Ensemble Discriminant Sparse Projections Applied to Music Genre Classification

Constantine Kotropoulos[†], Gonzalo R. Arce[‡], and Yannis Panagakis[†]
[†] *Department of Informatics, Aristotle University of Thessaloniki*
Box 451, Thessaloniki 541 24, GREECE
costas@aiia.csd.auth.gr, yannisp@csd.auth.gr
[‡] *Department of Electrical & Computer Engineering, University of Delaware*
Newark, DE 19716-3130, USA
arce@ece.udel.edu

Abstract

Resorting to the rich, psycho-physiologically grounded, properties of the slow temporal modulations of music recordings, a novel classifier ensemble is built, which applies discriminant sparse projections. More specifically, overcomplete dictionaries are learned and sparse coefficient vectors are extracted to optimally approximate the slow temporal modulations of the training music recordings. The sparse coefficient vectors are then projected to the principal subspaces of their within-class and between-class covariance matrices. Decisions are taken with respect to the minimum Euclidean distance from the class mean sparse coefficient vectors, which undergo the aforementioned projections. The application of majority voting to the decisions taken by 10 individual classifiers, which are trained on the 10 training folds defined by stratified 10-fold cross-validation on the GTZAN dataset, yields a music genre classification accuracy of 84.96% on average. The latter exceeds by 2.46% the highest accuracy previously reported without employing any sparse representations.

1 Introduction

Despite the lack of a commonly agreed definition of music genre, music genre is probably the most popular description of music content [13]. The majority of music genre classification algorithms model the music signal by the long-term statistical distribution of its short-time features. Commonly used feature sets represent either timbral texture, rhythmic content, pitch content, or combinations of the aforementioned characteristics [16]. The slow temporal modulations have appealing properties from the human perceptual point of view. This has motivated us to employ the auditory model proposed in [19] in order to map a given music recording to a two-dimensional (2D) representation of its slow temporal modulations.

Recently, the interest on sparse representations of signals has revived [4,5]. Resorting to the strong theoretical foundations of sparse representations, given a set of training auditory temporal modulations, the dictionary, that best represents each member of the training set under sparsity constraints, is extracted by means of the K-SVD algorithm [1]. To develop a supervised music genre classifier, the most discriminating features (MDF) [15] should be extracted by using the Fisher's multiclass linear discriminant analysis (LDA). In particular, the sparse coefficient vectors are projected to the principal subspaces of their within-class and between-class covariance matrices by applying dual LDA [17]. Decisions are then taken with respect to the minimum Euclidean distance from the class mean sparse coefficient vectors, which undergo the same projections. The cascade of sparse representation and LDA is termed *discriminant sparse projection*. The assessment of the proposed music genre classifier is conducted on the Tzanetakis' database (referred to as GTZAN dataset) [16]. However instead of implementing random discriminant analysis to boost performance, majority voting is applied to the decisions taken by 10 dual LDA classifiers trained on the 10 training subsets, defined by stratified 10-fold cross-validation, and applied to each

test subset. The proposed classifier ensemble yields an accuracy of 84.96% on average. The aforementioned genre classification accuracy is 2.46% higher than that reported in [3].

The novel contributions of the paper are in the formulation of the dual LDA of the auditory modulation representations as dual LDA applied to the coefficients of their sparse representations and the demonstration of the classification capability of the majority voting scheme that employs multiple dual LDA classifiers applied to the coefficients of the sparse auditory modulation representations.

2 Auditory Temporal Modulations of Music Signals and Their Sparse Approximation

Next, we briefly describe how a 2D representation of auditory temporal modulations can be obtained. Such a representation is a joint acoustic and modulation frequency representation [14], which discards much of the spectro-temporal details and focuses on the underlying slow temporal modulations of the music signal. In this paper, the mathematical model of Yang *et. al* [19] is adopted. Psychophysiological evidence justifies the choice of $r \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ (Hz) to represent the temporal modulation content of sound. The cochlear model, employed in the first stage, has 96 filters covering 4 octaves along the tonotopic axis (i.e. 24 filters per octave). Accordingly, the auditory temporal modulations of a set of music recordings (i.e. a dataset) are naturally represented by a third-order nonnegative real-valued tensor $\mathcal{Y} \in \mathbb{R}_+^{N_r \times N_f \times N_s}$, where $N_f = 96$, $N_r = 8$, and N_s denotes the number of music recordings. Let $\mathbf{Y}_{(3)} \in \mathbb{R}_+^{N_s \times (N_f \cdot N_r)}$ be the 3rd mode matrix unfolding of the aforementioned tensor. Obviously, $\mathbf{Y} = \mathbf{Y}_{(3)}^T = \left[\mathbf{y}_1 | \mathbf{y}_2 | \cdots | \mathbf{y}_{N_s}\right]$, where T denotes matrix transposition, is the data matrix. Each column $\mathbf{y}_j$, $j = 1, 2, \ldots, N_s$, is the lexicographically ordered vectorial representation of the 2D auditory temporal modulation of every sample in the dataset having originally size $N_f \cdot N_r = 768$, which is downsampled with ratio 1/8, 1/4, 1/3, and 1/2 in the frequency-rate 2D domain yielding finally a vector of size $M \in \{12, 48, 85, 192\}$, respectively.

