Non-Negative Multilinear Principal Component Analysis of Auditory Temporal Modulations for Music Genre Classification

Yannis Panagakis, Constantine Kotropoulos, *Senior Member, IEEE*, and Gonzalo R. Arce, *Fellow, IEEE*

***Abstract*—Motivated by psychophysiological investigations on the human auditory system, a bio-inspired two-dimensional auditory representation of music signals is exploited, that captures the slow temporal modulations. Although each recording is represented by a second-order tensor (i.e., a matrix), a third-order tensor is needed to represent a music corpus. Non-negative multilinear principal component analysis (NMPCA) is proposed for the unsupervised dimensionality reduction of the third-order tensors. The NMPCA maximizes the total tensor scatter while preserving the non-negativity of auditory representations. An algorithm for NMPCA is derived by exploiting the structure of the Grassmann manifold. The NMPCA is compared against three multilinear subspace analysis techniques, namely the non-negative tensor factorization, the high-order singular value decomposition, and the multilinear principal component analysis as well as their linear counterparts, i.e., the non-negative matrix factorization, the singular value decomposition, and the principal components analysis in extracting features that are subsequently classified by either support vector machine or nearest neighbor classifiers. Three different sets of experiments conducted on the GTZAN and the ISMIR2004 Genre datasets demonstrate the superiority of NMPCA against the aforementioned subspace analysis techniques in extracting more discriminating features, especially when the training set has small cardinality. The best classification accuracies reported in the paper exceed those obtained by the state-of-the-art music genre classification algorithms applied to both datasets.**

***Index Terms*—Auditory representations, music genre classification, nonnegative matrix factorization (NMF), non-negative multilinear principal components analysis (NMPCA), non-negative tensor factorization (NTF).**

I. Introduction

THE efficient organization of large music databases is of paramount importance for the electronic music distribution. Music genre is probably the most popular description of music content [1] to be exploited for the organization of music repositories despite it is not well-defined, since it may depend on cultural, artistic, or market factors, and the boundaries between genres are fuzzy [2].

Hopefully, there is evidence that the audio signal contains information about genre [2], [3]. Most of the music genre classification algorithms resort to the so-called *bag-of-features* (BOF) approach [2], which models the audio signals by the long-term statistical distribution of their short-time spectral features. These features can be roughly classified into three classes (i.e., timbral texture features, rhythmic features, and pitch content features) or their combinations [3]. Pattern recognition algorithms are employed next to classify the feature vectors extracted from short-time segments into genres. Frequently used classifiers include the nearest-neighbor (NN), the support vector machines (SVMs), or classifiers, which resort to Gaussian mixture models, linear discriminant analysis (LDA), non-negative matrix factorization (NMF), non-negative tensor factorization (NTF). Several common audio datasets have been used in experiments in order to make the reported classification accuracies comparable. Notable results on music genre classification are summarized in Table I.

Aucouturier *et al.* [13] have observed that recent systems, which assess audio similarity using the BOF approach, have failed to offer significant performance gains over early systems. Not to mention that their accuracy makes their practical use unrealistic.

Having the aforementioned remarks in mind, and motivated by the fact that most of the perceptual properties of both speech and music are encoded by slow temporal modulations [14]–[18], instead of the BOF approach, we propose here, a bio-inspired auditory representation, that maps a given sound to a 2-D representation of its slow temporal modulations. This is the first paper contribution. Such a representation extends the joint acoustic and modulation frequency analysis [14] by exploiting the properties of the human auditory system [16], [19]. By just feeding the auditory temporal modulations to an SVM with a radial basis function (RBF) kernel, music genre classification accuracy equal to 77.09% and 78.64% has been obtained on GTZAN dataset and ISMIR2004Genre one, respectively. The aforementioned classification accuracies provide a first hint that the auditory temporal modulations carry more discriminating information about music genre than many of BOF approaches included in Table I.

The proposed 2-D auditory representation can be treated as a second-order tensor (i.e., a matrix). Although each recording is represented by a matrix, a third-order tensor is needed to

Manuscript received December 15, 2008; revised September 29, 2009. Current version published February 10, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paris Smaragdis.

Y. Panagakis is with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 541 24, Greece (e-mail: yannisp@csd.auth.gr).

C. Kotropoulos is with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 541 24, Greece, on leave from the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716-3130 USA (e-mail: costas@aiia.csd.auth.gr).

G. R. Arce is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716-3130 USA. (e-mail: arce@ece.udel.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2036813

TABLE I
NOTABLE CLASSIFICATION ACCURACIES ACHIEVED BY MUSIC GENRE CLASSIFICATION APPROACHES

Reference	Dataset	Accuracy	Reference	Dataset	Accuracy
Bergstra *et al.* [4]	GTZAN	82.50%	Holzapfel *et al.* [5]	ISMIR2004	83.50%
Li *et al.* [6]	GTZAN	78.50%	Pampalk *et al.* [7]	ISMIR2004	82.30%
Panagakis *et al.* [8]	GTZAN	78.20%	Panagakis *et al.* [8]	ISMIR2004	80.95%
Lidy *et al.* [9]	GTZAN	76.80%	Lidy *et al.* [10]	ISMIR2004	79.70%
Benetos *et al.* [11]	GTZAN	75.00%	Bergstra *et al.* [4]	MIREX2005	82.34%
Holzapfel *et al.* [5]	GTZAN	74.00%	Lidy *et al.* [9]	MIREX2007	75.57%
Tzanetakis *et al.* [3]	GTZAN	61.00%	Mandel *et al.* [12]	MIREX2007	75.03%

represent a music corpus. Accordingly, music genres are expected to be defined on subspaces of the three-order (and in general high-order) tensor. In *multilinear algebra*, tensors are defined as the multidimensional equivalent of matrices or vectors [20]. In addition, auditory representations are constrained to be non-negative. They are highly redundant as well [21]. Therefore, it is reasonable to assume that the associated tensors are confined into a subspace of an intrinsically low dimension. Let $\{\mathcal{X}_i, i = 1, 2, \ldots, n\}$ be a set of n tensor samples $\mathcal{X}_i \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$. Subspace analysis methods can reveal such a low dimensional subspace by defining a transformation that maps the original tensor space $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ onto a tensor subspace $\mathbb{R}^{P_1 \times P_2 \times \ldots \times P_N}$ with $P_l < I_l, l = 1, 2, \ldots, N$. Indeed, linear subspace analysis methods (i.e., when $l = 1$) such as PCA, LDA, and NMF [22] have successfully been used for dimensionality reduction of vectorized data. By vectorizing a typical 2-D auditory representation, the dimensionality of the resulting feature space is usually much larger than the size of typical music corpora used for genre classification. This is known as *curse of dimensionality* or *small sample size problem* (SSS) [23]. Many classifiers, cannot cope with the high-dimensionality of a small number of feature vectors. One solution is to reduce the dimensionality of the feature space by using the aforementioned linear dimensionality reduction techniques. Unfortunately, these methods break the natural structure and any constraints on the data (e.g., positivity), while they assume either a Gaussian distribution of feature vectors or diagonal covariance matrices in order to reduce the number of unknown model parameters [23]–[25]. In addition, handling such high-dimensional feature vectors is computationally expensive. On the contrary, dimensionality reduction applied directly to tensors rather to vectors retains many of the data properties.

In this paper, we propose non-negative multilinear principal components analysis (NMPCA) in order to cope with the SSS problem of the non-negative auditory temporal modulation representations. NMPCA objective is to maximize the total variation of the given non-negative tensors while preserving their non-negativity. It is an unsupervised multilinear dimensionality reduction technique whose development is motivated by the success of NTF [26]–[28] in music genre classification [11] and the first promising results of the application of multilinear principal components analysis (MPCA) [29] in music genre classification [8]. However, the original MPCA does not preserves the non-negativity of the auditory temporal modulations, a property that is generally desirable in domains, where the underlying factors have physical or psychological interpretation [22], [28], [30]. This is the second contribution of the paper.

Furthermore, building on the notions of *homogeneous functions* (i.e., functions defined on the subspace spanned by the columns of a suitable orthonormal matrix) and the *Grassmann manifold* (i.e., the set of matrices defined over the aforementioned subspace), a novel framework for the maximization of homogenous, continuous, differentiable functions with non-negative constraints over the Grassmann manifold is proposed. This is the third contribution of the paper. This framework is employed in the derivation of an algorithm for NMPCA. Convergence and complexity analysis of the NMPCA algorithm is studied as well.

The NMPCA is applied to music genre classification. The performance of NMPCA in feature extraction against the state-of-the-art unsupervised multilinear subspace analysis techniques, namely the MPCA, the HOSVD, and the NTF as well as their linear counterparts (i.e., the PCA, the SVD, and NMF) is also investigated. The features extracted by the aforementioned multilinear and linear subspace analysis techniques are classified by the SVM with either an RBF or a linear kernel and the NN classifier, which employs distances, such as L_1, L_2, and the cosine similarity measure (CSM). In order to compare the reported genre classification rates with those achieved by the algorithms listed in Table I, two sets of experiments are conducted. First, stratified tenfold cross-validation tests are applied to the GTZAN dataset, which yield a classification accuracy of 84.3%. Second, experiments on the ISMIR2004Genre dataset are conducted by adhering to the setup employed during ISMIR2004 evaluation tests, which splits the dataset into two equal disjoint subsets with the first one used for training and the second one used for testing. The best classification accuracy is equal to 83.15% in the ISMIR2004Genre dataset. To the best of our knowledge, the reported classification accuracy is the highest for both datasets. Furthermore, in real world conditions the number of training samples per music genre is often limited. In order to simulate such conditions, experiments with training sets having small cardinality were conducted. Experimental results indicate that the classification accuracy exceeds 70% even when 100 training samples are employed in the GTZAN dataset.

The remainder of the paper is as follows. In Section II, the bio-inspired auditory representation based on a computational auditory model is described. Basic concepts of multilinear algebra are introduced in Section III, while multilinear subspace analysis techniques are briefly addressed in Section IV. The proposed NMPCA method is detailed in Section V. Convergence and computational complexity analysis of NMPCA algorithm are discussed in this section as well. Experimental results are demonstrated in Section VI. Conclusions are drawn and future research direction are indicated in Section VII.

II. Bio-Inspired Joint Acoustic and Modulation Frequency Representation of Music

The conventional spectrogram emphasizes many spectro-temporal details that are not directly germane to the music information encoded in the signal and does not take into account the perception and cognition of a human listener [31]. A key step for representing music signals in a psycho-physiologically consistent manner is to focus on how the audio information is encoded in the human *primary auditory cortex*. In this section, we develop a bio-inspired 2-D representation of audio, by modeling the path of auditory processing. The proposed 2-D auditory representation is a joint acoustic and modulation frequency representation [14], that discards much of the spectro-temporal details and focuses on the underlying slow temporal modulations of the music signal. There is evidence that important time-varying information is contained in the slow temporal modulation of audio signals [14]–[18].

The computational model of human auditory system consists of two basic processing stages. The first stage models the early auditory system, which converts the acoustic signal into a neural representation, the so-called *auditory spectrogram*. This representation is a time-frequency distribution along a tonotopic (logarithmic frequency) axis. At the second stage, the temporal modulation content of the auditory spectrogram is estimated by applying a wavelet transform to each row of the auditory spectrogram.

