Incremental Basis Estimation
Adopting Global k-means Algorithm
for NMF-Based Audio Source Separation

Kisoo Kwon\textsuperscript{a}, Jong Won Shin\textsuperscript{b}, and Nam Soo Kim\textsuperscript{a}

\textsuperscript{a}Department of Electrical and Computer Engineering and INMC, Seoul National
University, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea
\textsuperscript{b}School of Electrical Engineering and Computer Science, Gwangju Institute of Science
and Technology, 123 Cheomdan-gwagiro, Buk-gu, Gwangju, Korea

Abstract

Nonnegative matrix factorization (NMF) is a data decomposition technique enabling to discover meaningful latent nonnegative components. Since, however, the objective function of NMF is non-convex, the performance of the source separation can degrade when the iterative update of the basis matrix in the training procedure is stuck to a poor local minimum. In most of the previous studies, the whole basis matrix for a specific source is iteratively updated to minimize a certain objective function with random initialization although a few approaches have been proposed for the systematic initialization of the basis matrix such as the singular value decomposition and k-means clustering. In this paper, we propose an approach to robust bases estimation in which an incremental strategy is adopted. Based on an analogy between clustering and NMF analysis, we estimate the NMF bases in a similar way to the global k-means algorithm popular in the data clustering area. Experiments on audio source separation showed that the proposed methods outperformed the conventional NMF technique using random initialization by about 1.93 dB and 2.34 dB in signal-to-distortion ratio when the target source was speech and violin, respectively.

Keywords: Nonnegative matrix factorization, basis training, incremental approach, global k-means, audio source separation
1. Introduction

Over the recent years, nonnegative matrix factorization (NMF) has been widely applied to many tasks, and it has shown impressive performances particularly in image and audio signal processing [1]-[22]. NMF is a sort of latent factor analysis technique for which unsupervised learning algorithms are used to discover part-based representations underlying the given nonnegative data. NMF is conceptually similar to other well-known matrix factorization or even data clustering techniques which can be expressed in a unified formulation [16], [23], [24]. NMF has shown certain benefits compared with the other factorization schemes such as the independent component analysis and principal component analysis (PCA) [1], [2], in the area including source separation [3], [4], and document classification [5]. Since the publication of [1], a number of attempts have been made to improve NMF under some specific conditions, which include Itakura-Saito NMF [2], sparse NMF [6]-[8], convolutive NMF [10], discriminative NMF [11]-[13], and so on.

Though NMF shows an impressive performance in several fields, one of its weakness is that the final result is so sensitive to the initial values of the bases [24]. Because the specified objective function of NMF is not convex, the optimized solution obtained from iterative update of the basis matrix can be stuck to a local minimum, which implies that the overall performance may significantly depend on the initial parameter values. For this reason, several previous works attempt to provide systematic ways to initialize the basis and encoding matrices such as the centroids of k-means clustering, principal component analysis (PCA), and singular value decomposition (SVD)-based methods [25]-[31]. Though some of these methods show a lower reconstruction error and a faster convergence speed than the random value initialization, the source separation performance are not of primary concern. Moreover, the SVD- and PCA-based methods can not support over-complete bases in which the number of bases is larger than the dimension of the input vector. Recently, an incremental approach inspired by Linde-Buzo-Gray (LBG) algorithm is proposed which doubles the number of bases in each step [9]. It shows promising source separation performance, but the number of basis is strictly restricted to be a power of 2.

The conventional vector quantization task can be interpreted as a special case of the matrix factorization where each basis vector corresponds to a

\footnote{Corresponding author. E-mail address: jwshin@gist.ac.kr}
codeword and only a single basis is activated at each time [23]. This analogy implies that the data clustering techniques can provide some useful cues for the initialization of the NMF bases. Unfortunately, however, conventional codebook training approaches such as the k-means clustering can only guarantee suboptimal solutions similar to the case of NMF bases estimation and the final centroids are sensitive to the initialization of the code vectors. In order to alleviate this difficulty, several modified k-means algorithms have been developed [32]-[34]. The core idea of these algorithms is to increase the number of code vectors gradually while optimizing a certain criterion so that the final result can be less dependent on the initial parameter values.

In this paper, we propose a novel approach to estimate the basis and encoding matrices for the NMF analysis. Exploiting the analogy between NMF analysis and data clustering, a systematic method for estimating the NMF basis matrix is proposed by combining the standard NMF basis training procedure and an efficient codebook learning algorithm. The proposed methods borrow an idea from the global k-means algorithm [34]. One of the prominent features of this algorithm is that it estimates the parameters incrementally, i.e., increases the number of bases by one at each step. In order to evaluate the performance of the proposed techniques, we carried out an experiment on target source separation. In the experimental result, we can see that the proposed methods outperformed other bases estimation methods.

