CrossMark

Class-specific discriminant time-frequency analysis using novel jointly learnt non-negative matrix factorization

Behnaz Ghoraani

Abstract

Time-frequency (TF) representation has found wide use in many challenging signal processing tasks including classification, interference rejection, and retrieval. Advances in TF analysis methods have led to the development of powerful techniques, which use non-negative matrix factorization (NMF) to adaptively decompose the TF data into TF basis components and coefficients. In this paper, standard NMF is modified for TF data, such that the improved TF bases can be used for signal classification applications with overlapping classes and data retrieval. The new method, called jointly learnt NMF (JLNMF) method, identifies both distinct and shared TF bases and is able to use the decomposed bases to successfully retrieve and separate the class-specific information from data. The paper provides the framework of the proposed JLNMF cost function and proposes a projected gradient framework to solve for limit point stationarity solutions. The developed algorithm has been applied to a synthetic data retrieval experiment and epileptic spikes in EEG signals of infantile spasms and discrimination of pathological voice disorder. The experimental results verified that JLNMF successfully identified the class-specific information, thus enhancing data separation performance.

Keywords: Time-frequency analysis, Non-negative matrix factorization, Data retrieval, Data localization

1 Introduction

Automated decision-making systems have to deal with natural signals that are mostly nonstationary in nature, meaning that their statistics vary over time and frequency. Time-frequency (TF) analysis [1, 2] has been widely used over the past decades to characterize the nonstationary content of signals in the joint time and frequency domain. However, the main limitation of the TF-methods is the high dimensionality of the representation. Therefore, in order to effectively analyze natural signals in artificial intelligence and decision-making applications, it is necessary to develop dimensionality reduction techniques that can remove the redundant data and capture the essential features of the TF representation. Among such techniques, TF decomposition is an effective approach, where a matrix decomposition scheme is used to adaptively decompose the TF data into representing TF bases for further analysis [3–9].

Given the non-negative nature of TF data, non-negative matrix (NMF) decomposition has received favorable attention and been widely used for TF feature extraction applications. Several variants of NMF have been proposed over the past decade. Those methods impose additional constraints, such as localization or sparsity, on the NMF cost function to construct improved NMF decompositions. For example, Zafeiriou et al. [10] and Nikitidis et al. [11] have proposed a discriminative NMF, which adds additional constraints to maximize the between-class scatter and minimize the within-class scatter in the subspace spanned by the bases in order to improve the discrimination power of the decomposed TF bases. However, those methods are designed to classify the data to only one class at a time, and cannot separate the information as needed in data retrieval applications. In this paper, we developed a jointly learnt NMF (JLNMF) method to jointly learn the class-specific TF bases that are discriminant to each class as well as the ones that are shared between classes, and use the discriminant class-specific TF bases to identify how much of each class presents in a test data. To make this happen, a set of new constraints

Correspondence: bghoraani@fau.edu
Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA

is enforced on the NMF cost function as explained in the present work.

The rest of the paper is organized as follows. The background and theoretical framework of the proposed algorithm is reviewed in Section 2. Section 3 introduces the proposed JLNMF algorithm and develops an optimization strategy to solve the proposed JLNMF cost function. Section 4 investigates the properties of the proposed JLNMF algorithm and presents the conducted experimental study and verifies the efficiency of JLNMF for data retrieval and information localization applications. Finally, concluding remarks are drawn in Section 5.

2 Theoretical framework

2.1 Time-frequency matrix decomposition

The TF representation of a signal simultaneously represents the signal's energy distribution over time and frequency domains. For a signal, $x(m)$, its TF representation can be denoted by a matrix $\mathbf{V}_{N \times M}$, where M is the number of time samples in signal $x(m)$ and N is the number of frequency bins (i.e., frequency resolution). For example, the value of $\mathbf{V}(n, m)$, for $n = 1, \ldots, N$ and $m = 1, \ldots, M$, indicates the amount of signal energy at time sample m and frequency of n. The application of a matrix decomposition technique with r coefficients on the TF matrix of $\mathbf{V}$ can be written as following:

$$\mathbf{V}_{N \times M} = \mathbf{W}_{N \times r} \mathbf{H}_{r \times M} = \sum_{i=1}^{r} w_i h_i^T$$

$$\text{s.t. a given constraint on } \mathbf{W} \text{ and } \mathbf{H} \qquad (1)$$

where the decomposed TF matrices, $\mathbf{W}_{N \times r}$ and $\mathbf{H}_{r \times M}$ (with $r << N, M$) contain the TF bases and the corresponding coefficients of the linear combinations of the TF bases required to reconstruct each TF data, respectively, and are defined as $\mathbf{W}_{N \times r} = [w_1 \ w_2 \ \cdots \ w_r]$ and $\mathbf{H}_{r \times M} = [h_1 \ h_2 \ \cdots \ h_r]^T$. In Eq. (1), TF matrix ($\mathbf{V}$) is reduced to base and coefficient vectors ($\{w_i\}_{i=1,\ldots,r}$ and $\{h_i^T\}_{i=1,\ldots,r}$, respectively) subject to a given constraint on $\mathbf{W}$ and $\mathbf{H}$.

Among well-known matrix decomposition techniques (e.g., principal component analysis, independent component analysis, and NMF), NMF is the only technique that guarantees the non-negativity of the decomposed matrices (i.e., $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$) and has been widely used in TF decomposition applications. NMF derives the TF bases and coefficients (i.e., $\mathbf{W}$ and $\mathbf{H}$, respectively) by minimizing the least-square cost function as shown below:

$$\min_{\mathbf{W}, \mathbf{H}} f(\mathbf{W}, \mathbf{H}) = \|\mathbf{V} - \mathbf{W} \mathbf{H}\|_F^2, \qquad (2)$$

$$\text{subject to } \mathbf{W} \geq 0, \mathbf{H} \geq 0$$

NMF optimization starts with an initial estimate of $\mathbf{W}$ and $\mathbf{H}$, and performs an iterative optimization to minimize

the cost function in Eq. (2). In [12], Lee and Seung introduce two updating algorithms using the least square error and the Kullback-Leibler (KL) divergence as the cost functions. The equations, using the least square error method, are as follows:
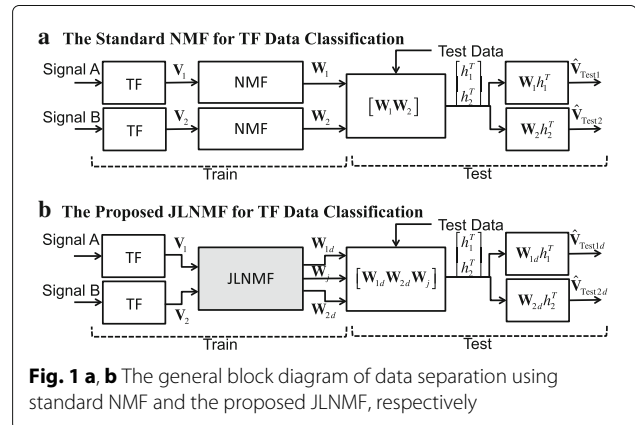
$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} \cdot \frac{\mathbf{V} \mathbf{H}^{(t)^T}}{\mathbf{W}^{(t)} \mathbf{H}^{(t)} \mathbf{H}^{(t)^T}}, \text{ and } \mathbf{H}^{(t+1)}$$

$$= \mathbf{H}^{(t)} \cdot \frac{\mathbf{W}^{(t+1)^T} \mathbf{V}}{\mathbf{W}^{(t+1)^T} \mathbf{W}^{(t+1)} \mathbf{H}^{(t)}}, \qquad (3)$$

In these equations, $\langle \ . \ \rangle$ and $\frac{\langle \ \rangle}{\langle \ \rangle}$ are element-wise multiplication and division of two matrices. Various alternative minimization strategies for NMF decomposition have been proposed [12–14]. In this work, we use a projected gradient bound-constrained optimization approach in [15]. The gradient-based NMF is computationally competitive and offers better convergence properties than [12].

2.2 Need for jointly learnt NMF (JLNMF) method

In order to apply the TF matrix decomposition approach for classification of overlapping information, and data retrieval applications, NMF is separately applied on the TF data of each class (i.e., $\mathbf{V}_1$ and $\mathbf{V}_2$ in Fig. 1a) to decompose class-specific TF bases (i.e., $\mathbf{W}_1$ and $\mathbf{W}_2$) and create an overall TF basis matrix, consisting of all the class-specific TF bases. Then, a new test TF data is decomposed over the overall TF basis matrix to obtain the corresponding coefficients (i.e., h_1^T and h_2^T) of the linear combinations of the class-specific TF bases required to reconstruct the new TF data. Finally, the TF bases of each class and the corresponding coefficients can be used to reconstruct the part of the test data that belongs to each class (i.e., $\hat{\mathbf{V}}_{\text{Test1}}$ and $\hat{\mathbf{V}}_{\text{Test2}}$). Figure 1a shows a general schematic of this approach for a two-group classification scenario.

The approach in Fig. 1a is effective if the decomposed TF bases are not only representative of the TF structure

Fig. 1 a, b The general block diagram of data separation using standard NMF and the proposed JLNMF, respectively

of each class but also discriminative to the TF structures from different classes. Inspired by image recognition applications, much outstanding research on the design of discriminant NMF techniques has been reported [11, 16]. For the most part, those methods have focused on increasing the discrimination by enforcing a penalty to maximize the discrimination between the decomposed bases of different classes. However, the decomposed bases could be shared between classes, and it is not always possible to identify class-specific TF bases as not necessarily there is a unique relationship between the decomposed bases and class labels. As a result, those methods could be effective in improving the accuracy rate in image classification applications, but cannot be employed in challenging signal processing applications, including classification of overlapping classes, and data retrieval. In this paper, a novel JLNMF method is proposed to decompose discriminant and class-specific TF bases. This new approach enforces the discrimination between the TF bases of different classes by modeling the similarities between classes with a set of shared TF bases (i.e., $\mathbf{W}_j$ in Fig. 1b) and decomposes class-specific TF bases (i.e., $\mathbf{W}_{1d}$ and $\mathbf{W}_{2d}$) to model the distinct TF structure of each class. Decomposing a new data, with overlapping classes, over the entire discriminant and shared TF bases will identify the portion of the data that belongs to each class, and the corresponding coefficients can be used to successfully retrieve and separate the data from each class (i.e., $\hat{\mathbf{V}}_{\text{Test}1d}$ and $\hat{\mathbf{V}}_{\text{Test}2d}$). Figure 1b shows a general schematic of the proposed JLNMF approach for a two-group classification scenario.

The idea of including shared TF bases to enhance the discrimination power is inspired by an assumption made by the classification techniques, where they find a discriminating pattern between classes by dividing the feature space into non-overlapping subspaces, which each represents one of the classes. Although, this approach might be satisfactory in cases that the signals can be assumed to be completely separable in the feature space, and it seems to be too optimistic in applications where a natural and unavoidable overlap exists between different classes. In most real-world applications, specially in biomedical applications, the nature of signals from different classes is very similar, and there could only be slight changes in the TF patterns of signals from different classes. Natural similarities between different classes may result in some overlaps in the feature space, and the extracted features may not necessarily represent the discriminating structures in each class. This may cause some overlaps in the feature space, and thereby may degrade the performance of the classification task. Hence, identifying and removing those shared components in the TF bases from a classification task can improve the discrimination power between the remaining class-specific features.