A downsampled representation of auditory temporal modulations $\mathbf{y}_j \in \mathbb{R}_+^M$, $j = 1, 2, \ldots, N_s$, called sample hereafter, admits a sparse approximation over a dictionary $\mathbf{D} \in \mathbb{R}^{M \times K}$ (whose columns are referred to as atoms), when $\mathbf{y}_j$ can be approximated either exactly or closely as a linear combination of a few only atoms of $\mathbf{D}$, i.e. $\mathbf{y}_j = \mathbf{D}\,\mathbf{x}_j$ or $||\mathbf{y}_j - \mathbf{D}\mathbf{x}_j||_p \leq \gamma$, where $||\,||_p$ denotes the ℓ_p vector norm for $p = 1, 2$ and ∞. Here, we are interested in $p = 2$, because ℓ_2 norms are employed in LDA as well. K-SVD [1] has been proposed for learning $\mathbf{D}$ with a fixed number of atoms K. Let $\mathbf{D}^*$ be the overcomplete dictionary $\mathbf{D}^*$ learned by the K-SVD.

3 Dual Linear Discriminant Analysis of Sparse Representations

Let the training set contains N_g genres and each genre class $\mathcal{Y}_i$ has n_i samples whose sample mean vector is denoted by $\mathbf{m}_i$, $i = 1, 2, \ldots, N_g$. If $\mathbf{S}_w$, $\mathbf{S}_b$, and $\mathbf{m}$ are the within-class sample covariance matrix, the between-class sample covariance matrix, and the gross sample mean vector of the whole training set, respectively, the MDFs are obtained by projecting the training samples using the columns of matrix $\mathbf{W}^*$, so that the ratio of the determinants is maximized [6]:

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{|\mathbf{W}^T\,\mathbf{S}_b\,\mathbf{W}|}{|\mathbf{W}^T\,\mathbf{S}_w\,\mathbf{W}|}. \tag{1}$$

To cope with the small sample size problem, we propose to apply LDA in the space of the sparse representations defined by the matrix $\mathbf{D}^*$. Let $\hat{\mathbf{m}}_i$ be the sample mean vector of the sparse coefficients associated to the training samples that belong to the ith class, i.e. $\hat{\mathbf{m}}_i$ is defined as $\hat{\mathbf{m}}_i = \frac{1}{n_i} \sum_{j:\,\mathbf{y}_j \in \mathcal{Y}_i} \mathbf{x}_j$. Then, $\mathbf{S}_w \approx \mathbf{D}^*\,\hat{\mathbf{S}}_w\,[\mathbf{D}^*]^T$, where $\hat{\mathbf{S}}_w$ is the within-class sample covariance matrix of the sparse coefficients. Similarly, $\mathbf{S}_b \approx \mathbf{D}^*\,\hat{\mathbf{S}}_b\,[\mathbf{D}^*]^T$, where $\hat{\mathbf{S}}_b$ is the between-class sample covariance matrix of the sparse coefficients. Let $\hat{\mathbf{W}} \triangleq [\mathbf{D}^*]^T\,\mathbf{W}$. The optimization problem (1) can be recast as

$$\max_{\hat{\mathbf{W}}} \frac{|\hat{\mathbf{W}}^T\,\hat{\mathbf{S}}_b\,\hat{\mathbf{W}}|}{|\hat{\mathbf{W}}^T\,\hat{\mathbf{S}}_w\,\hat{\mathbf{W}}|}. \tag{2}$$