2. NMF-based audio source separation

When NMF is applied to audio source separation, it approximates the magnitude or power spectra of a given mixture \( V \in \mathbb{R}_{+}^{M \times N} \) as the product of a basis matrix \( W \in \mathbb{R}_{+}^{M \times R} \) and an encoding matrix \( H \in \mathbb{R}_{+}^{R \times N} \) \((V \approx WH)\) where \( M \), \( N \), and \( R \) denote the number of frequency bins, short-time frames, and the number of basis vectors, respectively. In this paper, \( V \) consists of magnitude spectra of audio signals. In order to resolve the non-unique factorization problem, it is needed to impose some constraints on the structures of \( W \) or \( H \). In our work, all the column vectors of \( W \) are constrained to have a unit \( L_2 \)-norm. The process of NMF-based source separation is given in Fig. 1. In this case, the basis matrix \( W \) is considered as a concatenation of the target and noise basis matrices, \( W_S \in \mathbb{R}_+^{M \times R_s} \) and \( W_N \in \mathbb{R}_+^{M \times R_n} \) where \( R_s \) and \( R_n \) indicate the number of target signal and noise basis vectors, respectively. \( W_S \) and \( W_N \) are usually trained separately with clean target signal and noise DBs, respectively. The objective function
of NMF is given as the discrepancy between $V$ and $WH$, i.e.,

$$f(W,H) = D(V|WH)$$ (1)

where $D(a | b)$ denotes the divergence between $a$ and $b$. One of the popular choices for the divergence measure is Kullback-Leibler divergence (KLD) which is given as [1]

$$D(V|WH) = \sum_{m,n} V_{m,n} \log \frac{V_{m,n}}{(WH)_{m,n}} - V_{m,n} + (WH)_{m,n}$$ (2)

where $A_{m,n}$ denotes the $m$-th row and $n$-th column component of the matrix $A$. When the multiplicative update rule (MuR) [1] is used, $W$ and $H$ are updated as follows:

$$H_{r,n} \leftarrow H_{r,n} \frac{\sum_{k=1}^{M} W_{k,r} V_{k,n}}{\sum_{k=1}^{M} W_{k,p}}$$ (3)

$$W_{m,r} \leftarrow W_{m,r} \frac{\sum_{p=1}^{N} H_{r,p} V_{m,p}}{\sum_{p=1}^{N} H_{r,p}}$$ (4)

The final estimate for $H$ and $W$ are obtained by iterative application of the update rules (3) and (4) for a fixed number of iterations. The MuR is an well-known approach to estimate $W$ and $H$ which is simple to implement and shown to yield good results. Each basis matrix, $W_S$ and $W_N$, is obtained separately by (3) and (4).

In the separation phase, a noisy magnitude spectrum $|Y(t)|$ is approximated as $|Y(t)| \approx WH(t)$ for each frame with the fixed basis matrix $W = [W_S \ W_N]$ obtained during the training phase where $H(t) = [H_S(t)^T \ H_N(t)^T]^T \in \mathbb{R}^{(R_s+R_n) \times 1}$ denotes the encoding vector of the mixed signal in the $t$-th frame, $Y(t)$ represents the short-time Fourier transform (STFT) coefficients of the noisy input, and $| \cdot |$ denotes taking element-wise magnitude. Keeping $W$ fixed, $H(t)$ is computed by iterating (3) for a fixed number of times, in which $H_S(t)$ and $H_N(t)$ are initialized to nonnegative random numbers. After a fixed number of iterations, the magnitude spectra of the target and noise signals are estimated as follows:

$$|\hat{S}(t)| = W_S H_S(t), \quad |\hat{N}(t)| = W_N H_N(t).$$ (5)
Instead of directly using the estimated magnitude spectra in (5), a spectral gain function similar to the Wiener filter is adopted in [2]-[4] and [20]. In this scheme, the gain function is given by

\[
G(t) = \frac{|\hat{S}(t)|}{|\hat{S}(t)| + |\hat{N}(t)|}
\]

where \( \frac{A}{B} \) denotes element-wise division of vectors, and \( G(t) \in \mathbb{R}^{M \times 1} \) is a gain vector obtained at the \( t \)-th frame. Finally, the STFT coefficients of the enhanced speech signal at the \( t \)-th frame are obtained according to \( \hat{S}_{\text{final}}(t) = G(t) \otimes Y(t) \) where \( \otimes \) indicates an elementary-wise multiplication.