In order to improve the discrimination power of the TF bases of each class, in the present work, we account for those similarities between classes by inclusion of shared TF bases in the formulation of the NMF decomposition. It is worth mentioning that in a work by [17], the authors have proposed a non-negative matrix partial co-factorization approach for retrieving the information of drum track from a compound music of a drum and harmony track. The authors use pure drum music from a different source and assume that all the TF bases of the drum music are shared with the compound music. Then the two music (i.e., drum and compound music) are simultaneously decomposed by enforcing the constraint that all the bases of the drum track are shared with the music compound and there are only distinct bases of harmony tracks in the music compound. Basically, there is only $\mathbf{W}_j$ and $\mathbf{W}_d$ in the proposed model by Yoo et al. instead of $\mathbf{W}_j$, $\mathbf{W}_{1d}$, and $\mathbf{W}_{2d}$. Hence, the model by Yoo et al. can be considered as a subset of the developed framework in the present work and can be considered only if the information of one of the classes is fully shared with the other class.

3 Jointly learnt NMF (JLNMF) method
3.1 Formulation of JLNMF
In order to identify TF bases representing each class (see Fig. 1a), using standard NMF, each TF data is separately decomposed to its TF bases and coefficients as follows:

$$\mathbf{V}_1 = \mathbf{W}_1\mathbf{H}_1, \text{ and } \mathbf{V}_2 = \mathbf{W}_2\mathbf{H}_2, \tag{4}$$

The standard approach decomposes the data of each class without any information from the other class by performing the following least square cost functions:

$$\begin{aligned} &\min_{\mathbf{W}_1,\mathbf{H}_1} f(\mathbf{W}_1, \mathbf{H}_1) = \|\mathbf{V}_1 - \mathbf{W}_1\mathbf{H}_1\|_F^2, \text{ and} \\ &\qquad \text{subject to } \mathbf{W}_1 \geq 0, \mathbf{H}_1 \geq 0 \\ &\min_{\mathbf{W}_2,\mathbf{H}_2} f(\mathbf{W}_2, \mathbf{H}_2) = \|\mathbf{V}_2 - \mathbf{W}_2\mathbf{H}_2\|_F^2, \\ &\qquad \text{subject to } \mathbf{W}_2 \geq 0, \mathbf{H}_2 \geq 0 \end{aligned} \tag{5}$$

However, in most real-world applications, data from different classes share some common structures, which do not contribute to the distinguished characteristics of the classes. We account for such bases by dividing base matrix $\mathbf{W}$ into two parts of $\mathbf{W}_d$ and $\mathbf{W}_j$, where the former corresponds to the discriminant class-specific TF bases, and the latter represents the shared TF bases (see Fig. 1b). Using the new notation, Eq. (4) can be rewritten as follows:

$$\mathbf{V}_1 = \left[\mathbf{W}_{1d} + \mathbf{W}_j\right]\mathbf{H}_1, \text{ and } \mathbf{V}_2 = \left[\mathbf{W}_{2d} + \mathbf{W}_j\right]\mathbf{H}_2, \tag{6}$$

The relationship between $\mathbf{W}_1$, $\mathbf{W}_{1d}$, and $\mathbf{W}_j$ is formulated as follows:

$$
\begin{aligned}
\mathbf{W}_1^{N \times r} &= \mathbf{W}_{1d}^{N \times r} + \mathbf{W}_j^{N \times r} \\
&= \left[w_{1d}^1 \, w_{1d}^2 \cdots w_{1d}^{r_d} \, 0 \cdots 0 \right] + \left[0 \cdots 0 \, w_j^1 \, w_j^2 \cdots w_{r_j}^j \right]
\end{aligned}
\tag{7}
$$

where $r = r_d + r_j$ equals to the total number of discriminant and shared components (i.e., r_d and r_j, respectively). Then, we can get the following:

$$
\mathbf{W}_{1d}^{\mathsf{T}} \mathbf{W}_j = \mathbf{0} \text{ and } \mathbf{W}_{2d}^{\mathsf{T}} \mathbf{W}_j = \mathbf{0}
\tag{8}
$$

To find an approximate solution for Eq. (6), the cost function in Eq. (5) is modified to jointly learn for $\mathbf{W}_{1d}$, $\mathbf{W}_{2d}$, and $\mathbf{W}_j$ matrices as shown in Fig. 2.

In this figure, the dark gray boxes represent the shared structures in each class and the light gray boxes represent the distinct structure in each class; for example, the reconstructed data, $\hat{\mathbf{V}}_1$, consists of the discriminant and shared parts (i.e., $\hat{\mathbf{V}}_{1d}$ and $\hat{\mathbf{V}}_{1j}$, respectively). Similar to standard NMF, the modified cost function has to minimize the reconstruction error (see arrows I and II in Fig. 2). The between-class discrimination is enforced by maximizing the error between the discriminant components of the two classes (see arrow III in Fig. 2). To minimize any similarities between shared and discriminant components of each class, the error between discriminant and shared components was maximized (see arrows IV and V

in Fig. 2). Therefore, the new cost function is formulated as follows:

$$
\begin{aligned}
\min_{\mathbf{W}_{1d}, \mathbf{W}_{2d}, \mathbf{W}_j, \mathbf{H}_1, \mathbf{H}_2} \quad & g\left(\mathbf{W}_{d1}, \mathbf{W}_{2d}, \mathbf{W}_j, \mathbf{H}_1, \mathbf{H}_2\right) \\
= & f\left(\mathbf{V}_1, \left[\mathbf{W}_{1d} + \mathbf{W}_j\right]\mathbf{H}_1\right) \\
& + f\left(\mathbf{V}_2, \left[\mathbf{W}_{2d} + \mathbf{W}_j\right]\mathbf{H}_2\right) \\
& - \delta f\left(\mathbf{W}_{1d}\mathbf{H}_1, \mathbf{W}_{2d}\mathbf{H}_2\right) \\
& - \lambda f\left(\mathbf{W}_{1d}\mathbf{H}_1, \mathbf{W}_j\mathbf{H}_1\right) \\
& - \lambda f\left(\mathbf{W}_{2d}\mathbf{H}_2, \mathbf{W}_j\mathbf{H}_2\right),
\end{aligned}
$$

$$
\text{subject to } \mathbf{W}_{1d} \geq 0, \mathbf{W}_{2d} \geq 0, \mathbf{W}_j \geq 0, \mathbf{H}_1 \geq 0, \mathbf{H}_2 \geq 0
\tag{9}
$$

where δ and λ are positive constants that control the trade-off between minimizing the reconstruction error and maximizing the discrimination power. The order of the terms in Eq. (9) follows the numbering of the arrows in Fig. 2.

3.2 Optimization of JLNMF

To minimize the new cost function, g in Eq. (9), the partial derivatives of g with respect to each unknown variable is obtained and set equal to zero as follows:
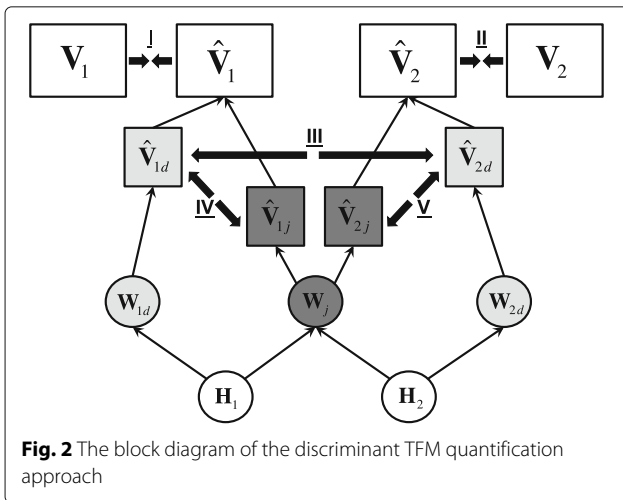
$$
\begin{aligned}
\nabla g(\mathbf{W}_j) &= \left(\left[\mathbf{W}_{1d} + \mathbf{W}_j\right]\mathbf{H}_1 - \mathbf{V}_1\right)\mathbf{H}_1^{\mathsf{T}} \\
&\quad + \left(\left[\mathbf{W}_{2d} + \mathbf{W}_j\right]\mathbf{H}_2 - \mathbf{V}_2\right)\mathbf{H}_2^{\mathsf{T}} \\
&\quad + \lambda\left(\mathbf{W}_{1d}\mathbf{H}_1 - \mathbf{W}_j\mathbf{H}_1\right)\mathbf{H}_1^{\mathsf{T}} \\
&\quad + \lambda\left(\mathbf{W}_{2d}\mathbf{H}_2 - \mathbf{W}_j\mathbf{H}_2\right)\mathbf{H}_2^{\mathsf{T}} = 0
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
\nabla g(\mathbf{W}_{1d}) &= \left(\left[\mathbf{W}_{1d} + \mathbf{W}_j\right]\mathbf{H}_1 - \mathbf{V}_1\right)\mathbf{H}_1^{\mathsf{T}} \\
&\quad - \delta\left(\mathbf{W}_{1d}\mathbf{H}_1 - \mathbf{W}_{2d}\mathbf{H}_2\right)\mathbf{H}_1^{\mathsf{T}} \\
&\quad - \lambda\left(\mathbf{W}_{1d}\mathbf{H}_1 - \mathbf{W}_j\mathbf{H}_1\right)\mathbf{H}_1^{\mathsf{T}} = 0
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
\nabla g(\mathbf{W}_{2d}) &= \left(\left[\mathbf{W}_{2d} + \mathbf{W}_j\right]\mathbf{H}_2 - \mathbf{V}_2\right)\mathbf{H}_2^{\mathsf{T}} \\
&\quad + \delta\left(\mathbf{W}_{1d}\mathbf{H}_1 - \mathbf{W}_{2d}\mathbf{H}_2\right)\mathbf{H}_2^{\mathsf{T}} \\
&\quad - \lambda\left(\mathbf{W}_{2d}\mathbf{H}_2 - \mathbf{W}_j\mathbf{H}_2\right)\mathbf{H}_2^{\mathsf{T}} = 0
\end{aligned}
\tag{12}
$$

$$
\begin{aligned}
\nabla g(\mathbf{H}_1) &= \left[\mathbf{W}_{1d} + \mathbf{W}_j\right]^{\mathsf{T}}\left(\left[\mathbf{W}_{1d} + \mathbf{W}_j\right]\mathbf{H}_1 - \mathbf{V}_1\right) \\
&\quad - \delta\mathbf{W}_{1d}^{\mathsf{T}}\left(\mathbf{W}_{1d}\mathbf{H}_1 - \mathbf{W}_{2d}\mathbf{H}_2\right) \\
&\quad - \lambda\mathbf{W}_{1d}^{\mathsf{T}}\left(\mathbf{W}_{1d}\mathbf{H}_1 - \mathbf{W}_j\mathbf{H}_1\right) = 0
\end{aligned}
\tag{13}
$$

$$
\begin{aligned}
\nabla g(\mathbf{H}_2) &= \left[\mathbf{W}_{2d} + \mathbf{W}_j\right]^{\mathsf{T}}\left(\left[\mathbf{W}_{2d} + \mathbf{W}_j\right]\mathbf{H}_2 - \mathbf{V}_2\right) \\
&\quad + \delta\mathbf{W}_{2d}^{\mathsf{T}}\left(\mathbf{W}_{1d}\mathbf{H}_1 - \mathbf{W}_{2d}\mathbf{H}_2\right) \\
&\quad - \lambda\mathbf{W}_{2d}^{\mathsf{T}}\left(\mathbf{W}_{2d}\mathbf{H}_2 - \mathbf{W}_j\mathbf{H}_2\right) = 0
\end{aligned}
\tag{14}
$$