The computation of the auditory spectrogram consists of three operations, which mimic the early stages of human auditory processing. In this paper, the mathematical model of Yang *et al.* [32] is adopted. First, a constant-Q transform is applied to the acoustic signal $s(t)$. The constant-Q transform applies a bank of filters, such that the ratio of each filter center frequency to its resolution is constant. Here, the constant-Q transform is implemented via a bank of 96 overlapping bandpass filters with center frequencies uniformly distributed along the tonotopic axis over four octaves. Let f denote the logarithmic frequency. The impulse response of each filter is denoted as $h_{\text{cochlea}}(t,f)$. The output of cochlear filter is given by $y_{\text{cochlea}}(t,f) = s(t) *_t h_{\text{cochlea}}(t,f)$, where $*_t$ denotes convolution in the time domain. It is transduced into an auditory nerve pattern $y_{\text{an}}(t,f)$ by a hair cell stage, which converts the cochlear output into inner hair cell intracellular potential. The just described process is modelled by high-pass filtering, i.e., $(\partial y_{\text{cochlea}}/\partial t)(t,f)$, corresponding to the fluid-cilia coupling, followed by an instantaneous nonlinear compression $g_{h_c}(.)$, which models the gated ionic channels, and finally low-pass filtering by $\mu_{h_c(t)}(.)$, that models the hair cell membrane leakage. That is, $y_{\text{an}}(t,f) = g_{h_c}((\partial y_{\text{cochlea}}/\partial t)(t,f)) *_t \mu_{h_c(t)}$. At a second step, a lateral inhibitory network (LIN) detects the discontinuities in the response along the tonotopic axis of the auditory nerve array. The LIN can be approximated by a first-order derivative with respect to the logarithmic frequency followed by a half-wave rectifier, i.e., $y_{\text{LIN}}(t,f) = \max((\partial y_{\text{an}}/\partial f)(t,f), 0)$. The final step of this stage is the integration of $y_{\text{LIN}}(t,f)$ over a short window $\mu_{\text{midbrain}}(t;\tau) = e^{-t/\tau}u(t)$, where $u(t)$ is the unit step function. The time constant τ of the order of a few milliseconds (typically 2–8 ms) accounts for the further loss of phase-locking observed in the midbrain. Thus, the auditory spectrogram $y(t,f)$ is obtained by $y(t,f) = y_{\text{LIN}}(t,f) *_t \mu_{\text{midbrain}}(t;\tau)$.

Higher central auditory stages, especially the primary auditory cortex, further analyze the auditory spectrogram by estimating the signal content in slow spectro-temporal modulations. In this paper, we are interested in the slow temporal modulations only. In order to mimic the human perception of temporal modulation, we apply the concept of *modulation scale analysis* [14] in order to derive a compact representation that captures the underlying temporal modulations of an audio signal. Recent psychoacoustic evidence suggests that a log frequency axis with a constant-Q resolution best mimics the human perception of modulation frequency [16]. In [14], a continuous wavelet transform is applied to the temporal rows of a standard spectrogram in order to efficiently approximate this constant-Q effect. Instead of the standard spectrogram, in this paper we use the auditory spectrogram as input to the modulation scale analysis.

The modulation scale analysis consist of two stages. First, for discrete rate r, the wavelet filter $\Psi(t)$ is applied along each temporal row of the auditory spectrogram $y(t,f)$, i.e.,

$$X^{SP}(r,t,f) = \frac{1}{r} y(t,f) *_t \Psi\left(-\frac{t}{r}\right). \tag{1}$$

Equation (1) can be interpreted as filtering the temporal envelope of each cochlear channel output. The multiresolution wavelet analysis is implemented via a bank of Gabor filters, that are selective to different temporal modulation parameters ranging from slow to fast temporal rates (in Hz). Since, the analysis yields a rate–time–frequency representation for each recording, the entire auditory spectrogram is modeled by a 3-D representation of rate, time, and frequency

In the final step, the power of the 3-D temporal modulation representation $X^{SP}(r,t,f)$ is obtained by integrating across the wavelet translation axis t. Thus, a joint rate–frequency representation results that has no uniform resolution in the modulation frequency indexed by the discrete rate r

$$X^{JF}(r,f) = \int \left|X^{SP}(r,t,f)\right|^2 dt. \tag{2}$$

The resulting 2-D representation (2) is referred to as *auditory temporal modulations* representation. The extraction of the auditory temporal modulations representation is depicted in Fig. 1. In Fig. 2, the auditory temporal modulations representations of ten music recordings that belong to ten different music genre classes are shown. Psychophysiological evidence [33] 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 four octaves along the tonotopic axis (i.e., 24 filters per octave). Accordingly, the auditory temporal modulation representation of an audio recording is naturally represented by a second-order tensor (matrix) $\mathbf{X} \in \mathbb{R}_+^{I_1 \times I_2}$, where $I_1 = I_{\text{frequency}} = 96$ and $I_2 = I_{\text{rate}} = 8$. Thus, an ensemble of audio recordings can be represented by a third-order tensor created by stacking the second-order tensors associated to the recordings. Then, the data

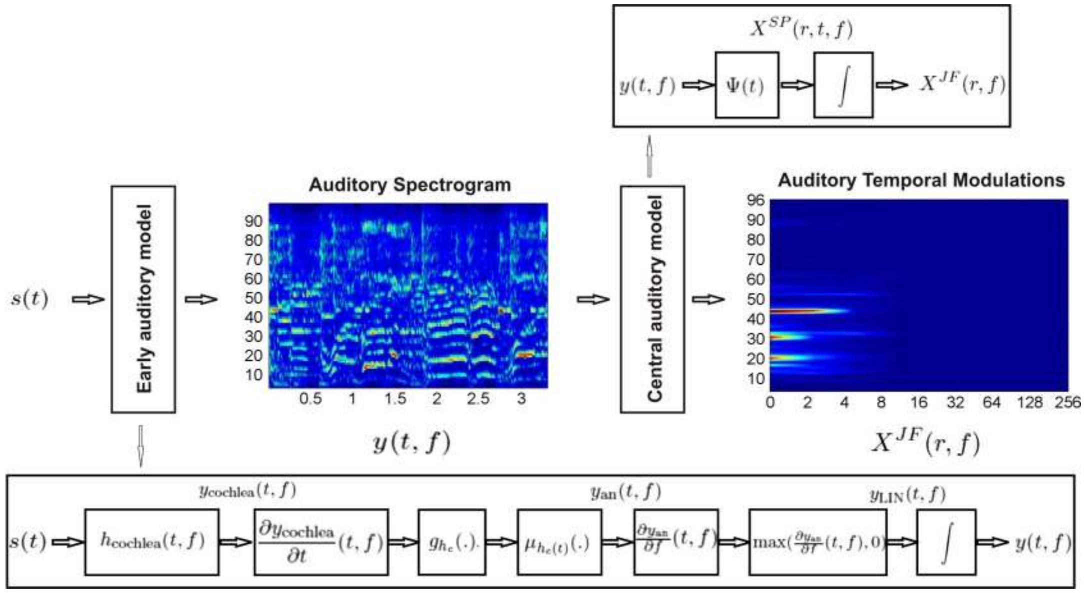


Fig. 1. Flow-chart of auditory temporal modulations representation extraction.

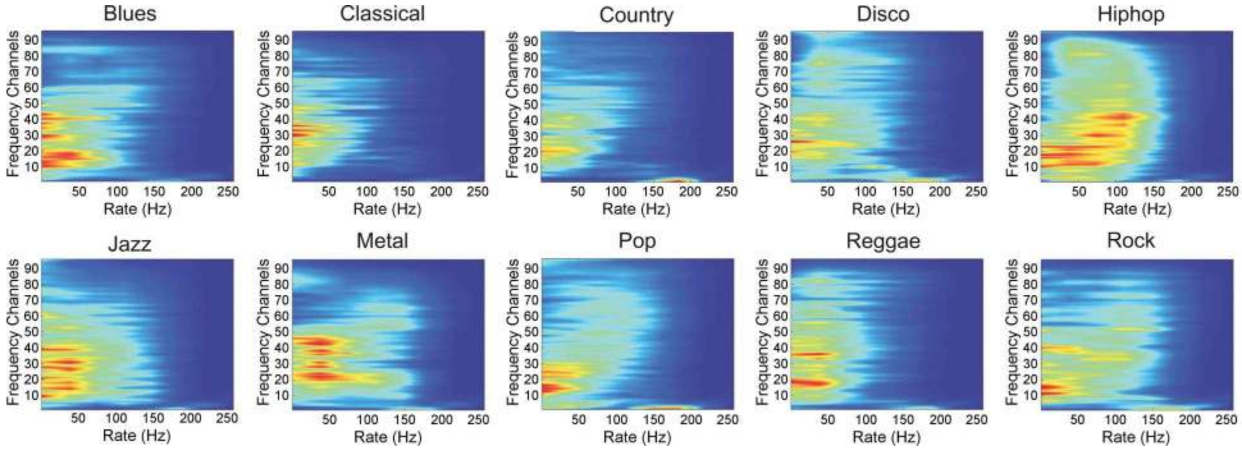


Fig. 2. Auditory temporal modulations representations of ten music recordings from the GTZAN dataset.

tensor $\mathcal{X} \in \mathbb{R}_+^{I_1 \times I_2 \times I_3}$ is obtained, where $I_3 = I_{\text{samples}}$ denotes the number of available recordings. Such multiway data can be handled by employing the mathematical tools of multilinear algebra [20], [34], which is briefly introduced in the next section to make the paper self-contained.

III. Multilinear Algebra Basics

Hereafter, vectors are denoted by lowercase boldface letters (e.g., $\mathbf{u}$) and matrices by uppercase boldface letters (e.g., $\mathbf{U}$). The ith entry of a vector $\mathbf{u}$ is denoted by $(\mathbf{u})_i$, while the (i,j) element of a matrix $\mathbf{U}$ is denoted by $(\mathbf{U})_{i,j}$. *Tensors* are considered as the multidimensional equivalent of matrices (second-order tensors) and vectors (first-order tensors) [20] and are denoted by calligraphic letters (e.g., $\mathcal{A}$). An Nth-order or an N-way real-valued tensor $\mathcal{A}$ is defined over the tensor space $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, where $I_l \in \mathbb{Z}$ for $l = 1, 2, \ldots, N$. The *order* of a tensor is the number of indices needed to address its elements. Consequently, each element of an Nth-order tensor $\mathcal{A}$ is addressed by N indices, $(\mathcal{A})_{i_1, i_2, \ldots, i_N}$.

The *mode-l matricization* of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ maps $\mathcal{A}$ to a matrix $\mathbf{A}_{(l)} \in \mathbb{R}^{I_l \times \bar{I}_l}$ with $\bar{I}_l = \prod_{\substack{m=1 \\ m \neq l}}^{N} I_m$ such that the tensor element $(\mathcal{A})_{i_1, i_2, \ldots, i_N}$ is mapped to the matrix element $(\mathbf{A})_{i_l, j}$, where $j = 1 + \sum_{\substack{k=1 \\ k \neq l}}^{N} (i_k - 1) J_k$ with $J_k = \prod_{\substack{m=1 \\ m \neq l}}^{k-1} I_m$. Furthermore, it is possible to vectorize a tensor $\mathcal{A}$ by reordering its elements to form a vector. The vectorized form of a tensor $\mathcal{A}$ is denoted as $vec(\mathcal{A})$.