Let $\hat{\mathbf{W}}^*$ be the solution of the optimization problem (2). The solution of the original LDA optimization problem (1) is obtained as

$$\mathbf{W}^* = \left[[\mathbf{D}^*]^\dagger\right]^T \hat{\mathbf{W}}^* \tag{3}$$

where $\dagger$ denotes the Moore-Penrose pseudoinversion. (3) suggests that the original training samples are projected to

$$\mathbf{z}_j = [\mathbf{W}^*]^T\,\mathbf{y}_j = [\hat{\mathbf{W}}^*]^T\,\mathbf{x}_j \tag{4}$$

which can be interpreted as an LDA applied to the coefficients of the sparse representation. Accordingly, using the terms most expressive features (MEFs) and MDFs, introduced in [15], we may claim that the coefficients of the sparse representation $\mathbf{x}_j$ are the MEFs, and the application of the LDA to them, producing $\mathbf{z}_j$, yields the MDFs. The cascade of sparse representation and LDA is the proposed discriminant sparse projection. Accordingly, the distance of any sample $\mathbf{y}$ from the ith class center (i.e. the class mean vector) $\mathbf{m}_i$, $i = 1, 2, \ldots, N_g$ is given by

$$\mathfrak{D}(\mathbf{y}, \mathbf{m}_i) = \sqrt{||[\hat{\mathbf{W}}^*]^T \left(\mathbf{x} - \hat{\mathbf{m}}_i\right)||^2}. \quad (5)$$

$\hat{\mathbf{W}}^*$ can be computed by applying the so-called *dual space LDA* proposed in [17], which performs LDA in the principal subspaces of $\hat{\mathbf{S}}_w$ and $\hat{\mathbf{S}}_b$ denoted by $\mathbf{U}_F^*$ and $\mathbf{U}_{\overline{F}}^*$, respectively. In (5), $||[\hat{\mathbf{W}}^*]^T \left(\mathbf{x} - \hat{\mathbf{m}}_i\right)||^2 = ||[\mathbf{U}_F^*]^T \left(\mathbf{x} - \hat{\mathbf{m}}_i\right)||^2 + \varrho ||[\mathbf{U}_{\overline{F}}^*]^T \left(\mathbf{x} - \hat{\mathbf{m}}_i\right)||^2$, where ϱ is used to ensure that the two ℓ_2 norms possess the same scale in the two principal subspaces. The test sample $\mathbf{y}$ is classified to genre

$$i^* = \underset{i}{\operatorname{argmin}}\, \mathfrak{D}(\mathbf{y}, \mathbf{m}_i). \quad (6)$$

Both theoretical and empirical studies have demonstrated the advantages of the classifier combination paradigm over the individual classifiers [8]. Our interest is in the design of a classifier ensemble, which resorts to discriminant sparse projections trained on different data subsets. In particular, the classifier ensemble has been constructed by exploiting the 10 folds the training dataset is split into by stratified 10 fold cross-validation in order to learn the overcomplete dictionary $[\mathbf{D}^*]_\tau$ and the projection matrices $[\mathbf{U}_F^*]_\tau$ and $[\mathbf{U}_{\overline{F}}^*]_\tau$ in each training dataset fold $\tau = 1, 2, \ldots, 10$. By doing so, 10 discriminant sparse projections are learned. Then, for each test sample, a voting is performed between the classification labels assigned to it by (6) using the aforementioned 10 discriminant sparse projections. The final decision is to classify the test sample to the class received the most votes.

4 Experimental Results

In order to assess the accuracy of the proposed discriminant sparse projections for genre classification, experiments are conducted on the publicly available dataset GTZAN, that has been collected by G. Tzanetakis [16]. The dataset consists of 10 genre classes, namely Blues, Classical, Country, Disco, HipHop, Jazz, Metal, Pop, Reggae, Rock. Each genre class contains 100 audio recordings 30 sec long. In [2, 9–12, 16], stratified 10-fold cross-validation has been employed. Thus 10 training subsets of 900 audio recordings are defined and by vectorizing the representation of auditory temporal modulations extracted from each recording in the training subset, the raw training pattern matrix of size 768×900 is created. By downsampling the raw vectorized representations of auditory temporal modulations with ratios 1/8, 1/4, 1/3, and 1/2, $\mathbf{Y} \in \mathbb{R}^{M \times N_{st}}$ is obtained for $M \in \{12, 48, 85, 192\}$ and $N_{st} = 900$, respectively. 100 samples (10 of each genre), that are not included in each training subset, form the raw test pattern matrix of size 768×100, which also undergoes downsampling. The overcomplete dictionary learned by KSVD is formed by $K = 400$ atoms. At most $L = 20$ coefficients for the sparse representation are derived by Orthogonal Matching Pursuit (OMP), which is repeated for 80 iterations. To estimate the null space of the within-class sample covariance matrices, the tolerance parameter in the function `rank.m` of MATLAB was set to 10^{-5}. The aforementioned experimental setup is adopted for all individual classifiers. However, for the classifier ensembles, we employ the 10 training subsets to learn 10 discriminant sparse projections, which are applied to each test subset as was explained in Section 3. The application of majority voting to the decisions taken by 10 individual classifiers, yields a music genre classification accuracy of 84.96% on average.