Fig. 2 The block diagram of the discriminant TFM quantification approach

Simplifying Eqs. (10)-(14) and using Eq. (8), the following equations are derived:

$$\nabla g(\mathbf{W}_j) = (1-\lambda)\,\mathbf{W}_j\left(\mathbf{H}_1\mathbf{H}_1^{\mathrm{T}} + \mathbf{H}_2\mathbf{H}_2^{\mathrm{T}}\right)$$
$$+ (1+\lambda)\,\mathbf{W}_{1d}\mathbf{H}_1\mathbf{H}_1^{\mathrm{T}}$$
$$+ (1+\lambda)\,\mathbf{W}_{2d}\mathbf{H}_2\mathbf{H}_2^{\mathrm{T}}$$
$$- \mathbf{V}_1\mathbf{H}_1^{\mathrm{T}} - \mathbf{V}_2\mathbf{H}_2^{\mathrm{T}} = 0 \tag{15}$$

$$\nabla g(\mathbf{W}_{1d}) = (1-\delta-\lambda)\,\mathbf{W}_{1d}\left(\mathbf{H}_1\mathbf{H}_1^{\mathrm{T}}\right)$$
$$+ (1+\lambda)\,\mathbf{W}_j\mathbf{H}_1\mathbf{H}_1^{\mathrm{T}}$$
$$+ \delta\mathbf{W}_{2d}\mathbf{H}_2\mathbf{H}_1^{\mathrm{T}} - \mathbf{V}_1\mathbf{H}_1^{\mathrm{T}} = 0 \tag{16}$$

$$\nabla g(\mathbf{W}_{2d}) = (1-\delta-\lambda)\,\mathbf{W}_{2d}\left(\mathbf{H}_2\mathbf{H}_2^{\mathrm{T}}\right)$$
$$+ (1+\lambda)\,\mathbf{W}_j\mathbf{H}_2\mathbf{H}_2^{\mathrm{T}}$$
$$+ \delta\mathbf{W}_{1d}\mathbf{H}_1\mathbf{H}_2^{\mathrm{T}} - \mathbf{V}_2\mathbf{H}_2^{\mathrm{T}} = 0 \tag{17}$$

$$\nabla g(\mathbf{H}_1) = \left((1-\delta-\lambda)\,\mathbf{W}_{1d}^{\mathrm{T}}\mathbf{W}_{1d} + \mathbf{W}_j^{\mathrm{T}}\mathbf{W}_j\right)\mathbf{H}_1$$
$$- \left(\mathbf{W}_{1d} + \mathbf{W}_j\right)^{\mathrm{T}}\mathbf{V}_1 + \delta\mathbf{W}_{1d}^{\mathrm{T}}\mathbf{W}_{1d}\mathbf{H}_2 = 0 \tag{18}$$

$$\nabla g(\mathbf{H}_2) = \left((1-\delta-\lambda)\,\mathbf{W}_{2d}^{\mathrm{T}}\mathbf{W}_{2d} + \mathbf{W}_j^{\mathrm{T}}\mathbf{W}_j\right)\mathbf{H}_2$$
$$- \left(\mathbf{W}_{2d} + \mathbf{W}_j\right)^{\mathrm{T}}\mathbf{V}_2 + \delta\mathbf{W}_{2d}^{\mathrm{T}}\mathbf{W}_{2d}\mathbf{H}_1 = 0 \tag{19}$$

When solving for each variable in Eqs. (15)-(19), we keep the rest of the variables fixed and optimize the cost function of Eq. (9). For example, in Eq. (15), $\mathbf{W}_j$ is the variable that the equation in being solved for, and the terms $\left(\mathbf{H}_1\mathbf{H}_1^{\mathrm{T}} + \mathbf{H}_2\mathbf{H}_2^{\mathrm{T}}\right)$ and $(1+\lambda)\,\mathbf{W}_{1d}\mathbf{H}_1\mathbf{H}_1^{\mathrm{T}} + (1+\lambda)\,\mathbf{W}_{2d}\mathbf{H}_2\mathbf{H}_2^{\mathrm{T}} - \mathbf{V}_1\mathbf{H}_1^{\mathrm{T}} - \mathbf{V}_2\mathbf{H}_2^{\mathrm{T}}$ are constant matrices. Thus, using the same proof that was shown in [12] to get the NMF solutions in Eq. (3), it can be concluded that Eq. (9) is convergent and non-increasing when alternatively one matrix is improved while the other matrices are fixed, and the solution of the optimization problem in Eq. (9) can be solved by iteratively solving the following sub-problems [18, 19]:

$$\mathbf{W}_{1d}^{(k+1)} = \underset{\text{subject to } \mathbf{W}_{1d}\geq 0}{\arg\min}\, g\left(\mathbf{W}_1, \mathbf{W}_2^{(k)}, \mathbf{W}_j^{(k)}, \mathbf{H}_1^{(k)}, \mathbf{H}_2^{(k)}\right) \tag{20}$$

$$\mathbf{W}_{2d}^{(k+1)} = \underset{\text{subject to } \mathbf{W}_{2d}\geq 0}{\arg\min}\, g\left(\mathbf{W}_1^{(k+1)}, \mathbf{W}_2, \mathbf{W}_j^{(k)}, \mathbf{H}_1^{(k)}, \mathbf{H}_2^{(k)}\right) \tag{21}$$

$$\mathbf{W}_j^{(k+1)} = \underset{\text{subject to } \mathbf{W}_j\geq 0}{\arg\min}\, g\left(\mathbf{W}_1^{(k+1)}, \mathbf{W}_2^{(k+1)}, \mathbf{W}_j, \mathbf{H}_1^{(k)}, \mathbf{H}_2^{(k)}\right) \tag{22}$$

$$\mathbf{H}_1^{(k+1)} = \underset{\text{subject to } \mathbf{H}_1\geq 0}{\arg\min}$$
$$g\left(\mathbf{W}_1^{(k+1)}, \mathbf{W}_2^{(k+1)}, \mathbf{W}_j^{(k+1)}, \mathbf{H}_1, \mathbf{H}_2^{(k)}\right) \tag{23}$$

$$\mathbf{H}_2^{(k+1)} = \underset{\text{subject to } \mathbf{H}_2\geq 0}{\arg\min}$$
$$g\left(\mathbf{W}_1^{(k+1)}, \mathbf{W}_2^{(k+1)}, \mathbf{W}_j^{(k+1)}, \mathbf{H}_1^{(k+1)}, \mathbf{H}_2\right) \tag{24}$$

Although the optimization problem of Eq. (9) is not convex over variables $\mathbf{W}_j, \mathbf{W}_{1d}, \mathbf{W}_{2d}, \mathbf{H}_1$, and $\mathbf{H}_2$, the sub-problems in Eqs. (20) to (24) are convex and can be solved using projected gradient methods as follows.

For a bounded-constrained optimization problem described as below:

$$\underset{\mathbf{U}\in R^n}{\min}\qquad g(\mathbf{U}): R^n \to R \tag{25}$$
$$\text{subject to}\quad 0 \leq x_i \leq \infty, i = 1,\ldots,n$$

projected gradient methods update the current solution $\mathbf{U}^{(t)}$ to $\mathbf{U}^{(t+1)}$ by the following updating rule:

$$\mathbf{U}^{(t+1)} = P\left[\mathbf{U}^{(t)} - \alpha^{(t)}\nabla g(\mathbf{U}^{(t)})\right]$$
$$= \max\left\{\left(\mathbf{U}^{(t)} - \alpha^{(t)}\nabla g\left(\mathbf{U}^{(t)}\right)\right), 0\right\} \tag{26}$$

where (t) is the iteration order, $\nabla g(\mathbf{U})$ is the projected gradient of the function g with respect to $\mathbf{U}$, while all the other matrices are constant, and $\alpha^{(t)}$ is the step size to update the matrix. The step size is found as $\alpha^{(t)} = \beta^{k_t}$, where k_t is the first non-negative integer k for which the following equation holds:

$$g\left(\mathbf{U}^{(t+1)}\right) - g\left(\mathbf{U}^{(t)}\right) \leq \sigma\nabla g\left(\mathbf{U}^{(t)}\right)\left(\mathbf{U}^{(t+1)} - \mathbf{U}^{(t)}\right) \tag{27}$$

where $0 < \sigma < 1$ is a scalar parameter that is defined based on the convexity of the function $g(\mathbf{U})$ [20]. The condition in Eq. (27) ensures the sufficient decrease of the function value per iteration. Bertsekas [20] has proved that $\alpha^{(t)} > 0$ satisfying Eq. (27) always exists and every limit point of the solution $\left\{\mathbf{U}^{(t)}\right\}_{t=1}^{\infty}$ is a stationary point of Eq. (25). Therefore, by iteratively updating each matrix using updating rule in Eq. (26), while the other matrices are constant, we can iteratively solve for $\mathbf{W}_j, \mathbf{W}_{1d}, \mathbf{W}_{2d}, \mathbf{H}_1$, and $\mathbf{H}_2$, in the sub-problems of Eqs. (20) to (24).

3.3 Computation cost considerations
The computation cost for each iteration consists of computation of the gradient functions of Eqs. (15)-(19) to

update current variables using Eq. (26) and the computational task of the sub-iteration to find a step size α such that the condition in Eq. (27) is satisfied. Let us first consider the computation of $\mathbf{W}_j$. The computation of constant matrices $\mathbf{H}_1\mathbf{H}_1^\mathrm{T}$, $\mathbf{H}_2\mathbf{H}_2^\mathrm{T}$, $\mathbf{W}_{1d}\left(\mathbf{H}_1\mathbf{H}_1^\mathrm{T}\right)$, $\mathbf{W}_{2d}\left(\mathbf{H}_2\mathbf{H}_2^\mathrm{T}\right)$, $\mathbf{V}_1\mathbf{H}_1^\mathrm{T}$, and $\mathbf{V}_2\mathbf{H}_2^\mathrm{T}$ is going to be performed only once per iteration and the computational cost is of the order of $O\left(\bar{r}^2M + Nr_d^2 + NMr\right)$, where $\bar{r} = max(r_d, r_j)$. Given that $r_d < r$, $r_j < r$, and r is a small value compared to N and M, the computational cost is of the order of $O(NMr)$. Similarly, the computational cost of the fix matrices at each iteration for $\mathbf{W}_d$ and $\mathbf{H}$ can be computed respectively in $O\left(\bar{r}^2M + \bar{(r)}^2N + NMr\right)$ and $O\left(\bar{r}^2N + r_d^2M + NMr\right)$ operations, which by considering the NMF criterion on r ($r << N, M$), are computed to be of the order of $O(NMr)$.