The *norm of tensor* $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, is denoted as $||\mathcal{A}||$, and it is defined as the square root of the sum of the squares of all its elements [34]. It can be shown that $||\mathcal{A}|| = ||\mathbf{A}_{(l)}||_F$, where $||.||_F$ denotes the Frobenious norm.

The *tensor product* of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ with a tensor $\mathcal{B} \in \mathbb{R}^{J_1 \times J_2 \times \ldots \times J_M}$, $\mathcal{A} \otimes \mathcal{B}$, is defined by $(\mathcal{A} \otimes \mathcal{B})_{i_1, i_2, \ldots, i_N j_1, j_2, \ldots, j_M} = (\mathcal{A})_{i_1, i_2, \ldots, i_N} (\mathcal{B})_{j_1, j_2, \ldots, j_M}$.

Given two tensors $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N \times K_1 \times K_2 \times \ldots \times K_M}$ and $\mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N \times J_1 \times J_2 \times \ldots \times J_P}$ the *contraction* on the tensor product $\mathcal{A}\otimes\mathcal{B}$ with respect to indices $i_1, i_2, \ldots, i_N$ is expressed as

$$[[\mathcal{A}\otimes\mathcal{B};(1:N),(1:N)]] = \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots \sum_{i_N=1}^{I_N}(\mathcal{A})_{i_1,i_2,\ldots,i_N,k_1,k_2,\ldots,k_M}\ldots(\mathcal{B})_{i_1,i_2,\ldots,i_N,j_1,j_2,\ldots,j_P}. \quad (3)$$

When the contraction on the tensor product of N-order tensors $\mathcal{A}$ and $\mathcal{B}$ is with respect to all indices but the lth index, we can express this procedure as in [25]

$$\begin{aligned}\left[\left[\mathcal{A}\otimes\mathcal{B};(\bar{l}),(\bar{l})\right]\right] &= [[\mathcal{A}\otimes\mathcal{B};(1:l-1,l+1:N),\\ &\qquad(1:l-1,l+1:N)]]\\ &= \mathbf{A}_{(l)}\mathbf{B}_{(l)}^T. \end{aligned}\quad (4)$$

From (4) it is seen that $[[\mathcal{A}\otimes\mathcal{B};(\bar{l}),(\bar{l})]] \in \mathbb{R}^{I_l\times I_l}$.

The *mode-l product* of a tensor $\mathcal{A} \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_N}$ with a matrix $\mathbf{U} \in \mathbb{R}^{J\times I_l}$, denoted by $\mathcal{A}\times_l\mathbf{U} \in \mathbb{R}^{I_1\times I_2\times\ldots\times I_{l-1}\times J\times I_{l+1}\times\ldots\times I_N}$, is defined as

$$\begin{aligned}&(\mathcal{A}\times_l\mathbf{U})_{i_1,\ldots,i_{l-1},j,i_{l+1},\ldots,i_N}\\ &\quad= \sum_{i_l=1}^{I_l}(\mathcal{A})_{i_1,\ldots,i_{l-1},j,i_{l+1},\ldots,i_N}(\mathbf{U})_{j,i_l}\\ &\quad= [[\mathcal{A}\otimes\mathbf{U};(l),(2)]]. \end{aligned}\quad (5)$$

In order to simplify the notation, we denote $\mathcal{A}\times_1\mathbf{U}_1\times_2\ldots\times_N\mathbf{U}_N = \mathcal{A}\prod_{i=1}^N\times_i\mathbf{U}_i$. Furthermore, $\mathcal{A}\times_1\mathbf{U}_1\ldots\times_{l-1}\mathbf{U}_{l-1}\times_{l+1}\mathbf{U}_{l+1}\ldots\times_N\mathbf{U}_N = \mathcal{A}\prod_{i=1,i\neq l}^N\times_i\mathbf{U}_i = \mathcal{A}\times_{\overline{i=l}}\mathbf{U}_i$.

An Nth-order tensor $\mathcal{A}\in\mathbb{R}^{I_1\times I_2\times\ldots\times I_N}$ has *rank*-1, when it is decomposed as the outer product of N vectors $\mathbf{u}_i$, $i = 1, 2, \ldots, N$, i.e., $\mathcal{A} = \mathbf{u}_1\circ\mathbf{u}_2\circ\ldots\circ\mathbf{u}_N = \bigcirc_{i=1}^N\mathbf{u}_i$, where $\circ$ stands for the vector outer product.

Let us close this section by introducing some matrix products that will be used next. The *Khatri-Rao* product of matrices $\mathbf{A}\in\mathbb{R}^{I\times K}$ and $\mathbf{B}\in\mathbb{R}^{J\times K}$ is denoted by $\mathbf{A}\odot\mathbf{B}$ and yields a matrix of dimensions $(IJ)\times K$. The *Hadamard* product is the element wise matrix product. Given two matrices $\mathbf{A}$ and $\mathbf{B}$ both of dimension $I\times J$, their Hadamard product, denoted by $\mathbf{A}*\mathbf{B}$, is also of dimensions $I\times J$. Definitions of the aforementioned matrix products can be found in [34] for example.

IV. Multilinear Subspace Analysis Techniques

Three unsupervised multilinear subspace analysis techniques, namely the NTF, the HOSVD, and the MPCA are briefly discussed. Clearly no class information is used by the just-mentioned techniques. Their linear counterparts, namely the NMF, the SVD, and the PCA can be viewed as special cases for first-order tensors (i.e., vectors). In the following, let $\{\mathcal{X}_i, i = 1, 2, \ldots, n\}$ be a set of n training tensor samples $\mathcal{X}_i\in\mathbb{R}^{I_1\times I_2\times\ldots\times I_N}$ which is represented by an $(N+1)$-order tensor $\mathcal{X}\in\mathbb{R}^{I_1\times I_2\times\ldots\times I_N\times I_{N+1}}$, where $I_{N+1} = n$.

A. NTF

NTF is a generalization of the NMF [22] for high-order tensors. It decomposes a given non-negative training tensor $\mathcal{X}\in\mathbb{R}_+^{I_1\times I_2\times\ldots\times I_N\times I_{N+1}}$ into a sum of k rank-1 tensors

$$\mathcal{X}\approx\sum_{j=1}^k\bigcirc_{i=1}^{N+1}\mathbf{u}_i^j. \quad (6)$$

Various NTF algorithms have been proposed [11], [26]–[28]. In this paper, we resort to the NTF algorithm, which minimizes the tensor norm of the difference between the given tensor and its expansion (6). In order to apply the NTF algorithm for an $(N+1)$th-order tensor, $N+1$ matrices $\mathbf{U}_l\in\mathbb{R}_+^{I_l\times k}$, $l = 1, 2, \ldots, N+1$ should be created and initialized randomly with non-negative values. Let t denote the iteration index. The following update rule in matrix form is applied to each $\mathbf{U}_l$ in an alternating fashion:

$$\mathbf{U}_l^{t+1} = \mathbf{U}_l^t * \frac{\mathbf{X}_{(l)}\mathbf{Z}_l}{\mathbf{U}_l^t\mathbf{Z}_l^T\mathbf{Z}_l} \quad (7)$$

where $\mathbf{Z}_l \triangleq \mathbf{U}_{N+1}^t\odot\ldots\odot\mathbf{U}_{l+1}^t\odot\mathbf{U}_{l-1}^t\odot\ldots\odot\mathbf{U}_1^t$ and $\mathbf{U}_l^t$ refers to the matrix before updating. It is worth noting, that operators such as Khatri–Rao product preserve the inner structure of data. By applying NTF to $\mathcal{X}$, the decomposition (in matricized form)

$$\begin{aligned}\mathbf{X}_{(N+1)} &= \mathbf{U}_{N+1}(\mathbf{U}_N\odot\mathbf{U}_{N-1}\odot\ldots\odot\mathbf{U}_1)^T \Leftrightarrow\\ \mathbf{X}_{(N+1)}^T &= (\mathbf{U}_N\odot\mathbf{U}_{N-1}\odot\ldots\odot\mathbf{U}_1)\mathbf{U}_{N+1}^T \end{aligned}\quad (8)$$

is obtained, where $\mathbf{X}_{(N+1)}$ is the unfolding of tensor $\mathcal{X}$ to the samples mode. It is clear that (8) implies that every column of $\mathbf{X}_{(N+1)}^T\in\mathbb{R}_+^{(I_1I_2\ldots I_N)\times I_{N+1}}$, (i.e., a vectorized training sample), is a linear combination of the basis vectors, which span the columns of the basis matrix $\mathbf{W}\triangleq(\mathbf{U}_N\odot\mathbf{U}_{N-1}\odot\ldots\odot\mathbf{U}_1)$ with coefficients taken from the columns of coefficient matrix $\mathbf{U}_{N+1}^T$. Let $\tilde{\mathcal{X}}$ be a test (new) tensor sample, then the feature vector $\check{\mathbf{y}}$ derived by the projection $\check{\mathbf{y}} = \mathbf{W}^Tvec(\tilde{\mathcal{X}})$ can be used in place of $\tilde{\mathcal{X}}$ for either representation or classification. Following the strategy employed in [8], Gram–Schmidt orthonormalization is performed to basis matrix $\mathbf{W}$. Accordingly, $\mathbf{W} = \mathbf{QR}$, where $\mathbf{Q}$ is an orthogonal matrix whose columns define a basis that spans the same vector space with that of the learned basis $\mathbf{W}$ and $\mathbf{R}$ is an upper triangular matrix. Accordingly, the feature vector $\tilde{\mathbf{y}}$ is derived by the projection $\tilde{\mathbf{y}} = \mathbf{Q}^Tvec(\tilde{\mathcal{X}})$, because orthogonality increases the discriminative power of the projections [35].

B. HOSVD

HOSVD is a generalization of singular value decomposition (SVD) applied to high-order tensors [36]. The training tensor $\mathcal{X}$ can be decomposed as

$$\mathcal{X} = \mathcal{S}\times_1\mathbf{U}_1\times_2\mathbf{U}_2\times_3\ldots\times_{N+1}\mathbf{U}_{N+1} \quad (9)$$

where $\mathbf{U}_l$, $l = 1, 2, \ldots, N+1$ is a unitary matrix containing the left singular vectors of the mode-l unfolding of tensor $\mathcal{X}$ computed by applying SVD to $\mathbf{X}_{(l)}$. The tensor $\mathcal{S}\in\mathbb{R}^{I_1\times I_2\times\ldots\times I_{N+1}}$, known as *core tensor*, is given by

$\mathcal{S} = \mathcal{X} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \times_3 \ldots \times_{N+1} \mathbf{U}_{N+1}^T$ and has the properties of all orthogonality and ordering.

HOSVD results in a new ordered orthogonal basis for data representation in subspaces spanned by each tensor mode. Dimensionality reduction in each subspace is obtained by projecting the data onto the principal axes and keeping only the components that correspond to the largest singular values. The vectorized version of the lower dimension tensor obtained by HOSVD is used here for classification.

