learning*. Unsupervised learning of low-level audio features for music similarity estimation (Bellevue, 28 June–2 July 2011)
62. E Pampalk, A Flexer, G Widmer, in *Proc. 6th Int. Conf. Music Information Retrieval*. Improvements of audio-based music similarity and genre classification (London, 11–15 Sept 2005), pp. 628–633
63. B Xie, W Bian, D Tao, P Chordia, in *Proc. 12th Int. Conf. Music Information Retrieval*. Music tagging with regularized logistic regression (Miami, 24–28 Oct 2011), pp. 711–716
64. D Eck, P Lamere, Bertin-T Mahieux, S Green, *Automatic generation of social tags for music recommendation* (MIT, Cambridge, 2007), pp. 385–392

doi:10.1186/1687-4722-2013-13
Cite this article as: Panagakis and Kotropoulos: **Music classification by low-rank semantic mappings.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:13.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com