To reduce the computational cost of the decreasing condition in Eq. (27), a similar strategy to [15] was employed. For any quadratic function $g(U)$ and vector U, using Taylor series, the following approximation can be considered:

$$
\begin{aligned}
g\left(U^{(k+1)}\right) = {} & g\left(U^{(k)}\right) + \nabla g\left(U^{(k)}\right)^\mathrm{T}\left(U^{(k+1)} - U^{(k)}\right) \\
& + 0.5\left(U^{(k+1)} - U^{(k)}\right)^\mathrm{T}\nabla^2 g\left(U^{(k)}\right) \\
& \times \left(U^{(k+1)} - U^{(k)}\right)
\end{aligned}
\tag{28}
$$

and re-write Eq. (27) as follows:

$$
\begin{aligned}
& (1 - \sigma)\nabla g\left(U^{(k)}\right)^\mathrm{T}\left(U^{(k+1)} - U^{(k)}\right) \\
& + 0.5\left(U^{(k+1)} - U^{(k)}\right)^\mathrm{T}\nabla^2 g\left(U^{(k)}\right)\left(U^{(k+1)} - U^{(k)}\right) \leq 0
\end{aligned}
\tag{29}
$$

See Appendix for $\nabla^2 g(\mathbf{W}_j)$, $\nabla^2 g(\mathbf{W}_{1d})$, $\nabla^2 g(\mathbf{W}_{2d})$, $\nabla^2 g(\mathbf{H}_1)$, and $\nabla^2 g(\mathbf{H}_2)$. Using the decreasing condition of Eq. (29) instead of Eq. (27), $g(U^{(k+1)})$ does not need to be computed at every iteration to find a suitable step size. Although $\nabla^2 g\left(U^{(k)}\right)$ has to be computed but it will happen only once. Hence, the computational cost at each sub-iteration can be computed in $O\left(kNr_d^2\right)$, $O\left(kNr^2\right)$, and $O(kMr^2)$ for $\mathbf{W}_j$, $\mathbf{W}_d$, and $\mathbf{H}$, respectively, where k is the average number of checking the decreasing condition in Eq. (27) at each iteration. Since $r_d < r$, the overall computational cost to solve Eq. (9) is as follows: *#iterations* $\times$ $\left(O(NMr) + \textit{#sub\_iterations} \times O(kNr^2 + kMr^2)\right)$.

3.4 JLNMF algorithms

The optimization algorithm of the proposed JLNMF method is summarized in Algorithms 1 and 2. Regarding the initialization of the algorithm, a common approach for

NMF is initialization of the values of $\mathbf{W}$ and $\mathbf{H}$ matrices as non-negative random values. Following the same initialization method, all matrices $\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_j, \mathbf{H}_1$, and $\mathbf{H}_2$ are initialized to non-negative random values.

Algorithm 1 Updating base and coefficient matrices using a projected gradient method

---

Given r, r_j, λ and γ
Input: Data $\{\mathbf{V}_1, \mathbf{V}_2\}$
1. Initialize $\mathbf{W}_1^{(1)} \geq 0, \mathbf{W}_2^{(1)} \geq 0, \mathbf{W}_j^{(1)} \geq 0, \mathbf{H}_1^{(1)} \geq 0, \mathbf{H}_2^{(1)} \geq 0$ with random values
2. **Repeat** for $t = 1, 2, \ldots$ (Jointly update the class-specific and shared TF bases.)
3. Update $\left\{\mathbf{W}_{1d}^{(t)}, \mathbf{W}_{2d}^{(t)}\right\}$ by optimizing Eqs. (20) and (21) using Algorithm 2.
4. Update $W_j^{(t)}$ by optimizing Eq. (22) using Algorithm 2.
5. Update $\left\{\mathbf{H}_1^{(t)}, \mathbf{H}_2^{(t)}\right\}$ by optimizing Eq. (23) and (24) using Algorithm 2.
6. **Until** convergence.
Output: The learned class-specific bases $\{\mathbf{W}_{1d}, \mathbf{W}_{2d}\}$, shared common bases $\left\{\mathbf{W}_j\right\}$

Algorithm 2 Step size selection of the projected gradient method

---

Given $0 < \beta, \alpha < 1$
Input: $\mathbf{U}^{(t)}$
1. Initialize $\alpha_t^{(1)}$ by α_{t-1} or 1 if $t = 1$.
2. **if** $\alpha_t^{(k)}$ satisfies Eq. (29)
3. **Repeat** for $k = 1, 2, \ldots$ (update the step size)
4. $\alpha_t^{(k+1)} = \alpha_t^{(k)}/\beta$
5. **until** Eq. (29) is not satisfied
6. **else** $\alpha_t^{(k)}$ does not satisfy Eq. (29)
7. **Repeat** for $k = 1, 2, \ldots$
8. $\alpha_t^{(k+1)} = \alpha_t^{(k)} . \beta$
9. **Until** Eq. (29) is satisfied.
10. **end if**
11. Compute the updated matrix:
$$
\mathbf{U}^{(t+1)} = P\left[\mathbf{U}^{(t)} - \alpha_t^{(k)}\nabla g(\mathbf{U}^{(t)})\right].
$$
Output: The updated matrix $\mathbf{U}^{(t+1)}$

3.5 Parameter selection considerations

There are several parameters that need to be selected in the proposed JLNMF algorithm. The values of σ and β are set to 0.01 and 0.1, respectively, as commonly used in literature [15], and the values of λ and γ are dependent on the application. A suitable value of λ and γ can be selected

from the range of [0.05 to 0.20] in order to balance off between the reconstruction error (i.e., the first two terms in Eq. 9) and the discrimination power (i.e., the remaining terms). Selection of r is also application depended and considerations similar to those for NMF can be applied to adjust r [21–23]. The selection of r_j can be adjusted to increase the discrimination power between classes (i.e., maximize the difference between $\hat{\mathbf{V}}_{1d}$ and $\hat{\mathbf{V}}_{2d}$). Furthermore, it is worth mentioning that although the proposed JLNMF algorithm has been formulated in the context of two classes, the extension of the method to multiple classes can be simply done. However, since the objective here is to address the challenges in biomedical signal classification applications, the paper has focused on two-class cases (i.e., abnormal vs. normal).

4 Experiments with JLNMF
4.1 Synthetic data
4.1.1 Visualization of JLNMF
The performance of the proposed JLNMF is visualized and compared to the standard NMF through a synthetic example. In this example, two non-stationary signals are generated, $x(m)$ and $y(m)$, by combining a set of non-stationary components in the form of $\alpha g(\mu, \sigma) \sin(2\pi(f + \theta m)m)$ where $g(\mu, \sigma)$ is a Gaussian function with mean μ and variance of σ, and the set $(\alpha, \mu, \sigma, f, \theta)$ is the parameter for each component (a) to (g). The parameters of each component are manipulated to generate components (a) through (d) (see Fig. 3) as the shared structure by the two signals, components (e) and (f) as the distinct structure of signal $x(m)$, and the chirp component (g) as the distinct structure of signal $y(m)$. In this figure, parameters

$(\alpha, \mu, \sigma, f, \theta)$ for components (a) to (g) are, respectively, as follows: (1, 1.2, 0.05, 0.325, 0), (1,1.2, 0.05, 0.075, 0), (1, 3.4, 0.04, 0.125, 0), (1, 5.6, 0.03, 0.325, 0), (3, 0.9, 0.001, 0.450, 0), (3, 3.5, 0.001,0.213, 0), and (1, 5.3, 0.05, 0.005, 0.245), and Spectrogram with FFT size of 1024 points and Kaiser window with parameter of 5, length of 256 samples, and 220 samples overlap, is used to construct the TF representations. Each TF data (i.e., $\mathbf{V}_1$ and $\mathbf{V}_2$) was decomposed using standard NMF method with $r = 6$, and the decomposed TF basis and coefficient matrices were obtained and displayed in Fig. 3c, d. It can be seen that NMF adaptively breaks down the non-stationary signal into TF basis and coefficient matrices representing the spectral and temporal information of the signal, respectively. For example, the first TF basis in $\mathbf{W}_1$ represents the spectral characteristics of components (f) and (c) in the TF signal $\mathbf{V}_1$, and the corresponding TF coefficient in $\mathbf{H}_1$ represents the temporal locations of those components in the TF signal. The best scenario for this decomposition would be if the decomposed TF bases represent the discriminant structures of each TF signal separate from the shared ones. Then the discriminant TF bases would be powerful in differentiating between the two signals. However, there is no separation between the TF bases that represent the shared structure (i.e., components (a)–(d)) and the distinct structure of each signal; for example, TF basis 1 of $x(m)$, contains the frequency structure of components (e) and (b), and similarly TF basis 2 models the components (a), (b), and (e). A similar behavior can be seen in the TF bases of signal $y(m)$.

We repeated the process using the proposed JLNMF with $r = 6$ and $r_j = 3$. The values of σ and β were set to

Fig. 3 The TF representations (**b**) of two non-stationary signals from class 1 and class 2 data (**a**) are decomposed into TF bases and coefficients (**c** and **d**, respectively) using standard NMF
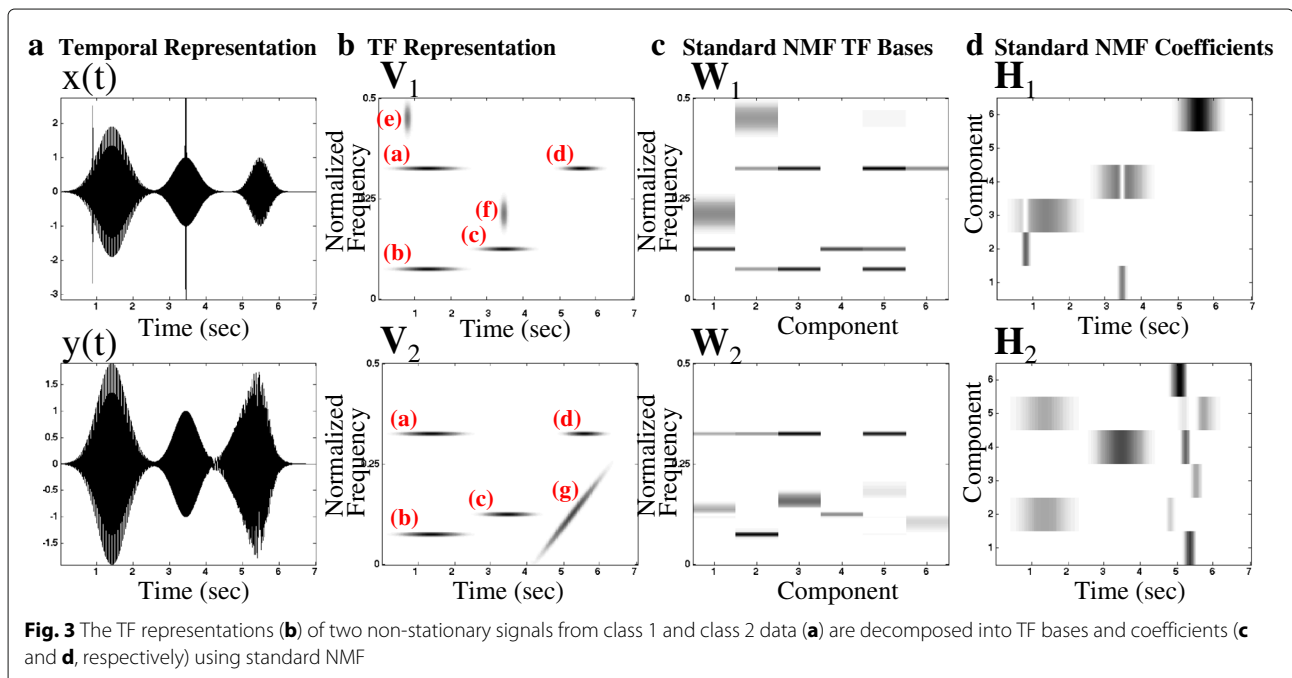
0.01 and 0.1, respectively, as commonly used in literature [15], and a value of 0.05 was selected for λ and γ. As can be seen in Fig. 4a, the first three (i.e., $r_d = 3$) components of the decomposed TF bases $\mathbf{W}_{1d}$ and $\mathbf{W}_{2d}$, respectively, represent the distinct structure of signals $x(m)$ and $y(m)$, and the last three (r_j) are zero (see Eq. (7) for the structures of $\mathbf{W}_d$ and $\mathbf{W}_j$ with respect to $\mathbf{W}$). The shared structure is represented in the last three ($r_j=3$) of the shared TF bases $\mathbf{W}_j$. It is interesting to point out that although r_d was set to be three, the method adaptively identifies that only two distinct TF bases are sufficient to model the discriminate structure in $x(m)$, and finds an empty (all zero) for the one of the TF basis vectors in $\mathbf{W}_{2d}$. Additionally, unlike standard NMF TF bases, none of the JLNMF TF bases represents both the shared and distinct TF structure of the original signals $x(m)$ and $y(m)$. Comparing the decomposed TF basis and coefficient matrices of the proposed JLNMF and standard NMF (Figs. 3c, d and 4a, b), it can be observed that the JLNMF coefficients are noticeably more localized. Furthermore, the discriminant TF structures of signals $x(m)$ and $y(m)$ are successfully reconstructed as $\mathbf{W}_{1d}\mathbf{H}_1$ and $\mathbf{W}_j\mathbf{H}_2$, respectively, as shown in Fig. 4c. The shared structures are also successfully reconstructed as