C. MPCA

Recently, Lu *et al.* [29] proposed MPCA as a multilinear equivalent of PCA. Given a set of training tensor samples $\{\mathcal{X}_i, i = 1,2,\ldots,n\}$, MPCA defines a multilinear transformation $\{\mathbf{U}_l \in \mathbb{R}^{I_l \times P_l}, l = 1,2,\ldots,N\}$ that maps the original tensor space $\mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ onto a tensor subspace $\mathbb{R}^{P_1 \times P_2 \times \ldots \times P_N}$ with $P_l < I_l$, $l = 1,2,\ldots,N$. That is, $\mathcal{Y}_i = \mathcal{X}_i \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \ldots \times_N \mathbf{U}_N^T$, $i = 1,2,\ldots,n\}$ are derived such that most of the variation observed between the original tensor samples is captured. A measure of the variation is the total tensor scatter defined as

$$\mathcal{S}_{\mathcal{T}} = \sum_{i=1}^{n} ||\mathcal{Y}_i - \mathcal{M}_{\mathcal{Y}}||^2 \tag{10}$$

where $\mathcal{M}_{\mathcal{Y}}$ is the mean tensor given by $\mathcal{M}_{\mathcal{Y}} = (1/n)\sum_{i=1}^{n} \mathcal{Y}_i$. The N projection matrices $\mathbf{U}_l \in \mathbb{R}^{I_l \times P_l}$, $l = 1,2,\ldots,N$, that maximize the total tensor scatter $\mathcal{S}_{\mathcal{T}}$, are obtained by solving the optimization problem

$$\{\mathbf{U}_l, l = 1,2,\ldots,N\} = \underset{\mathbf{U}_l^T\mathbf{U}_l = \mathbf{I}}{\arg\max} \mathcal{S}_{\mathcal{T}}. \tag{11}$$

Since there is no known optimal solution to the optimization problem (11), the problem was solved in [29] iteratively by employing the alternating projection scheme [37]. Let $\tilde{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ be a test (new) tensor. The lower dimension tensor obtained by MPCA $\tilde{\mathcal{Y}} = \tilde{\mathcal{X}} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \times_3 \ldots \times_N \mathbf{U}_N^T$ can be used in place of $\tilde{\mathcal{X}}$ for either representation or classification. In this paper, $vec(\tilde{\mathcal{Y}})$ is used for classification.

V. Non-Negative Multilinear Principal Components Analysis

In this section, we propose a novel extension of MPCA, that incorporates the non-negativity of the projection matrices in MPCA. The proposed extension is referred to as Non-Negative Multilinear Principal Components Analysis (NMPCA). Unlike the MPCA, the NMPCA preserves the non-negativity of the original tensor samples, a property which is crucial, when the underlying data factors (i.e., the learned basis) have physical or psychological interpretation [22], [28], [30]. Furthermore, in this section, we derive a novel multiplicative type algorithm for maximizing *homogenous functions over the Grassmann manifold* subject to the non-negativity constraints by incorporating the *natural gradient* [38] of the Grassmann manifold [39], [40] into the multiplicative updates.

A. Multiplicative Updates on the Grassmann Manifold

Let $f(\mathbf{U})$ be any continuous, differentiable, concave function of $\mathbf{U}$ for which the *homogeneity assumption* holds, i.e., $f(\mathbf{U}) = f(\mathbf{UQ})$ with $\mathbf{Q}$ being a $P \times P$ orthonormal matrix. Furthermore, it is assumed that $f(\mathbf{U})$ can be expressed as the difference of two non-negative terms. To begin with, let us determine a real-valued column-orthonormal non-negative matrix (i.e., a permutation matrix) $\mathbf{U}$ that maximizes the function $f(\mathbf{U})$. That is,

$$\mathbf{U}^* = \underset{\substack{\mathbf{U}^T\mathbf{U}=\mathbf{I} \\ \mathbf{U} \geqslant \mathbf{0}}}{\arg\max} f(\mathbf{U}) \tag{12}$$

where $\mathbf{U} \geqslant \mathbf{0}$ stands for $\mathbf{U} \in \mathbb{R}_+^{I \times P}$. Due to the homogeneity assumption, maximizing $f(\mathbf{U})$ under the orthogonality constraint is over-parameterized, and any conventional constrained optimization algorithm would meet difficulties. It follows that we should maximize (12) not just over matrices with orthonormal columns, but over an equivalence class of such matrices, i.e., for $[\mathbf{U}] = \{\mathbf{UQ}|\mathbf{Q}\ \text{orthonormal}\}$. Such an equivalence class spans the same subspace with the columns of an orthonormal matrix and is known as the *Grassmann manifold* [39], [40]. If $\mathbf{U}$ is a point on the Grassmann manifold, denoted as $Gr(I,P)$, we write $\mathbf{U} \in Gr(I,P)$. With this notation in mind the optimization problem (12) can be formulated as

$$\mathbf{U}^* = \underset{\substack{\mathbf{U} \in Gr(I,P) \\ \mathbf{U} \geqslant \mathbf{0}}}{\arg\max} f(\mathbf{U}). \tag{13}$$

A tangent vector $\boldsymbol{\Delta} \in \mathbb{R}^{I \times P}$ at $\mathbf{U}$ satisfies $\mathbf{U}^T\boldsymbol{\Delta} = 0$. Let $\mathbb{T}_{\mathbf{U}}$ be the tangent space at $\mathbf{U}$, which consists of all tangent vectors at $\mathbf{U}$. The projection onto the tangent space is $\mathbf{\Pi}_{\mathbf{U}} = \mathbf{I} - \mathbf{UU}^T$. The natural gradient of $f(\mathbf{U})$ on the Grassmann manifold at $\mathbf{U}$ denoted as $\tilde{\nabla}_{\mathbf{U}} f$ is given by

$$\tilde{\nabla}_{\mathbf{U}} f = \mathbf{\Pi}_{\mathbf{U}} \nabla_{\mathbf{U}} f = \nabla_{\mathbf{U}} f - \mathbf{UU}^T \nabla_{\mathbf{U}} f. \tag{14}$$

where $(\nabla_{\mathbf{U}} f)_{i,j} = (\partial f / \partial u_{ij})$ is the ordinary gradient. It is seen that $\tilde{\nabla}_{\mathbf{U}} f$ is the projection of $\nabla_{\mathbf{U}} f$ onto the tangent space at $\mathbf{U}$.

Amari in [38] has proved that when a parameter space has a certain underlying structure, such as the Grassman manifold, the ordinary gradient of a function does not represent its steepest direction, while the natural gradient does. Therefore, the natural gradient ascent algorithm for the maximization problem (12) over the Grassman manifold takes the form

$$\mathbf{U}^{t+1} = \mathbf{U}^t + n_t \tilde{\nabla}_{\mathbf{U}} f \tag{15}$$

where n_t is a step size controlling the learning rate. However, the update rule (15) does not maintain the non-negativity constraint after each iteration. To preserve non-negativity, we employ the strategy described in [41] and [42] in order to choose an appropriate step size n_t that ensures $\mathbf{U}^{t+1}$ is non-negative in each iteration. Following [41], it is possible to decompose the natural gradient into two non-negative parts, i.e., $\tilde{\nabla}_{\mathbf{U}} f = \tilde{\nabla}_{\mathbf{U}} f^+ - \tilde{\nabla}_{\mathbf{U}} f^-$, where

$$(\tilde{\nabla}_{\mathbf{U}} f^+)_{i,j} = \begin{cases} (\tilde{\nabla}_{\mathbf{U}} f)_{i,j}, & \text{if } (\tilde{\nabla}_{\mathbf{U}} f)_{i,j} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

$$(\tilde{\nabla}_{\mathbf{U}} f^-)_{i,j} = \begin{cases} -(\tilde{\nabla}_{\mathbf{U}} f)_{i,j}, & \text{if } (\tilde{\nabla}_{\mathbf{U}} f)_{i,j} < 0 \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

The step size n_t can be chosen to be data-dependent, i.e.,

$$n_{t_{i,j}} = \left(\frac{\mathbf{U}^t}{\tilde{\nabla}_{\mathbf{U}} f^-} \right)_{i,j}. \quad (18)$$

By inserting (18) into (15), the update rule becomes

$$\mathbf{U}^{t+1} = \mathbf{U}^t + \frac{\mathbf{U}^t}{\tilde{\nabla}_{\mathbf{U}} f^-} * (\tilde{\nabla}_{\mathbf{U}} f^+ - \tilde{\nabla}_{\mathbf{U}} f^-) = \mathbf{U}^t * \frac{\tilde{\nabla}_{\mathbf{U}} f^+}{\tilde{\nabla}_{\mathbf{U}} f^- + \epsilon} \quad (19)$$

where ϵ is a small positive number (i.e., typically 10^{-8}) used in order to ensure that the denominator in (19) is always nonzero. The *multiplicative update* (19) preserves the non-negativity of $\mathbf{U}^{t+1}$, while $\tilde{\nabla}_{\mathbf{U}} f = \mathbf{0}$ at convergence. Furthermore, the multiplicative update maintains the orthogonality constraint, since the natural gradient on the Grassmann manifold is employed. The monotonic convergence of $f(\mathbf{U})$ is guaranteed by

Theorem 1: If $\{\mathbf{U}^t\}$ is the sequence of updates generated by (19), then the function $f(\mathbf{U})$ increases monotonically to its global maximum.

Proof: As in [41].

B. NMPCA Algorithm

In order to preserve the non-negativity property of the original tensor samples, NMPCA aims to define a non-negative multilinear transformation $\{\mathbf{U}_l \in \mathbb{R}_+^{I_l \times P_l}, l = 1, 2, \ldots, N\}$ that maps the original tensor space $\mathbb{R}_+^{I_1 \times I_2 \times \ldots \times I_N}$ onto a non-negative tensor subspace $\mathbb{R}_+^{P_1 \times P_2 \times \ldots \times P_N}$ with $P_l < I_l$, $l = 1, 2, \ldots, N$. That is to derive $\{\mathcal{Y}_i, i = 1, 2, \ldots, n\}$ with $\mathcal{Y}_i = \mathcal{X}_i \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \ldots \times_N \mathbf{U}_N^T$ such that most of the variation observed between the original tensor samples is captured, assuming that the variation can be measured by the total tensor scatter (10). Clearly the objective function to be maximized in NMPCA is the same as in MPCA, i.e.,

$$f_{\text{NMPCA}} \left(\mathbf{U}_l|_{l=1}^N \right) = \mathcal{S}_T = \sum_{i=1}^n \|\mathcal{Y}_i - \mathcal{M}_{\mathcal{Y}}\|^2. \quad (20)$$

Let $\mathcal{M}_{\mathcal{X}}$ be the mean tensor of $\{\mathcal{X}_i, i = 1, 2, \ldots, n\}$. (20) can be reformulated as follows:

$$\begin{aligned} &f_{\text{NMPCA}} \left(\mathbf{U}_l|_{l=1}^N \right) \\ &= \frac{1}{2} \sum_{i=1}^n \left[\left[\left((\mathcal{X}_i - \mathcal{M}_{\mathcal{X}}) \prod_{j=1}^N \times_j \mathbf{U}_j^T \right) \right.\right. \\ &\qquad \otimes \left((\mathcal{X}_i - \mathcal{M}_{\mathcal{X}}) \prod_{j=1}^N \times_j \mathbf{U}_j^T \right), \\ &\qquad \left.\left. (1:N)(1:N) \right]\right] \\ &= \frac{1}{2} \sum_{i=1}^n \left[\left[\left((\mathcal{X}_i - \mathcal{M}_{\mathcal{X}}) \times_{\overline{j=l}} \mathbf{U}_j^T \times_l \mathbf{U}_l^T \right) \right.\right. \\ &\qquad \left. \otimes (\mathcal{X}_i - \mathcal{M}_{\mathcal{X}}) \times_{\overline{j=l}} \mathbf{U}_j^T \times_l \mathbf{U}_l^T \right); \\ &\qquad (1:N)(1:N)]]. \end{aligned} \quad (21)$$