Notable music genre classification accuracies reported on the GTZAN dataset are summarized in Table 1. Although the table entries are sorted in a decreasing order of the reported accuracy, it is worth noting that best accuracies have been reported by the more recent techniques in a course of 8 years after the release of the GTZAN dataset and the first music genre classification accuracy reported on it. In the following, due to lack of space, we comment only three table entries. The best accuracy ever reported (e.g. 91%) was obtained by applying the sparse representation-based classification proposed in [18] to the slow auditory temporal modulations [12]. Bergstra et al. tested mel-frequency cepstral coefficients, fast Fourier transform coefficients, linear prediction coefficients, and zero-crossing rate and reported classification accuracy reaching 82.5% for the Adaboost meta-classifier. It is seen that the proposed discriminant sparse projections classifier ensemble, that builds also on sparse approximations, offers an accuracy

that exceeds by 2.46% the best accuracy reported in [3] without exploiting sparseness. For comparison purposes, a music genre classification accuracy of 71.3% was reported for LDA applied to Daubechies wavelet coefficient histograms on the same dataset [9]. However, Li et al. reported 78.5% by employing support vector machines (SVMs) and LDA for classification [9].

Table 1. Notable classification accuracies achieved by music genre classification techniques (in %).

Reference	Accuracy
[12]	91
[3]	82.5
[9]	78.5
[11]	78.2
[10]	76.8
[2]	75
[7]	74
[16]	61

The reported accuracy of 84.96% in addition to the accuracy 91% disclosed in [12] justifies paying more research effort toward the extraction of sparse representations and the assimilation of such sparse representations within classifiers.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Processing*, 54(11):4311–4322, November 2006.

[2] E. Benetos and C. Kotropoulos. A tensor-based approach for automatic music genre classification. In *Proc. XVI European Signal Processing Conf.*, Lausanne, Switzerland, 2008.

[3] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kegl. Aggregate features and Adaboost for music classification. *Machine Learning*, 65(2-3):473–484, December 2006.

[4] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Information Theory*, 52(2):489–509, February 2006.

[5] D. L. Donoho. Compressed sensing. *IEEE Trans. Information Theory*, 52(4):1289–1306, April 2006.

[6] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, N.Y., 2nd edition, 1990.

[7] A. Holzapfel and Y. Stylianou. Musical genre classification using nonnegative matrix factorization-based features. *IEEE Trans. Audio, Speech, and Language Processing*, 16(2):424–434, February 2008.

[8] L. I. Kuncheva. *Combining Pattern Classifiers. Methods and Algorithms*. John Wiley & Sons, Chichester, U.K., 2004.

[9] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proc. 26th Annual ACM Conf. Research and Development in Information Retrieval*, pages 282–289, Toronto, Canada, 2003.

[10] T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. 6th Int. Symposium Music Information Retrieval*, London, U.K., 2005.

[11] I. Panagakis, E. Benetos, and C. Kotropoulos. Music genre classification: A multilinear approach. In *Proc. 9th Int. Symposium Music Information Retrieval*, Philadelphia, PA, 2008.

[12] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music genre classification via sparse representations of auditory temporal modulations. In *Proc. XVII European Signal Processing Conf.*, Glasgow, Scotland, 2009.

[13] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: A survey. *IEEE Signal Processing Magazine*, 23(2):133–141, March 2006.

[14] S. Sukittanon, L. E. Atlas, and J. W. Pitton. Modulation-scale analysis for content identification. *IEEE Trans. Signal Processing*, 52(10):3023–3035, October 2004.

[15] D. L. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(8):831–836, August 1996.

[16] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Speech and Audio Processing*, 10(5):293–302, July 2002.

[17] X. Wang and X. Tang. Dual space linear discriminant analysis for face recognition. In *Proc. 2004 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, volume 2, pages 564–569, 2004.

[18] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(2):210–227, February 2009.

[19] X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE Trans. Information Theory*, 38(2):824–839, March 1992.