$\mathbf{W}_j\mathbf{H}_1$ and $\mathbf{W}_{1d}\mathbf{H}_2$ for each signal as shown in Fig. 4d. This example demonstrates that the proposed JLNMF successfully separates the shared components and the distinct class-specific structures of each TF data. The only way to make this separation happen in the standard NMF is by identifying the TF bases of $\mathbf{W}_1$ and $\mathbf{W}_2$ that are correlated as the shared TF bases, and the uncorrelated ones as the class-specific TF bases. However, as can be seen in Fig. 3c, the standard NMF TF bases are a mixture of shared and distinct structures and the TF bases of the two classes are correlated, and thereby, the standard NMF cannot achieve the results shown in Fig. 4.

4.1.2 Properties of JLNMF

It is desirable that the extracted discriminant bases are robust to signal processing operations, which are not expected to affect the classification task. This section examines the properties of JLNMF for amplitude scale and time shift.

Amplitude scaling: If $y(m)$ is the amplitude scaled version of signal $x(m)$, such that $y(m) = ax(m)$, the TFM of the signal $y(m)$, $\mathbf{V}_2$, can be written as $\mathbf{V}_2 = a^2\mathbf{V}_1$, where
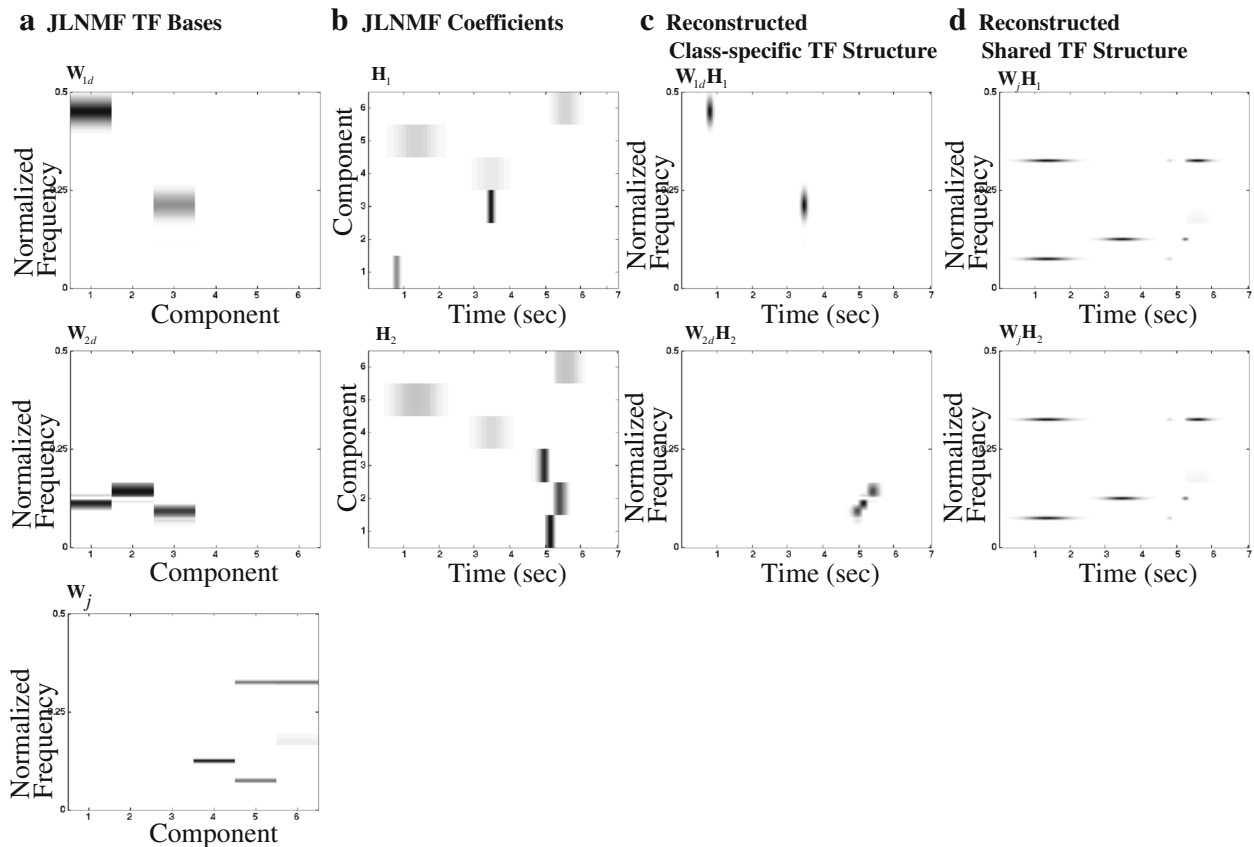

Fig. 4 Visualization of the JLNMF method on data shown in Fig. 3a, b. **a** The decomposed TF bases. **b** The corresponding coefficients. **c** The reconstructed class-specific structure in each class ($\hat{\mathbf{V}}_{1d}$ and $\hat{\mathbf{V}}_{2d}$). **d** The reconstructed shared TF structure in each class ($\hat{\mathbf{V}}_{1j}$ and $\hat{\mathbf{V}}_{2j}$)

$\mathbf{V}_1$ is the TFM data of signal $x(m)$ and a is the amplitude scale. Since the amplitude scaling does not affect the TF structure of the signal, it is expected that JLNMF does not identify any discriminant TF bases. To put it into test, an amplified version of signal $x(m)$ in Fig. 3a was generated and JLNMF was applied to the original and scaled TF data, and as expected, both of the discriminant matrices, $\mathbf{W}_{1d}$ and $\mathbf{W}_{2d}$, were obtained empty, demonstrating that JLNMF was successful in identifying that there were no distinct differences between the two data, and verifying that JLNMF is invariant to amplitude scaling.

Time shift: The property of JLNMF under temporal shift is examined using the following example. Let us consider two signals $x(m)$ and $y(m) = x(m - \tau)$. The temporal shift only shifts the TF structure of the TF data in time and does not change the TF components of the data (i.e., $\mathbf{V}_2(n, m) = \mathbf{V}_1(n, m - \tau)$. Hence, JLNMF is not expected to identify any distinct structural differences between $x(m)$ and $y(m)$. A signal $x(m)$ and $x(m - \tau)$ with a time shift of $\tau = 1.5$ s is generated as shown in Fig. 5a, and the TF data is displayed in Fig. 5b.

JLNMF with $r = 9$ and $r_j = 6$ was performed on the TF data, and the decomposed TF bases and coefficients are displayed in Fig. 5c. The method did not identify any discriminant TF matrices (i.e., $\mathbf{W}_{1d}$ and $\mathbf{W}_{2d}$ are both zero). Hence, $\hat{\mathbf{V}}_{1d}$ and $\hat{\mathbf{V}}_{2d}$ are empty. $\mathbf{W}_j$ contained the spectral vectors that are shared between $x(m)$ and $x(m - \tau)$, and $\mathbf{H}_2$ was equal to $\mathbf{H}_1(m - \tau)$. As can be seen, the JLNMF algorithm is invariant to time shifting.

4.1.3 *Data retrieval application using JLNMF algorithm*
The developed JLNMF is applied for a data retrieval case, where there is an overlap of information in the test data. In order to evaluate the results, a synthetic non-stationary dataset is generated (see Fig. 6) as inspired by the previous work in the literature [24, 25]. In this data, the triangles in the TF representation of classes x and y could be considered as the distinct structure in each class, and the box represents the shared structures between the two. Hence, each class consists of a class-specific distinct structure and one shared structure. Each signal is defined as the sum of two components as defined below:

$$x_{\text{train}}(m) = g(0.5, 0.18)\sin\left[2 * \pi\left(a_0 + a_1 m\right)\right]$$
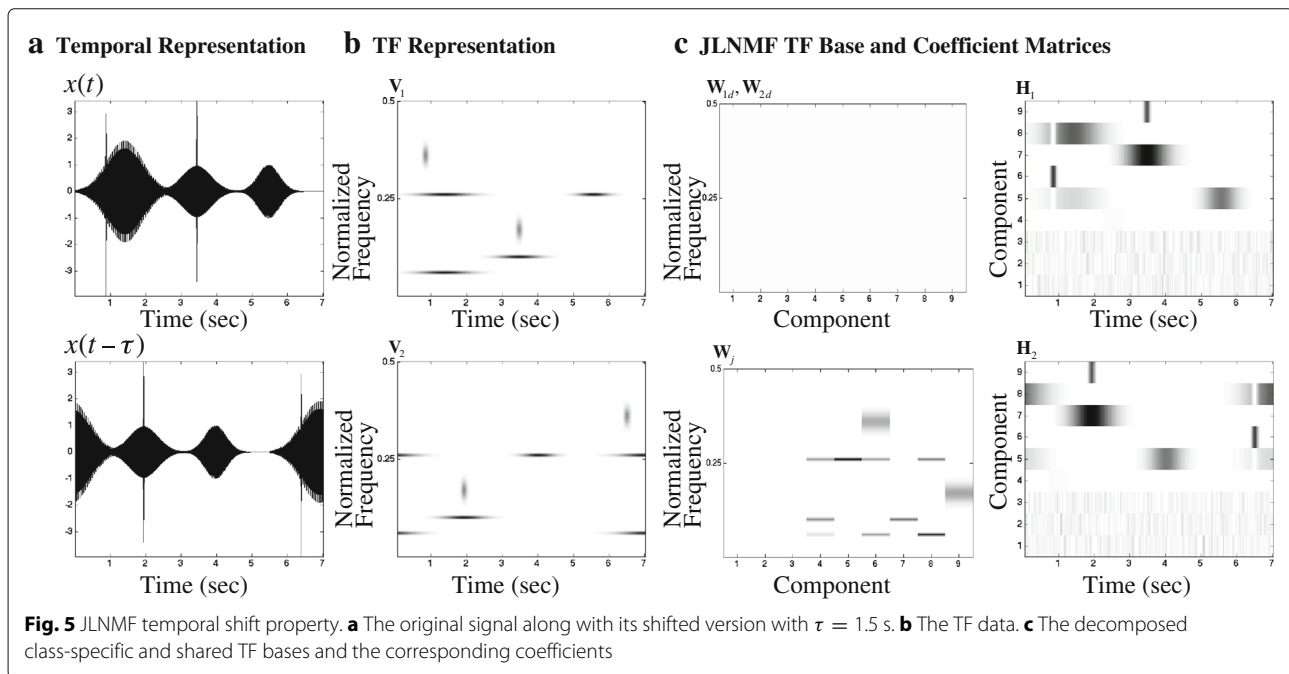$$+ g(0.5, 0.18)\sin\left[2 * \pi\left(b_0 + b_1 m + b_2 m^2\right)\right], \quad (30)$$