Setting $\mathcal{D}_i \triangleq (\mathcal{X}_i - \mathcal{M}_{\mathcal{X}}) \times_{\overline{j=l}} \mathbf{U}_j^T$, (20) is simplified as

$$\begin{aligned} f_{\text{NMPCA}} \left(\mathbf{U}_l|_{l=1}^N \right) &= \frac{1}{2} \text{tr} \left(\mathbf{U}_l^T \left(\sum_{i=1}^n \mathbf{D}_{i(l)} \mathbf{D}_{i(l)}^T \right) \mathbf{U}_l \right) \\ &= \frac{1}{2} \text{tr} \left(\mathbf{U}_l^T \mathbf{S}_{T(l)} \mathbf{U}_l \right) \end{aligned} \quad (22)$$

where $\mathbf{S}_{T(l)} = \sum_{i=1}^n \mathbf{D}_{i(l)} \mathbf{D}_{i(l)}^T$ is the mode-l unfolding of the total scatter tensor of $\{\mathcal{X}_i, i = 1, 2, \ldots, n\}$. Accordingly, the projection matrices $\{\mathbf{U}_l, l = 1, 2, \ldots, N\}$ are obtained by solving the optimization problem

$$\{\mathbf{U}_l^*, l = 1, 2, \ldots, N\} = \underset{\substack{\mathbf{U}_l \in Gr(I_l, P_l) \\ \mathbf{U}_l \geqslant \mathbf{0}}}{\arg\max} f_{\text{NMPCA}} \left(\mathbf{U}_l|_{l=1}^N \right). \quad (23)$$

The maximization problem (23) is a concave problem [30] and thus it is not possible to find a global maximum. However, since $f_{\text{NMPCA}}(\mathbf{U}_l|_{l=1}^N)$ in (22) is a homogenous function [39], we can define N different mappings $f_{\text{NMPCA}}^{(l)}(\mathbf{U}_l) = f_{\text{NMPCA}}(\mathbf{U}_l; \mathbf{U}_k|_{k=1}^{l-1}, \mathbf{U}_k|_{k=l+1}^N)$, $l = 1, 2 \ldots, N$. The notation $f_{\text{NMPCA}}^{(l)}(\mathbf{U}_l)$ indicates that only $\mathbf{U}_l$ varies, while the matrices $\mathbf{U}_k|_{k=1}^{l-1}$ and $\mathbf{U}_k|_{k=l+1}^N$ are kept fixed. Accordingly, N independent optimization subproblems can be defined as follows:

$$\mathbf{U}_l^* = \arg \max_{\substack{\mathbf{U}_l \in Gr(I_l, P_l) \\ \mathbf{U}_l \geqslant \mathbf{0}}} f_{\text{NMPCA}}^{(l)}(\mathbf{U}_l). \quad (24)$$

The N projection matrices can be computed iteratively by a local optimization procedure in a similar manner to the Alternating Least Squares [29], [37], until a convergency criterion is met or a maximum number of iteration is reached.

It is obvious that the homogeneity condition holds for $f_{\text{NMPCA}}^{(l)}(\mathbf{U}_l)$. Thus, each optimization subproblem (24) has the form of the general optimization problem (12). Consequently, it can be solved using the proposed multiplicative updates (19). The first-order partial derivative of $f_{\text{NMPCA}}^{(l)}(\mathbf{U}_l)$ with respect to $\mathbf{U}_l$ is given by $\nabla_{\mathbf{U}_l} f_{\text{NMPCA}}^{(l)} = (1/2)\mathbf{S}_{T(l)} \mathbf{U}_l$. Invoking (14), the natural gradient of $f_{\text{NMPCA}}^{(l)}(\mathbf{U}_l)$ is calculated as

$$\tilde{\nabla}_{\mathbf{U}_l} f_{\text{NMPCA}}^{(l)} = \frac{1}{2} \mathbf{S}_{T(l)} \mathbf{U}_l - \frac{1}{2} \mathbf{U}_l \mathbf{U}_l^T \mathbf{S}_{T(l)} \mathbf{U}_l. \quad (25)$$

By defining two non-negative matrices $\mathbf{S}_{T(l)}^+$ and $\mathbf{S}_{T(l)}^-$ as in (16) and (17), the mode-l unfolding of the total scatter tensor can be decomposed as $\mathbf{S}_{T(l)} = \mathbf{S}_{T(l)}^+ - \mathbf{S}_{T(l)}^-$. Since $\mathbf{U}_l$ is a non negative matrix, it follows

$$\begin{aligned} \tilde{\nabla}_{\mathbf{U}_l} f_{\text{NMPCA}}^{(l)} = &\underbrace{\frac{1}{2} \left(\mathbf{S}_{T(l)}^+ \mathbf{U}_l + \mathbf{U}_l \mathbf{U}_l^T \mathbf{S}_{T(l)}^- \mathbf{U}_l \right)}_{\left(\tilde{\nabla}_{\mathbf{U}_l} f_{\text{NMPCA}}^{(l)} \right)^+} \\ &- \underbrace{\frac{1}{2} \left(\mathbf{S}_{T(l)}^- \mathbf{U}_l + \mathbf{U}_l \mathbf{U}_l^T \mathbf{S}_{T(l)}^+ \mathbf{U}_l \right)}_{\left(\tilde{\nabla}_{\mathbf{U}_l} f_{\text{NMPCA}}^{(l)} \right)^-}. \end{aligned} \quad (26)$$

Inserting (26) into (19) and setting $\mathbf{U}_l = \mathbf{U}_l^t$, the multiplicative update for each projection matrix $\mathbf{U}_l$ is given by

$$\mathbf{U}_l^{t+1} = \mathbf{U}_l * \frac{\mathbf{S}_{T(l)}^+ \mathbf{U}_l + \mathbf{U}_l \mathbf{U}_l^T \mathbf{S}_{T(l)}^- \mathbf{U}_l}{\mathbf{S}_{T(l)}^- \mathbf{U}_l + \mathbf{U}_l \mathbf{U}_l^T \mathbf{S}_{T(l)}^+ \mathbf{U}_l + \epsilon}. \quad (27)$$

Let $\mathbf{u}_{l_j}$ indicates the jth column of $\mathbf{U}_l$. The solution of the N subproblems is iteratively repeated until the following global convergence criterion is met

$$\sum_{l=1}^{N}\sum_{j}\left|\mathbf{u}_{l_j}^{t\;T}\mathbf{u}_{l_j}^{t-1}-1\right|\leq\epsilon \tag{28}$$

which checks the stationarity of the solution $\mathbf{U}_1^t,\ldots,\mathbf{U}_N^t$. The convergence criterion in (28) indicates that the value of f_{NMPCA} between two successive iterations is quite small. That is, the procedure halts, when f_{NMPCA} admits a local maximum.

NMPCA Interpretation: Ding *et al.* have proved that the optimization subproblem (24) is equivalent to k-means clustering [43]. Since the N projection matrices $\mathbf{U}_l$ in NMPCA are obtained by solving N optimization sub-problems (24), one for each mode-l matrix, NMPCA could be interpreted as simultaneous mode-l k-means clustering. Accordingly, $\mathbf{U}_l$ is the mode-l cluster indicator matrix.

Convergence Analysis: The convergence of NMPCA algorithm is governed by Theorem 2. The proof of Theorem 2 is based on the global convergence theorem [44] and depends on several lemmata. The proofs of Lemma 3, Lemma 4, and Theorem 2 can be found in the Appendix.

Theorem 2: The limit point of any convergent subsequence of $\{\mathbf{U}_1^t,\ldots,\mathbf{U}_N^t\}$ generated by (27) is a stationary point of the optimization problem (22). □

Lemma 1: If $\{\mathbf{U}_l^t\}$ is the sequence of updates generated by (27) then

$$f_{\mathrm{NMPCA}}^{(l)}\left(\mathbf{U}_l^{t+1}\right)\geqslant f_{\mathrm{NMPCA}}^{(l)}\left(\mathbf{U}_l^{t}\right). \tag{29}$$

□

Proof: It is straightforward from Theorem 1.

Lemma 2: If $\{\mathbf{U}_1^t,\ldots,\mathbf{U}_N^t\}$ is an infinite sequence, generated by the alternating updates (27), then

$$f_{\mathrm{NMPCA}}\left(\mathbf{U}_1^{t+1},\ldots,\mathbf{U}_N^{t+1}\right)\geqslant f_{\mathrm{NMPCA}}\left(\mathbf{U}_1^{t},\ldots,\mathbf{U}_N^{t}\right). \tag{30}$$

□

Proof: Straightforward by applying Lemma 4.1.

Lemma 3: Let $\{\mathbf{U}_1^t,\ldots,\mathbf{U}_N^t\}$ be an infinite sequence generated by the alternating updates (27). This sequence lies on a compact set. □

Lemma 4: Let $\Omega_l : Gr(I_l,P_l)\mapsto Gr(I_l,P_l)$ be the mapping that generates $\mathbf{U}_l^{t+1}$ from $\mathbf{U}_l^t$. That is, $\mathbf{U}_l^{t+1}=\Omega_l(\mathbf{U}_l^t)$. Ω_l is closed. □

NMPCA Computational Complexity: Let $\{\mathcal{X}_i\in\mathbb{R}_+^{I_1\times I_2\times\ldots\times I_N}, i=1,2,\ldots,n\}$ be a set of n tensor samples. Assume $I_1=I_2=\ldots=I_N=I$, for simplicity. The time complexity of the iterative NMPCA algorithm during the training phase is as follows. At each iteration, the time complexity of computing the mode-l unfolding of the total scatter tensor, $\mathbf{S}_{T(l)}$, has an upper bound of the order $\mathcal{O}(n\;N\;I^{(N+1)})$, while the time complexity of the N multiplicative updates rule in (27) is in order of $\mathcal{O}(N\;I^3)$. The NMPCA algorithm does not require all the training tensor samples to be kept in the memory, since it can compute $\mathbf{S}_{T(l)}$ incrementally by loading each $\mathcal{X}_i$ sequentially. Hence, the memory requirements of NMPCA is of the order $\mathcal{O}(I^N)$. This is a significant advantage of NMPCA comparing to subspace analysis algorithms, such as NTF and HOSVD, which require the entire training set to be loaded in the memory.

VI. Experimental Evaluation

The projection matrices $\{\mathbf{U}_l, l=1,2,\ldots,N\}$ obtained by NMPCA from a set of non-negative training tensor samples $\{\mathcal{X}_i, i=1,2,\ldots,n\}$, can be used to extract features. In order to assess the discriminating power of both auditory temporal modulations and NMPCA, experiments in automatic music genre classification were conducted. To compare the classification accuracy obtained with that of state of the art music genre classification algorithms, two different experimental settings were considered. In addition, a third set of experiments was conducted in order to simulate the case when the training set has small cardinality.

A. Datasets and Preprocessing

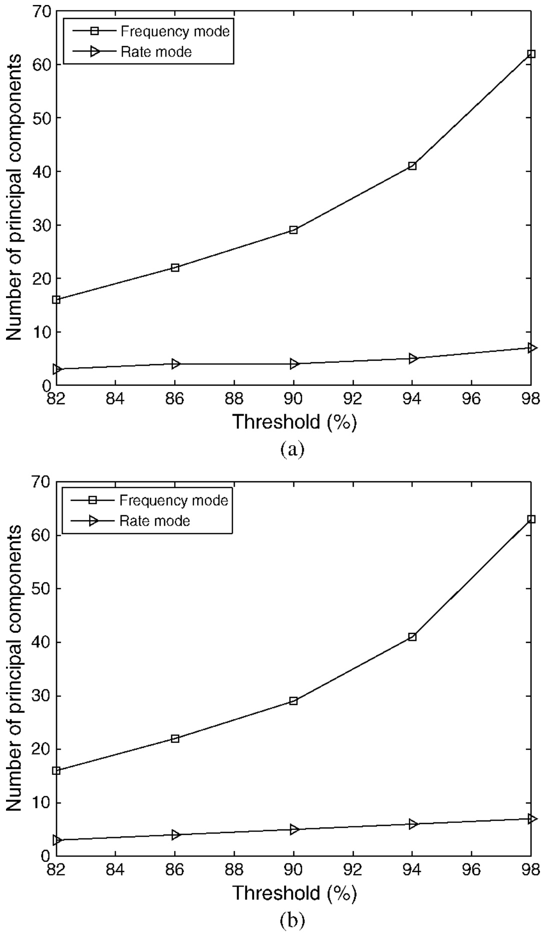
Experiments were performed on two widely used datasets for music genre classification [3]–[7], [10]. The first dataset, abbreviated as GTZAN, was collected by Tzanetakis [3] and consists of ten genre classes, namely Blues, Classical, Country, Disco, HipHop, Jazz, Metal, Pop, Reggae, and Rock. Each genre class contains 100 audio recordings 30 s long. The second dataset, abbreviated as ISMIR 2004 Genre, comes from the ISMIR 2004 Genre classification contest and contains 1458 full audio recordings distributed over six genre classes as follows. Classical (640), Electronic (229), JazzBlues (52), MetalPunk (90), RockPop (203), World (244), where the number within parentheses refers to the number of recordings which belong to each genre class.

All the audio recordings were converted to monaural wave format at a sampling frequency of 16 kHz and quantized with 16 bits. Moreover, the audio signals have been normalized, so that they have zero mean amplitude with unit variance, in order to remove any factors related to the recording conditions. Since the ISMIR 2004 Genre dataset, consists of full length tracks, we extracted a segment of 30 s just after the first 30 s of a recording to exclude any introductory parts that may not be directly related to the music genre the recording belongs to. The auditory temporal modulations representation is computed over a segment of 30-s duration for any recording of both datasets.

B. Evaluation Procedure and Experimental Results

Following the experimental setup used in [3], [6], [8], and [9], stratified tenfold cross validation is employed for experiments conducted on the GTZAN dataset. Thus each training set consists of 900 audio files. By stacking the associated auditory temporal modulations a training tensor $\mathcal{X}_{\mathrm{GTZAN}}\in\mathbb{R}_+^{I_1\times I_2\times I_3}$ is constructed, where $I_1=I_{\mathrm{frequency}}=96$, $I_2=I_{\mathrm{rates}}=8$, and $I_3=I_{\mathrm{samples}}=900$.

The experiments on ISMIR 2004 Genre dataset were conducted according to the ISMIR2004 Audio Description Contest protocol. The protocol defines training and evaluation sets, which consist of 729 audio files each. Thus the corresponding



(a)

(b)

Fig. 3. Total number of retained principal components in each mode (e.g., frequency and rate) as a function of the portion of total scatter retained (threshold) for the (a) GTZAN dataset and (b) ISMIR 2004 Genre dataset.

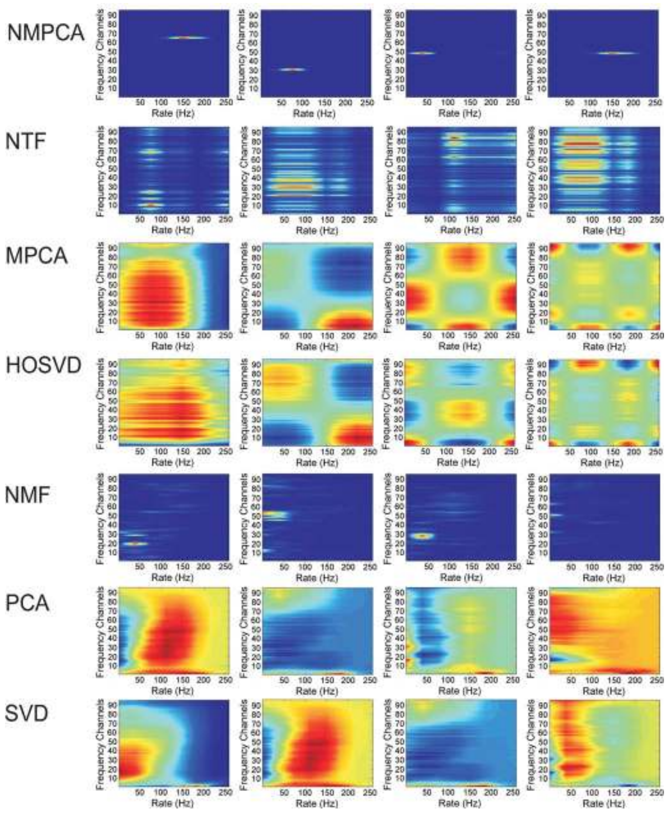


Fig. 4. Bases obtained by applying various subspace analysis methods to $\mathcal{X}_{\text{GTZAN}}$, when 98% of the total scatter is retained.

training tensor $\mathcal{X}_{\text{ISMIR}} \in \mathbb{R}_{+}^{I_1 \times I_2 \times I_3}$ is constructed for $I_1 = I_{\text{frequency}} = 96$, $I_2 = I_{\text{rates}} = 8$, and $I_3 = I_{\text{samples}} = 729$.

Compact feature vectors were extracted by applying NMPCA, NTF, MPCA, PCA, NMF, and SVD on each training tensor, as described in Section IV and [8].

In order to determine the reduced dimensions of tensors after subspace projections, the ratio of the sum of eigenvalues retained over the sum of all eigenvalues of each mode-l tensor unfolding is employed as in [29]. By using this ratio as a specification threshold, the number of retained principal components for each mode (e.g., frequency and rate) was determined, as is demonstrated in Fig. 3 for the GTZAN and the ISMIR Genre 2004 datasets. The different subspace analysis methods are compared for equal dimensionality reduction. That is, the same $P_1 = P_{\text{frequency}}$ and $P_2 = P_{\text{rates}}$ were used in MPCA, HOSVD, and NMPCA, while $k = P_1 P_2$ for PCA, SVD, NTF, and NMF. Several bases obtained by each subspace analysis method are visualized in Fig. 4. Classification was performed by the SVM with an RBF and a linear kernel. In order to tune the RBF kernel parameters, a grid search algorithm similar to the algorithm proposed in [45] was used. In addition, the NN classifier with three commonly used distances defined in Table II was tested. Given a test vector $\mathbf{y}$, its distance $d(\mathbf{y}, \mathbf{x}_i)$ from all $\mathbf{x}_i$, which belong to the training set, is calculated. The label assigned to $\mathbf{y}$ is that of $\mathbf{x}_i^*$ being closer to $\mathbf{y}$, i.e., $d(\mathbf{y}, \mathbf{x}_i^*)$ is minimal.

In Figs. 5 and 6 the classification accuracy achieved by the NN classifier and the SVMs, when the various subspace analysis methods are employed on both GTZAN and ISMIR 2004 Genre datasets, is plotted as a function of the threshold. The best classification results for each classifier and each dataset are summarized in Table III. Especially for the GTZAN dataset, the best classification results reported in Table III were calculated by applying tenfold stratified cross-validation.

For both GTZAN and ISMIR 2004 Genre datasets, the best classification accuracy obtained by auditory temporal modulations without any dimensionality reduction was achieved by the SVM with an RBF kernel and was equal to 77.09% and 78.64%, respectively. Therefore, the auditory temporal modulations carry a significant amount of information about music genre, outperforming many music genre classification algorithms based on BOF approach [3], [5], [9].

For both datasets, the classification accuracies obtained by the multilinear subspace analysis techniques outperform those obtained by their linear counterparts. For example, the best classification accuracies were obtained by the SVM with an RBF kernel. The SVM with a linear kernel did not perform well for features extracted by the subspace analysis methods.

On the GTZAN dataset the best classification accuracy (84.3%) was obtained when NMPCA extracts features that are

TABLE II
THREE DIFFERENT DISTANCES EMPLOYED IN THE NN CLASSIFIER

Distance	L_1	L_2	CSM
$d(\mathbf{y},\mathbf{x})$	$\sum_{h=1}^{H} \lvert(\mathbf{y})_h - (\mathbf{x})_h\rvert$	$\sqrt{\sum_{h=1}^{H}[(\mathbf{y})_h - (\mathbf{x})_h]^2}$	$-\frac{\sum_{h=1}^{H}(\mathbf{y})_h\,(\mathbf{x})_h}{\sqrt{\sum_{h=1}^{H}(\mathbf{y})_h^2 \sum_{h=1}^{H}(\mathbf{x})_h^2}}$

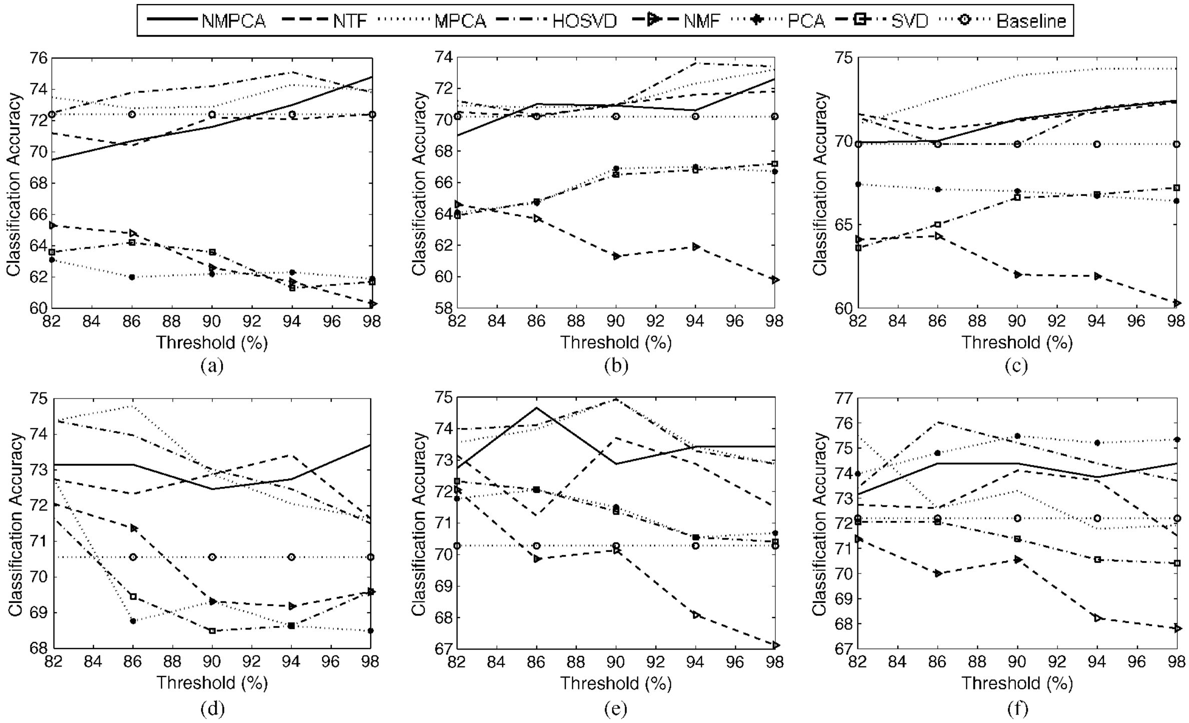


Fig. 5. Classification accuracy for the various subspace analysis methods used to extract features that are classified next by the NN classifier for various distances. The accuracy is calculated by ten n-fold stratified cross-validation. Classification accuracy on GTZAN dataset obtained by NN classifier with (a) L_1 distance, (b) L_2 distance, (c) CSM distance. Classification accuracy on ISMIR2004 Genre dataset obtained by NN classifier with (d) L_1 distance, (e) L_2 distance, (f) CSM distance.

classified by SVM with an RBF kernel. The reported classification accuracy outperforms those listed in Table I.