where the parameters of the shared component (i.e., a_0, b_0) belong to a uniform distribution U(0,1), $a_1 = 0.25$, $b_1 = 0.40$, and $N = 1,000$ is the signal length in samples with a sampling frequency of 1 kHz. Two classes are generated by selecting b_2 from one of the following uniform distributions:

$$\text{Class 1: U}\left(\frac{-0.30}{2(N-1)}, \frac{-0.20}{2(N-1)}\right) \quad (31a)$$

and

$$\text{Class 2: U}\left(\frac{-0.15}{2(N-1)}, \frac{-0.05}{2(N-1)}\right) \quad (31b)$$

The TF representation for signals in each class is plotted in Fig. 6a, b. For training purposes, a total of 300 signals
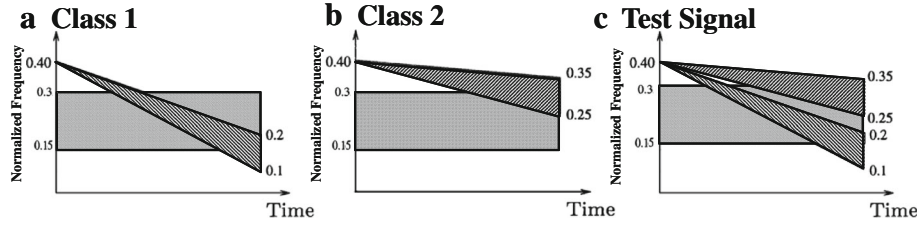

Fig. 5 JLNMF temporal shift property. **a** The original signal along with its shifted version with $\tau = 1.5$ s. **b** The TF data. **c** The decomposed class-specific and shared TF bases and the corresponding coefficients

Fig. 6 The TF structures of class 1 (**a**) and class 2 (**b**) are shown in this figure. Each class consists of a tone at random frequency uniformly distributed between 0.15 and 0.3, and a linear chirp signal starting at normalized frequency 0.40 and ending at a random frequency uniformly distributed between 0.1 and 0.2 for class 1 and between 0.25 and 0.35 for class 2. The test signals **c** with 100 % overlap between class 1 and 2 signals are generated

are generated in each class. A test data set is generated by 100 % overlap between the two classes as follows:

$$
\begin{aligned}
x_{\text{test}}(m) &= g(0.5, 0.18)\sin\left[2 * \pi\,(a_0 + a_1 m)\right] \\
&\quad + g(0.5, 0.18)\sin\left[2 * \pi\,\left(b_0 + b_1 m + b_{21}m^2\right)\right] \\
&\quad + g(0.5, 0.18)\sin\left[2 * \pi\,\left(b_0 + b_1 m + b_{22}m^2\right)\right]
\end{aligned}
\tag{32}
$$

where b_{21} and b_{22} are selected using Eqs. (31a) and (31b), respectively. The TF representation for signals in each class is plotted in Fig. 6c.

A total of 40 test signals are generated, and the training/classification tasks are performed as follows: (1) spectrogram with FFT size of 128 points and Kaiser window with parameter of 5, length of 128 samples, and 125 samples overlap, is used to construct the TF matrices of each signal. The dimension of the TF matrix is 65×291. (2) each TF matrix is collapsed into a vector, $v_{18915 \times 1}$, and the training data sets are created as: $\mathbf{V}_1^{(18915 \times 300)} = [v_1(1)\, v_1(2)\, \cdots\, v_1(300)]$ and $\mathbf{V}_2^{(18915 \times 300)} = [v_2(1)\, v_2(2)\, \cdots\, v_2(300)]$. (3) JLNMF is applied to the TF data $\mathbf{V}_1$ and $\mathbf{V}_2$ and class-specific and shared TF bases are decomposed. The parameters r and r_j are selected 40 and 20, respectively, and λ and γ are varied over $\{0.05, 0.1, 0.15, 0.2\}$. (4) for classification purposes, each test data, $v_{\text{test}}(i)_{i=1:40}$, is projected over the JLNMF class-specific and shared TF bases obtained in the previous step. To make this happen, an overall TF bases, $\mathbf{W}$, is constructed in the form of $\left[w_{1d}(1)\, \cdots\, w_{1d}(r_d)\, w_{2d}(1)\, \cdots\, w_{2d}(r_d)\, w_j(r - r_j + 1)\, \cdots\, w_j(r)\right]$, and coefficient vector $h_{\text{test}}^T(i)_{i=1:40}$ is computed using the updating rule in Eq. (3). The first $r_d = 20$ elements in $h_{\text{test}}^T(i)$ contain the corresponding coefficients of the class-specific TF bases of class 1, the next $r_d = 20$ elements contain the corresponding coefficients of the class-specific TF bases of class 2, and the last $r_j = 20$ elements contain the corresponding coefficients of the linear combinations of the shared TF bases. Hence, the

distinct structures of class 1 and class 2 in test data i are reconstructed as follows:

$$
\hat{v}_{1d}(i) = [w(1)\, \cdots\, w(r_d)]\, h^T(1 : r_d) \text{ and} \tag{33}
$$
$$
\hat{v}_{2d}(i) = [w(r_d + 1)\, \cdots\, w(2r_d)]\, h^T(r_d + 1 : 2r_d)
$$

(5) The correlation coefficients between $\hat{v}_{1d}(i)$ and the original $v_{1d}(i)$, and $\hat{v}_{2d}(i)$ and the original $v_{2d}(i)$, are computed. The average of the correlation values, which we denote as class-specific recovery percentage (CRP), is computed as shown in the following equation to assess the method's success in identifying and recovering the distinct structures of each class in the presence of 100 % overlap (see Fig. 6c for the test data).

$$
\text{CRP}_1(\%) = \sum_{i=1}^{i=300} \frac{\text{Cor}\left(\hat{v}_{1d}(i), v_{1d}(i)\right)}{300} \times 100 \tag{34}
$$

$$
\text{CRP}_2(\%) = \sum_{i=1}^{i=300} \frac{\text{Cor}\left(\hat{v}_{2d}(i), v_{2d}(i)\right)}{300} \times 100 \tag{35}
$$

(6) Steps (1)–(5) are repeated 20 times and the average CPR is reported for each set of λ and γ over $\{0.05, 0.1, 0.15, 0.2\}$. The average CRP_1 and CRP_2 values are shown in Fig. 7.

As can be seen, there is a high correlation between the reconstructed and the original distinct structures of
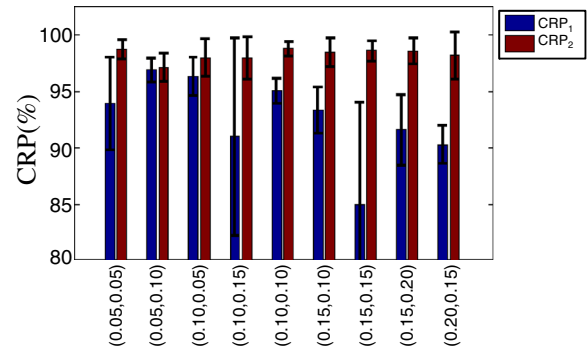

Fig. 7 The class-specific recovery percentage (CRP) using the proposed JLNMF algorithm

each class, meaning that the proposed JLNMF successfully separated the distinct structure of each class from the shared structure. The performance slightly varied by the selected value of λ and γ with (0.10 and 0.05), being the best average CRP value (97 %) for the given example. Figure 8 displays the reconstructed class-specific structures for two examples with $(\mathrm{CRP}_1, \mathrm{CRP}_2)$ values of (96 %, 98 %) and (72 %, 99 %) for the example in the top and bottom, respectively. As can be seen in this plot, although there is a relatively low correlation between $\hat{v}_{1d}(i)$ and $v_{1d}(i)$, the distinct structure of class 1 has been mostly recovered. To compare JLNMF with NMF, the experiment is repeated using standard NMF, where NMF with $r = 30$ is separately applied to each TF data and the shared TF bases are identified as the TF bases with a correlation value of greater than 0.9. The average CRP value of 88 % is obtained, which is significantly less the CRP value of 97 %, obtained using the proposed JLNMF method.

4.2 Real data

4.2.1 Localization of epileptic spikes in EEGs

The application of the proposed JLNMF method to localize the epileptic discharges associated with infantile spasms in hypsarrhythmia (HYPS) is explored. Infantile spasms refer to a catastrophic form of epilepsy occurring in infancy that is diagnosed based on the findings of HYPS in EEG recordings combined with epileptic spasms [26]. HYPS is characterized by a chaotic and high voltage background with multifocal, discharges [27]. However, identifying this pattern of activity in a conventional EEG recording is challenging in the presence of HYPS due to the abundance of epileptiform discharges with varying focality and morphology [28]. An experienced electroencephalographer interprets an EEG by inspecting

and approximating the characteristics of HYPS subjectively rather than through objective quantification. Due to complex nature of these signals, even experienced EEG readers tend to interpret HYPS differently, which can have serious implications in the treatment of the infant [28]. Several algorithms have been developed to detect epileptic discharges during epilepsy. Some of those algorithms include temporal analysis based on template matching and mimetic analysis methods, or TF method based on wavelet analysis. However, those methods have been developed for epileptic discharge detection in EEG signals associated with other types of epilepsy and not in the presence HYPS. The existing methods are either based on supervised classifiers in which the presence of true spikes has to be readily available and identifiable to train the algorithm during a learning phase, or they are template based, which rely on pre-specified spike characteristics such as amplitude and duration of the discharges. Given the chaotic appearance of EEG during HYPS, the manual localization of true spikes is not always possible. The spikes of interest are characteristically non-uniform, which presents a challenge for temporal-based methods. Hence, there is a need to develop an semi-supervised feature extraction and classification method to assist in localization of spikes with multiple foci and varying morphologies during HYPS.