On the ISMIR 2004 genre dataset the best classification accuracy (83.15%) was obtained when NTF extracts features that are classified by SVM with an RBF kernel. Again, the achieved classification accuracy outperforms all previously reported rates as shown in Table I. The classification accuracy achieved when NMPCA extracts features, that are classified by the SVM with an RBF kernel equals 82.19% and is comparable to the previous best classification rate obtained by Pampalk *et al.* [7] on this dataset. It is not possible to compare directly our results with the results obtained by Holzapfel *et al.* in [5] and Panagakis *et al.* in [8], on this dataset, because of the quite different experimental setup that was used.

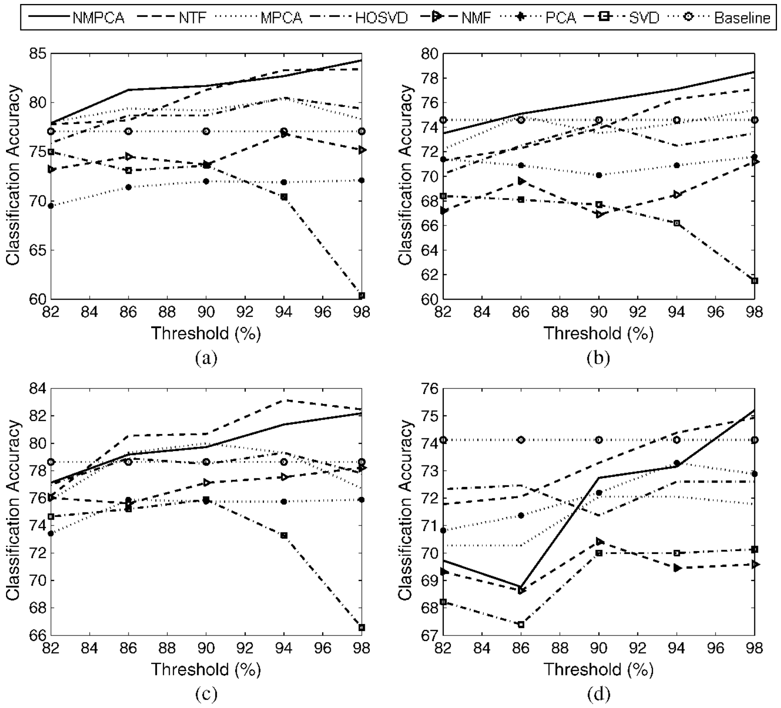
C. Experimental Results on a SSS Problem

In many real applications, both commercial and private, the number of available audio recordings per genre is limited. In order to investigate the performance of NMPCA against the other subspace analysis techniques, five different small training subsets were extracted from the GTZAN dataset. The first training set consist of the 10% of the total number of files contained in the GTZAN dataset, i.e., 100 recordings. The remaining 900 songs were used for testing. Similarly, the other four training sets consist of 20%, 30%, 40%, and 50% of the total number of files in the GTZAN dataset, respectively.

The classification accuracies obtained in this experiment are plotted in Fig. 7 as a function of the number of training samples employed. From Fig. 7, it is obvious that the NMPCA outperforms the other subspace analysis techniques either multilinear or linear ones when it is used to extract features that are next classified by NN and SVM classifiers. The classification accuracy obtained when NMPCA selects auditory temporal modulations that are next classified by the SVM with an RBF kernel exceeds 70% even when 100 training samples are exploited. This observation further supports the claim that the proposed representation of auditory temporal modulations when combined with the NMPCA has a potential for a viable music genre classification in real world conditions.

VII. CONCLUSION AND FUTURE WORK

Two-dimensional auditory temporal modulations have been proposed for music representation. Furthermore, a novel unsupervised multilinear subspace analysis method, the NMPCA, has been derived in order to preserve the non-negativity of N-order tensor representations. An algorithm for NMPCA has been developed by exploiting the structure of the Grassmann manifold. The NMPCA has been applied to the auditory temporal modulations in order to extract features of reduced



(a) (b) (c) (d)

Fig. 6. Classification accuracy for the various subspace analysis methods used to extract features that are classified next by the SVM classifier with an RBF and a linear kernel. The accuracy is calculated by tenfold stratified cross-validation. Classification accuracy on GTZAN dataset obtained by (a) SVM with RBF kernel and (b) SVM with linear kernel. Classification accuracy on ISMIR 2004 Genre dataset obtained by (c) SVM with RBF kernel and (d) SVM with linear kernel.

TABLE III
BEST CLASSIFICATION ACCURACIES FOR BOTH DATASETS

Classifier	Kernel/Distance	Dataset	Best Accuracy	Method	Dataset	Best Accuracy	Method
NN	L_1	GTZAN	74.8%	NMPCA	ISMIR2004Genre	74.79%	MPCA
NN	L_2	GTZAN	73.6%	HOSVD	ISMIR2004Genre	74.93%	MPCA/HOSVD
NN	CSM	GTZAN	74.3%	MPCA	ISMIR2004Genre	76.03%	HOSVD
SVM	RBF	GTZAN	84.3%	NMPCA	ISMIR2004Genre	83.15%	NTF
SVM	Linear	GTZAN	78.5%	NMPCA	ISMIR2004Genre	75.2%	NMPCA
NN	L_1	GTZAN	65.3%	NMF	ISMIR2004Genre	72.73%	PCA
NN	L_2	GTZAN	67.2%	SVD	ISMIR2004Genre	72.32%	SVD
NN	CSM	GTZAN	67.4%	PCA	ISMIR2004Genre	75.47%	PCA
SVM	RBF	GTZAN	76.8%	NMF	ISMIR2004Genre	78.21%	NMF
SVM	Linear	GTZAN	71.6%	PCA	ISMIR2004Genre	73.28%	PCA

dimensionality for music genre classification. The efficiency of the extracted features for music genre classification has been demonstrated. Moreover, the gains of multilinear subspace analysis techniques against the linear counterparts have been shown. The NMPCA outperforms the other subspace analysis methods for a SSS music genre classification problem highlighting the potential of the proposed method for viable practical music genre classification systems.

The multilinear dimensionality reduction techniques employed in this paper are unsupervised. In the future, supervised multilinear subspace analysis techniques based on preserving the non-negativeness of bio-inspired auditory representations will be developed and tested for music genre classification. In our experiments, we have considered that each song belongs exclusively to only one genre class. Obviously, it is more realistic to use overlapping class labels, e.g., labeling music by style [4]. In general, high-order tensors are structures that are suitable for a such multi-labeling classification problem.

Appendix

Proof of Lemma 3: In order to prove that the sequence $\{\mathbf{U}_1^t, \ldots, \mathbf{U}_N^t\}$ lies on a compact set, it suffices to prove that $\{\mathbf{U}_l^t\}$, for $l = 1, 2, \ldots, N$ is both closed and bounded. Since sequence $\{\mathbf{U}_l^t\}$ lies on the Grassmann manifold $Gr(I_l, P_l)$ in order to prove that $\{\mathbf{U}_l^t\}$, for $l = 1, 2, \ldots, N$ is closed it suffices to prove that the Grassmann manifold $Gr(I_l, P_l)$ is closed. Let $\phi(\mathbf{U}_l) = \mathbf{U}_l^T \mathbf{U}_l$. $\phi(\mathbf{U}_l)$ is a continuous function as product of two continuous functions, namely $\phi_1(\mathbf{U}_l) = \mathbf{U}_l^T$ and $\phi_2(\mathbf{U}_l) = \mathbf{U}_l$. By definition $Gr(I_l, P_l) = \phi^{-1}(\mathbf{I})$ and since $\mathbf{I}$ is closed and $\phi(\mathbf{U}_l)$ is continuous, $Gr(I_l, P_l)$ is closed, because it is the inverse image of a closed set. Consequently, $\{\mathbf{U}_l^t\}$ is a closed set.

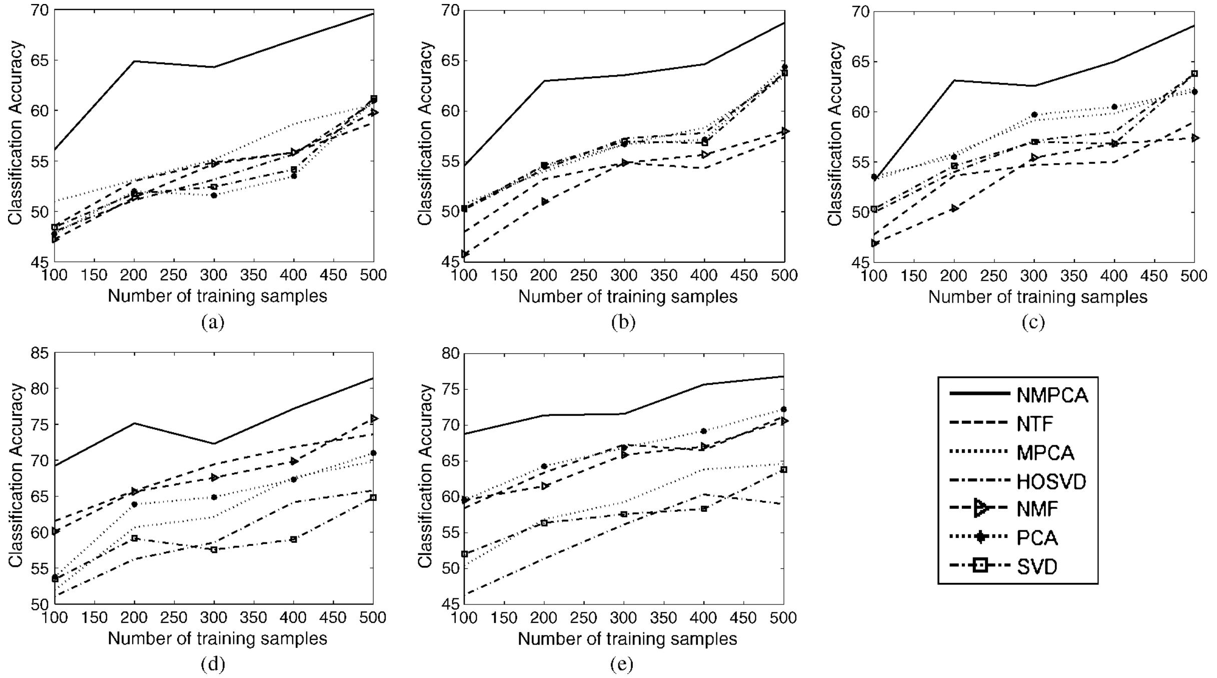
Let $\mathbf{U}_l = [\mathbf{u}_{l_1} | \mathbf{u}_{l_2} | \ldots | \mathbf{u}_{l_{P_l}}]$. Since $\mathbf{U}_l$ is an orthonormal matrix, each column $\mathbf{u}_{l_j}$, $j = 1, 2, \ldots, P_l$ has unit length. Thus, $||\mathbf{U}_l||^2 = P_l$. Consequently, the norm of every orthonormal matrix is $\sqrt{P_l}$ and thus $\{\mathbf{U}_l^t\}$ is bounded.

Therefore, $\{\mathbf{U}_1^t, \ldots, \mathbf{U}_N^t\}$ being closed and bounded, lies on a compact set, as union of compact sets.

Proof of Lemma 4: In Lemma 3, we have proven that $Gr(I_l, P_l)$ is closed. Since the update rule (27) is a continuous function, the mapping Ω_l is closed.

Proof of Theorem 2: The following conditions hold.

1) Let $\Omega = \Omega_1 \circ \Omega_2 \circ \ldots \circ \Omega_N$ be the NMPCA algorithm illustrated as a composition of N sub-algorithms. Since



(a) (b) (c) (d) (e)

Fig. 7. Classification accuracy for small numbers of training samples when various subspace analysis methods extract features that are next classified by either an SVM or a NN classifier, which employs different distance measures. Classification accuracy obtained by NN classifier with (a) NN classifier with L_1 distance, (b) NN classifier with L_2 distance, (c) NN classifier with CSM distance. (d) SVM with an RBF kernel, and (e) SVM with a linear kernel.

from Lemma 4, Ω_l, $l = 1, 2, \ldots, N$ is a closed algorithm, Ω is closed too.

2) By Lemma 3, the infinite sequence $\{\mathbf{U}_1^t, \ldots, \mathbf{U}_N^t\}$, generated by the alternating updates (27), lies on a compact set.

3) Furthermore, Ω strictly increases the objective function f_{NMPCA} unless a solution is reached according to Lemma 2.

Consequently, the conditions of the general convergence theorem are met. Thus, the limit point of any convergent subsequence of $\{\mathbf{U}_1^t, \ldots, \mathbf{U}_N^t\}$ generated by (27) is a stationary point of the optimization problem (22).

References

[1] J. J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *J. New Music Res.*, pp. 83–93, 2003.

[2] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 133–141, Mar. 2006.

[3] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[4] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kegl, "Aggregate features and ADABOOST for music classification," *Mach. Learn.*, vol. 65, no. 2–3, pp. 473–484, 2006.

[5] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 424–434, Feb. 2008.

[6] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proc. 26th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2003, pp. 282–289.

[7] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proc. 6th Int. Symp. Music Inf. Retrieval*, 2005.

[8] I. Panagakis, E. Benetos, and C. Kotropoulos, "Music genre classification: A multilinear approach," in *Proc. 9th Int. Symp. Music Inf. Retrieval*, 2008, pp. 583–588.

[9] T. Lidy, A. Rauber, A. Pertusa, and J. Inesta, "Combining audio and symbolic descriptors for music classification from audio," in *Proc. Music Inf. Retrieval Inf. Exchange (MIREX)*, 2007.

[10] T. Lidy and A. Rauber, "Evaluation of feature extractors and psychoacoustic transformations for music genre classification," in *Proc. 6th Int. Symp. Music Inf. Retrieval*, London, U.K., 2005.

[11] E. Benetos and C. Kotropoulos, "A tensor-based approach for automatic music genre classification," in *Proc. 16th Eur. Signal Process. Conf.*, 2008.

[12] M. Mandel and D. Ellis, "LABROSA's audio music similarity and classification submissions," in *Proc. Music Inf. Retrieval Inf. Exchange (MIREX)*, 2007.

[13] J. J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?," *J. Negative Results in Speech Audio Sci.* , vol. 1, no. 1, 2004.

[14] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 3023–3035, October 2004.

[15] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer function and speech intelligibility," *J. Acoust. Soc. Amer.*, no. 5, pp. 2719–2732, Nov. 1999.

[16] S. D. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *J. Acoust. Soc. Amer.*, vol. 108, pp. 1181–1196, 2000.

[17] M. Slaney and R. F. Lyon, "On the importance of time-a temporal representation of sound," in *Visual Representations of Speech Signals*, M. Cooke, S. Beet, and M. Crawford, Eds. New York: Wiley, 1993, pp. 95–116.

[18] N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," *J. Acoust. Soc. Amer.*, vol. 114, no. 6, pp. 3394–3411, 2003.

[19] S. Woolley, T. Fremouw, A. Hsu, and F. Theunissen, "Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds," *Nature Neurosci.*, vol. 8, no. 10, pp. 1371–1379, 2005.

[20] L. De Lathauwer, "Signal processing based on multilinear algebra," Ph.D. dissertation, Faculty of Eng., K. U. Leuven, Leuven, Belgium, Sep. 1997.

[21] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 920–930, May 2006.

[22] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[23] J. Liu, S. Chen, and X. Tan, "A study on three linear discriminant analysis based methods in small sample size problem," *Pattern Recognition*, vol. 41, no. 1, pp. 102–116, 2008.

[24] J. Frankel and S. King, "Factoring Gaussian precision matrices for linear dynamic models," *Pattern Recogn. Lett.* , vol. 28, no. 16, pp. 2264–2272, 2007.

[25] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[26] M. Welling and M. Weber, "Positive tensor factorization," *Pattern Recognition Lett.*, vol. 22, no. 12, pp. 1255–1261, 2001.

[27] A. Shashua and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 792–799.

[28] A. Cichocki, R. Zdunek, and S. Amari, "Nonnegative matrix and tensor factorization," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 142–145, Jan. 2008.

[29] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 18–39, Jan. 2008.

[30] R. Zass and A. Shashua, "Nonnegative sparse PCA," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2007.

[31] S. Greenberg, E. D. Brian, and Y. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, pp. 1647–1650.

[32] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 824–839, Mar. 1992.

[33] S. A. Shamma, "Encoding sound timbre in the auditory system," *IETE J. Res.*, vol. 49, no. 2–3, pp. 145–156, 2003.

[34] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, Sep. 2008.

[35] J. Duchene and S. Leclercq, "An optimal transformation for discriminant and principal component analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 6, pp. 978–983, Nov. 1988.

[36] L. Lathauwer, B. D. Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM J. Matrix Anal. Applicat.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[37] P. Kroonenberg and J. Leeuw, "Principal component analysis of three-mode data by means of alternating least squares algorithms," *Psychometrika*, vol. 45, no. 1, pp. 69–97, Mar. 1980.

[38] S. I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.

[39] A. Edelman, A. T. Arias, and T. S. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Applicat.*, vol. 20, no. 2, pp. 303–353, 1999.

[40] E. Lars and S. Berkant, A Newton–Grassmann method for computing the best multi-linear rank-(r_1, r_2, r_3) approximation of a tensor Dept. of Math., Linköpings Univ., 2007, Tech. Rep..

[41] Z. Yang and J. Laaksonen, "Multiplicative updates for non-negative projections," *Neurocomputing*, vol. 71, no. 1–3, pp. 363–373, 2007.

[42] J. Yoo and S. Choi, "Orthogonal nonnegative matrix factorization: Multiplicative updates on Stiefel manifolds," in *Proc. 9th Int. Conf. Intell. Data Eng. and Autom. Learn. (IDEAL)*, 2008.

[43] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix tri-factorizations for clustering," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2006, pp. 126–135.

[44] D. Luenberger*, Linear and Nonlinear Programming*, 3rd ed. New York: Springer, 2008.

[45] C. Hsu, C. C. Chang, and C. J. Lin, "A Practical guide to support vector classification," Dept. of Comput. Sci., National Taiwan Univ., 2003, Tech. Rep..

Yannis Panagakis was born in Grevena, Greece. He received the B.Sc. degree in informatics and telecommunication from the National and Kapodistrian University of Athens, Athens, Greece, and the M.Sc. degree in digital media from the department of informatics, Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece. He is currently pursuing the Ph.D. degree at the Department of Informatics, AUTH.

His current research interests include audio signal processing, music information retrieval, computational intelligence, and numerical linear and multilinear algebra.

Constantine Kotropoulos (SM'06) was born in Kavala, Greece, in 1965. He received the Diploma degree (with honors) in electrical engineering and the Ph.D. degree in electrical and computer engineering from the Aristotle University of Thessaloniki, in 1988 and 1993, respectively.

He is currently an Associate Professor in the Department of Informatics at the Aristotle University of Thessaloniki. From 1989 to 1993, he was a Research and Teaching Assistant in the Department of Electrical and Computer Engineering at the same university. In 1995, he joined the Department of Informatics at the Aristotle University of Thessaloniki as a Senior Researcher and served then as a Lecturer from 1997 to 2001 and as an Assistant Professor from 2002 to 2007. He was a Visiting Research Scholar in the Department of Electrical and Computer Engineering at the University of Delaware, Newark, during the academic year 2008–2009 and he conducted research in the Signal Processing Laboratory at Tampere University of Technology, Tampere, Finland, during the summer of 1993. He has coauthored 42 journal papers, 147 conference papers, and contributed six chapters to edited books in his areas of expertise. He is coeditor of the book *Nonlinear Model-Based Image/Video Processing and Analysis* (Wiley, 2001). His current research interests include audio, speech, and language processing; signal processing; pattern recognition; multimedia information retrieval; biometric authentication techniques, and human-centered multimodal computer interaction.

Prof. Kotropoulos was a scholar of the State Scholarship Foundation of Greece and the Bodossaki Foundation. He is a member of EURASIP, IAPR, and the Technical Chamber of Greece. He is a member of the Editorial Board of the *Advances in Multimedia* journal and serves as a EURASIP local liaison officer for Greece.

Gonzalo R. Arce (M'82–SM'93–F'00) received the Ph.D. degree from Purdue University, West Lafayette, IN, in 1982.

Since 1982, he has been with the faculty of the Department of Electrical and Computer Engineering, University of Delaware, Newark, where he is the Charles Black Evans Distinguished Professor. His research interests include statistical and nonlinear signal processing and their applications. He is author or coauthor of the textbooks *Modern Digital Halftoning* (Marcel Dekker, 2001), *Modern Digital Halftoning 2nd edition* (CRC, 2006), *Nonlinear Signal Processing and Applications* (CRC, 2003), *Nonlinear Signal Processing: A Statistical Approach* (Wiley, 2004), and *Resolution Enhancement Optimization: In Optical Lithography* (Wiley, to appear). He is a frequent consultant to industry and holds ten U.S. patents.

Prof. Arce was elected Fellow of the IEEE for his contributions on nonlinear signal processing. Dr. Arce served as an Associate Editor for several IEEE and OSA journals.