A 5-min section of awake EEG recording from an infant with infantile spasms is used to explore the application of JLNMF method for semi-supervised localization of epileptic spikes. The subject consent was obtained through the Infantile Spasms Registry and Genetic Studies via a protocol approved by the University of Rochester's Research Subjects Review Board. The EEG signals were recorded based on the international standard 10–20
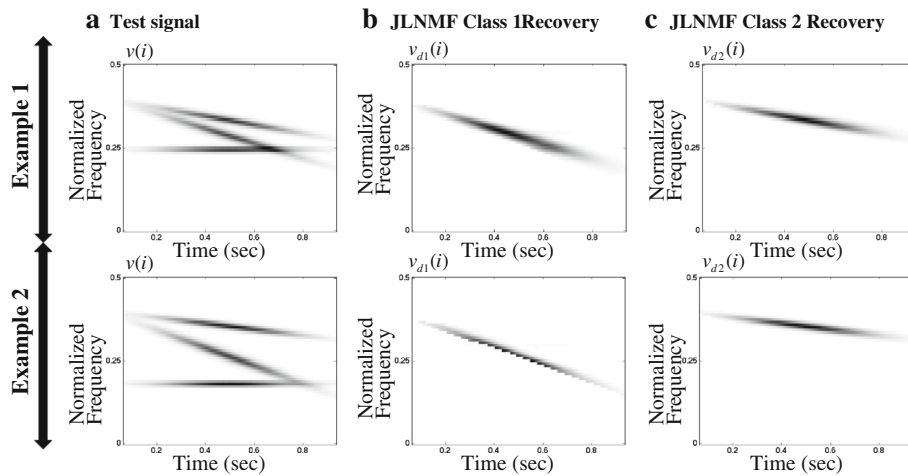

Fig. 8 The JLNMF reconstructed class-specific structures for two examples: Example 1 (CRP_1=96 % and CRP_2=98 %), Example 2 (CRP_1=72 %, CRP_2=99 %). See Fig. 6 for the structures of the shared and class-specific components

system with sampling frequency of 512 Hz. The recording EEGs were imported to Persyst EEG software (Persyst, San Diego, CA) for artifact reduction and then were imported into MATLAB and bandpass filtered (0.5–30 Hz) for further analysis. All the epilepticform discharges were manually marked by an epileptologist.

The data is divided into 10-s windows and an epileptologist labels each 10-s EEG window as non-spike (NS) if there are no epiletiform spikes in that window; otherwise, the window is marked as possibly-spike (PS). The objective is to characterize the structure of the epileptiform spikes and localize them during each 10-s window. The localization task is formulated as a classification task, where the objective is to decompose the class-specific (i.e., epileptic spikes) and shared bases (i.e., the common EEG baseline), which can be used to reconstruct the class-specific data in each class. The distinct structure of the PS recordings is expected to indicate the spike locations during the EEG recordings. There are two main challenges: the first challenge is that the EEG recordings are strongly non-stationary, which can be addressed by using the TF data of the EEG recordings to improve the representation of the non-stationary information. The second challenge is the substantial amount of similarity (i.e., overlap) between the two NS and PS classes. This is mainly because the exact locations of the epileptiform spikes are not specified, and the only available information is that the PS class contains several spikes at some unknown locations. The proposed JLNMF method is used to address this challenge, as it is able to decompose the TF data to the spike-specific TF bases from the common EEG baseline in a semi-supervised fashion. The details of the application of JLNMF on the EEG recording is as follows:

(1) The TF data is constructed using matching-pursuit TF (MP-TF) method. MP-TF has a much higher TF resolution compared to spectrogram [2, 8], and can better represent the spike-related transients and non-stationarity of the EEG recordings. The resolution of the MP-TF is selected to be 0.15 Hz in frequency and 2 ms in time. Since there is no meaningful physiological information beyond 30 Hz, the frequency domain is limited to that value. The dimension of the TF matrix is 200×5120 for each 10-s segment. (2) Each 10-s TF matrix is divided into 64 sections and each section is collapsed into a vector, $v_{16,000 \times 1}$. The training data set is created by collecting the collapsed TF vectors of half of the NS and PS data. (3) JLNMF is applied to the TF data and class-specific and shared TF bases are decomposed. The parameters r and r_j are selected as 40 and 20, respectively, and the values of 0.10 and 0.05 are, respectively, selected for λ and γ. Figure 9a shows the shared (on top) and PS-specific (on bottom) TF bases. To visualize the TF bases, each decomposed base is restructured back to the original size. The NS-specific TF matrix is found empty, which means that

JLNMF does not identify any distinct structure to the NS class. (4) For classification purposes, each test data is projected over the JLNMF PS-specific and shared TF bases obtained in the previous step, and the coefficient vector is computed using the updating rule in Eq. (3), and the distinct structures corresponding to the PS-specific TF bases are reconstructed for all the test data.

Figure 9c shows the PS-specific reconstructed TF data for the EEG signals shown in Fig. 9b. The figure shows two examples: the one on top belongs to a case, where two distinct spikes are located on the PS-specific reconstructed TF data (see the two dark vertical lines in Fig. 9c), and are marked by two red arrows in Fig. 9b. An epileptologist confirmed that those identified locations are indicating epileptic spikes. The plots on the bottom of Fig. 9b, c show a case, where the PS-specific reconstructed TF data does not indicate any spikes as also are confirmed by a epileptologist. Comparing the two plots in Fig. 9b, it can be seen that the two EEG signals look very similar; however, the proposed JLNMF is able to successfully locate the epileptic spikes. To compare the JLNMF algorithm with the standard NMF, the TF data of NS and PS classes are separately decomposed using NMF with $r = 40$. The TF bases of each class is shown in Fig. 10. The NS and PS TF bases are compared to separate the spike-specific TF bases from the shared ones, but none had a correlation value of greater than 0.9 and only three had a value of greater than 0.8. Hence, the standard NMF is unable to locate any spike-related TF bases without any further analysis, while JLNMF showed to be successful in identifying the epileptic spikes. Such a method is deemed to be necessary for reliable evaluation of features associated with HYPS, which could potentially improve the assessment of infantile spasms, which is of significant importance in the therapy, management and ultimately the success of the prescribed treatment.

4.2.2 *Discrimination of pathological voice disorder*
Dysphonia or pathological voice disorder refers to speech problems resulting from a damage to or malformation of the speech organs. Pathological voice disorder is more common in people who use their voice professionally, for example, teachers, lawyers, salespeople, actors, and singers, and it dramatically affects these professional groups' lives both financially and psychosocially. The purpose of the discrimination of pathological voice disorder is to help patients with pathological problems for monitoring their progress over the course of voice therapy. We applied the developed JLNMF method to the Massachusetts Eye and Ear Infirmary (MEEI) voice disorders database, distributed by Kay Elemetrics Corporation [29]. The database consists of 51 normal and 161 pathological speakers whose disorders spanned a variety of organic, neurological, traumatic, and psychogenic factors. The
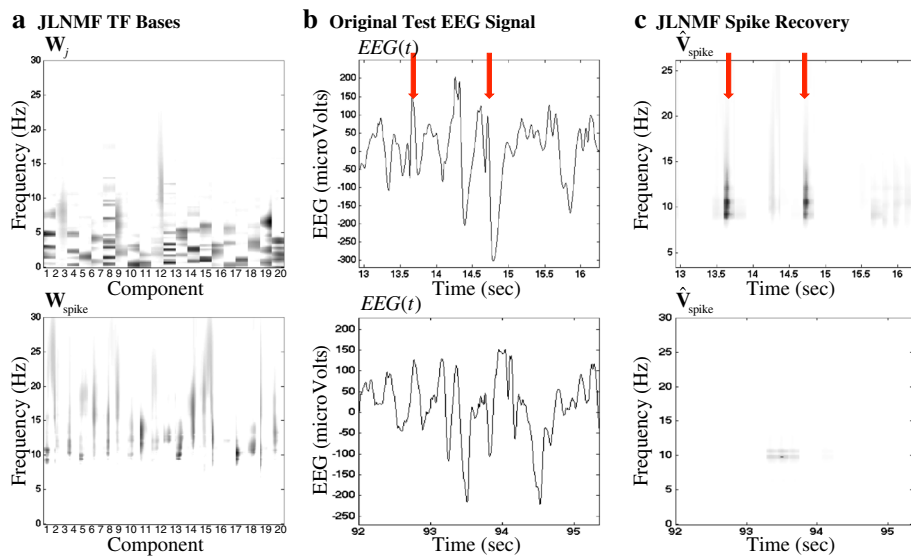
Fig. 9 The TF signals of two classes (NS and PS) are simultaneously decomposed using the JLNMF algorithm. **a** The shared TF bases between PS and NS classes (i.e., $\mathbf{W}_j$) and the PS-specific TF bases ($\mathbf{W}_{spike}$). **b** Two EEG recording samples (*top with two spikes* and *bottom with no spikes*). **c** The PS-specific reconstructed TF data. The *red arrows* indicate epileptic spikes

speech signal is sampled at 25 kHz and quantized at a resolution of 16 bits/sample. In this exploratory experiment, one speech signal from a normal subject and one from a pathological subject were used as two classes: normal and pathology. The TF signals of each class was constructed by computing the spectrogram with FFT size of 1024 points and Kaiser window with parameter of 5, length of 256 samples and 220 samples overlap. The two TF signals were then fed to the JLNMF algorithm to generate three sets of TF bases: normal-specific TF bases, pathological-specific TF bases, and shared TF bases. Figure 11 shows the above procedure along with the decomposed TF bases. The speech samples and their corresponding TF data are shown in Fig. 11a–d. The decomposed normal-specific TF bases, pathological-specific TF bases, and shared TF

bases are shown in Fig. 11e–g, respectively. For a successful decomposition, it is expected that the normal discriminant bases represent stronger formants compared to the pathological discriminant bases and the shared bases represent the natural structure of a speech signal [9]. The success of the JLNMF for discrimination of pathological voice disorder was visually investigated from the decomposed TF bases shown in Fig. 11. As expected, the pathological discriminant bases (Fig. 11f) present weak formants, while the normal discriminant bases (Fig. 11e) have more periodicity in low frequencies and introduce stronger formants. The shared TF bases (Fig. 11g) represent the low-frequency TF structures that is common to any natural speech regardless of being normal or pathological.
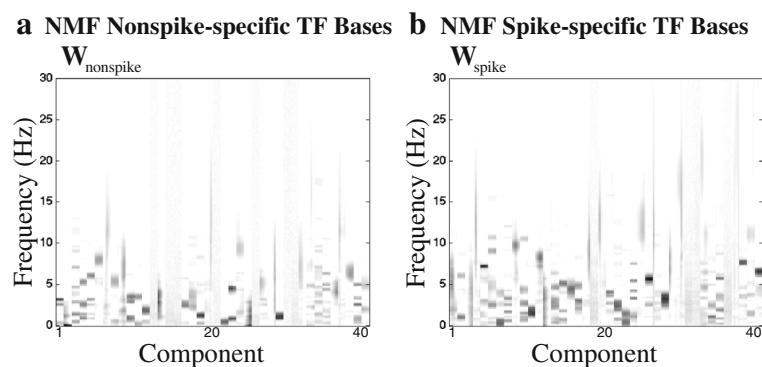

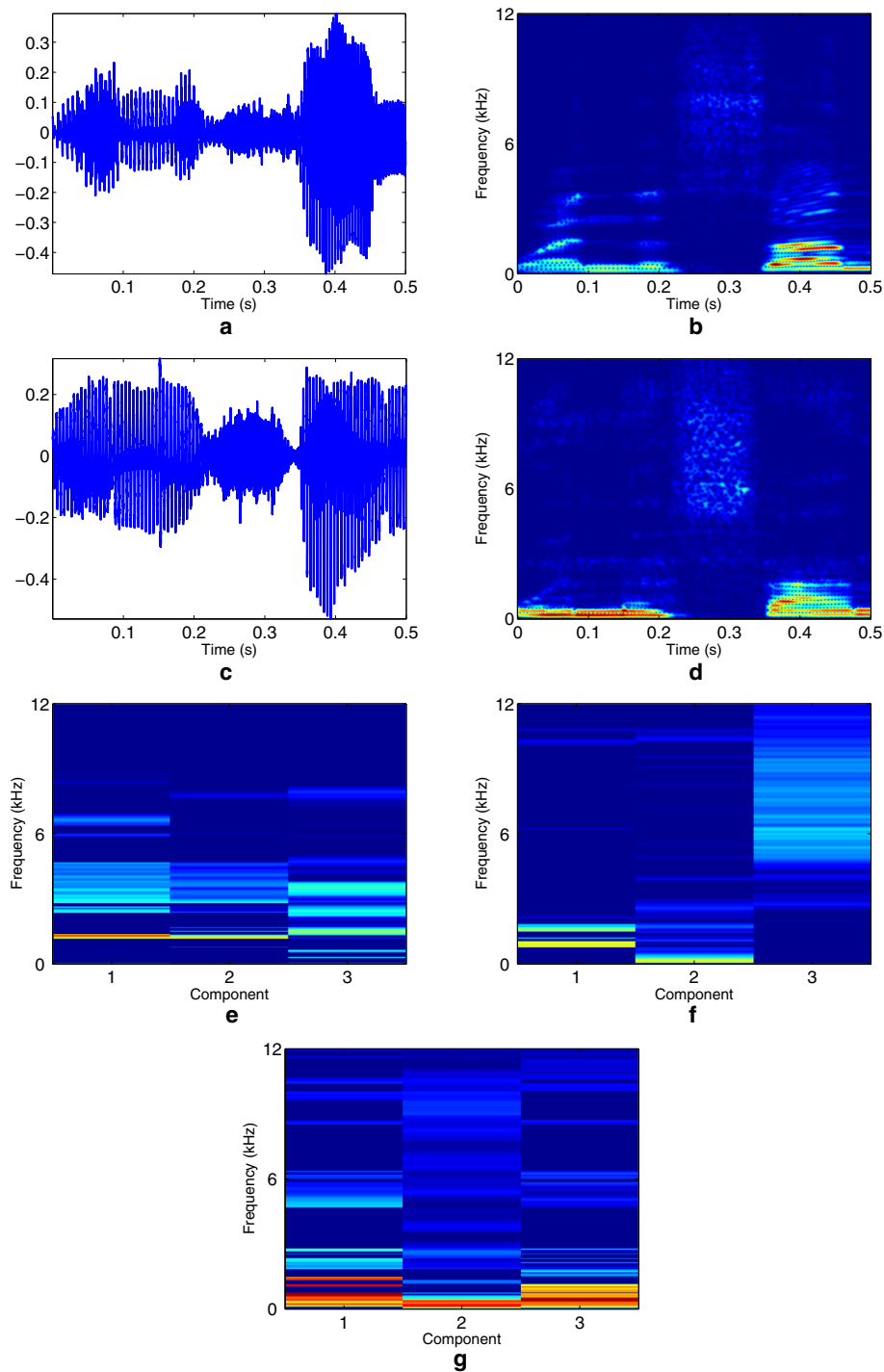**Fig. 10** NMF decomposed TF bases for NS and PS classes

Fig. 11 The JLNMF method was applied to two signals: one from a normal and one form a pathological voice subject. **a** A 0.5-s segment of a normal subject. **b** TF signal of the segment shown in (**a**). **c** A 0.5-s segment of a pathological voice disorder subject. **d** The TF signal of the pathological subject shown in (**c**). **e** Normal-specific TF bases. **f** Pathological-specific TF bases. **g** Shared-TF bases

5 Conclusions

In most real-world applications, the nature of signals from different classes are very similar, and there could only be slight changes in the TF patterns of signals from different classes. Inspired by this observation, a set of shared TF bases along with class-specific TF bases, was considered in the proposed JLNMF algorithm, and a new cost function was introduced to enforce discrimination between

shared and class-specific TF bases. A projected gradients framework was performed to optimization the JLNMF cost function. It was shown that the proposed JLNMF is invariant to amplitude scaling and time shifting, which is required for signal recognition applications. The performance of the proposed JLNMF was evaluated and compared to the results of standard NMF for data retrieval and localization applications, verifying the effectiveness of JLNMF.

Appendix

The second derivative of Eq. (9) with respect to one variable while keeping the rest of matrices constant is found by taking the derivative of Eqs. (15)–(19):

$$\nabla^2 g(\mathbf{W}_j) = (1 - \lambda) \left(\mathbf{H}_{1d}\mathbf{H}_{1d}^{\mathrm{T}} + \mathbf{H}_{2d}\mathbf{H}_{2d}^{\mathrm{T}} \right) \text{ and}$$

$$\nabla^2 g(\mathbf{W}_{1d}) = (1 - \delta - \lambda) \, \mathbf{H}_{1d}\mathbf{H}_{1d}^{\mathrm{T}}$$

$$\nabla^2 g(\mathbf{W}_{2d}) = (1 - \delta - \lambda) \, \mathbf{H}_{2d}\mathbf{H}_{2d}^{\mathrm{T}} \text{ and}$$

$$\nabla^2 g(\mathbf{H}_{1d}) = (1 - \delta - \lambda) \, \mathbf{W}_{1d}^{\mathrm{T}}\mathbf{W}_{1d} + \mathbf{W}_j^{\mathrm{T}}\mathbf{W}_j$$

$$\nabla^2 g(\mathbf{H}_{2d}) = (1 - \delta - \lambda) \, \mathbf{W}_{2d}^{\mathrm{T}}\mathbf{W}_{2d} + \mathbf{W}_j^{\mathrm{T}}\mathbf{W}_j$$

References
1. L Cohen, Time-frequency distributions—a review. Proc. IEEE. **77**, 941–981 (1989)
2. SG Mallat, Z Zhifeng, Matching Pursuits with Time-Frequency Dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)
3. B Boashash, G Azemi, NA Khan, Principles of time–frequency feature extraction for change detection in non-stationary signals: Applications to newborn EEG abnormality detection. Pattern Recognit. **48**(3), 616–627 (2015)
4. O Yilmaz, S Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. **52**(7), 1830–1847 (2004)
5. R Hennequin, B David, R Badeau, in *Proc. ICASSP*. Score informed audio source separation using a parametric model of non-negative spectrogram, (Prague, Czech Republic, 2011), pp. 45–48
6. T Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio, Speech Lang. Process. **15**(3), 1066–1074 (2007)
7. C Févotte, N Bertin, J-L Durrieu, Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. Neural Comput. **21**(3), 793–830 (2009)
8. B Ghoraani, S Krishnan, Time-frequency feature extraction and classification of environmental audio signals. IEEE Trans. Audio, Speech Lang. Process. **19**(7), 2197–2209 (2011)
9. B Ghoraani, S Krishnan, A joint time-frequency and matrix decomposition feature extraction methodology for pathological voice classification. EURASIP J. Adv. Signal Process. **2009**(ID 928974), 11–1011552009928974 (2009)
10. S Zafeiriou, A Tefas, I Buciu, I Pitas, Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. IEEE Trans. Neural Netw. **17**(3), 683–695 (2006)
11. S Nikitidis, A Tefas, I Pitas, Projected gradients for subclass discriminant nonnegative subspace learning. IEEE Trans. Cybernet. **44**(12), 2806–2819 (2014)
12. DD Lee, HS Seung, *Algorithms for non-negative matrix factorization.* (Adv Neural Inf Proces Syst, 2001), pp. 556–562
13. MW Berry, M Browne, AN Langville, VP Pauca, RJ Plemmons, Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. **52**(1), 155–173 (2007)
14. I Buciu, Non-negative matrix factorization, a new tool for feature extraction: Theory and applications. Proc. Int. J. Comput. Commun. Control. **3**, 67–74 (2008)
15. C-J Lin, Projected gradient methods for nonnegative matrix factorization. Neural Comput. **19**(10), 2756–2779 (2007)
16. N Guan, X Zhang, Z Luo, D Tao, X Yang, Discriminant projective non-negative matrix factorization. PloS one. **8**(12), 8329 (2013)
17. J Yoo, M Kim, K Kang, S Choi, *Nonnegative matrix partial co-factorization for drum source separation*, (Dallas, Texas, USA, 2010), pp. 1942–1945
18. P Paatero, The multilinear enginea table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. J. Comput. Graph. Stat. **8**, 854–888 (1999)
19. M Chu, F Diele, R Plemmons, S Ragni, Optimality, computation, and interpretation of nonnegative matrix factorizations. SIAM J Matrix Anal, 4–8030 (2004). Citeseer
20. DP Bertsekas, On the goldstein-levitin-polyak gradient projection method. IEEE Trans. Automatic Control. **21**, 174–184 (1976)
21. AT Cemgil, Bayesian inference for nonnegative matrix factorisation models. Computational Intelligence and Neuroscience. **2009**, 17 (2009). article ID 785152
22. M Morup, LK Hansen, in *Proc. 17th European Signal Processing Conference (EUSIPCOŠ09)*. Tuning pruning in sparse non-negative matrix factorization, (Glasgow, Scotland, 2009), pp. 1923–1927
23. VY Tan, C Févotte, Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7), 1592–1605 (2013)
24. M Davy, C Doncarli, GF Boudreaux-Bartels, Improved optimization of time-frequency-based signal classifiers. Signal Process. Lett. IEEE. **8**(2), 52–57 (2001)
25. B Ghoraani, S Krishnan, Discriminant non-stationary signal features' clustering using hard and fuzzy cluster labeling. EURASIP J. Adv. Signal Process. **2012**(1), 1–20 (2012)
26. JM Pellock, R Hrachovy, *et al*, Infantile spasms: a us consensus report. Epilepsia. **51**(10), 2175–2189 (2010)
27. RA Hrachovy, JD Frost, P Kellaway, Hypsarrhythmia variations on the theme. Epilepsia. **25**(3), 317–325 (1984)
28. SA Hussain, G Kwong, SA Hussain, G Kwong, JJ Millichap, JR Mytinger, N Ryan, JH Matsumoto, JY Wu, JT Lerner, R Sankar, Hypsarrhythmia assessment exhibits poor interrater reliability: a threat to clinical trial validity. Epilepsia. **56**(1), 77–81 (2014)
29. K Lee, HF Schuknecht, Results of tympanoplasty and mastoidectomy at the Massachusetts Eye and Ear Infirmary. Laryngoscope. **81**, 529–43 (1971)