3. Global k-means clustering based NMF bases estimation

In this section, we propose a novel approach to estimate NMF bases, which is based on an analogy between the NMF basis training and the codebook design in vector quantization. If the encoding vector of the NMF analysis is allowed to have only one non-zero component, then each NMF basis
can be viewed as a codeword vector and the reconstruction error can be treated as the distance between the input vector and its nearest codeword. Our approach to NMF bases estimation is motivated by the global k-means (GKM) clustering technique [34], which has demonstrated smaller clustering error than several other variants of the k-means clustering approach. In general, for data clustering, we need to find $R$ codewords and a rule to map any $M$-dimensional input vector into one of the $R$ codewords for the sake of minimizing the sum of the squared Euclidean distances between each input vector and the corresponding code vector. GKM starts with one cluster ($R = 1$) for which the optimal codeword is set as the centroid of the whole data set. At each iteration of the GKM algorithm, a new codeword is added to refine the clusters and the conventional k-means algorithm is run until convergence.

In the proposed approach, the NMF bases are trained incrementally like the GKM technique while the iterative update rules in (3) and (4) are adopted instead of the k-means clustering operation. To illustrate the proposed method, the geometric analysis of NMF proposed in [35] and [36] is employed in Fig. 2. The left figure shows an example for the data space with $M = 3$, where all the nonnegative data vectors and basis vectors can be mapped onto the points on the standard simplex if they are normalized to have unit $L_1$ norms. The figure on the right side shows the standard simplex, where red stars represent current basis vectors and blue dots denote normalized training data. The data vectors that reside in the convex hull formed by the basis vectors can be perfectly reconstructed by the NMF analysis. In contrast, the normalized data vector far from this convex hull
Figure 3: Pseudo code for the proposed incremental method to the NMF basis estimation.

**Input:** Matrix $V = (V_1, \cdots, V_N) \in \mathbb{R}^{M \times N}$, integer $R$

**Output:** Matrix $W \in \mathbb{R}^{M \times R}$

1. $W^1 = \frac{c(V)}{\|c(V)\|_2} 1$
   
   $H^1 = W^1 T V$

   for $r = 1 : R - 1$

   2. Initialize $W^{r+1}$ and $H^{r+1}$

   1) Find the vector $V_n^*$ with the maximum Euclidean distance from the reconstructed vector

   2) $W^{new} = \frac{V_n^*}{\|V_n^*\|_2} 1$
   
   $H^{new} = \|V_n^*\|_1 e^T_n$

   3) $W^{r+1} = [W^r, W^{new}] \in \mathbb{R}^{M \times (r+1)}_+$,
   
   $H^{r+1} = [H^r T, H^{new} T] T \in \mathbb{R}^{(r+1) \times N}_+$

3. Iteratively update $W^{r+1}$ and $H^{r+1}$ with (3) and (4) while keeping $(r + 1 - L)$ leftmost columns of $W^{r+1}$ unchanged

4. $W = W^R$

will have a large reconstruction error in the NMF analysis. In the proposed algorithms, the training data that shows the maximum Euclidean distance from the reconstructed vector is appended to the current basis matrix $W$ as a new basis vector. With this augmented basis matrix as an initial value, the conventional update algorithm in (3) and (4) is iterated until convergence. One possible drawback may be that the basis vectors added earlier will be updated much more times than newer basis vectors. In [37], the basis matrix updated to many times degrades the performance of the sources separation. To alleviate this problem, only recent $L$ basis vectors are updated in each step.

The pseudo code of the proposed incremental method to NMF basis estimation is given in Fig. 3. The input of the algorithm is the training data matrix $V = (V_1, \cdots, V_N) \in \mathbb{R}^{M \times N}_+$ and the number of bases $R$, while the output is the trained basis matrix $W$. Let $W^r$ and $H^r$ respectively denote the basis matrix and the corresponding encoding matrix when the number of bases is $r$. We begin with $r = 1$ and increase it at each iteration until $r = R$. 

7
The proposed method proceeds as follows:

1. Initialize $W^1$ and $H^1$; Compute the centroid $c(V)$ of $V$, the entire training set. The centroid is computed as the mean of the whole training data vectors. Set $W^1$ as the normalized centroid, i.e.,

$$W^1 = \frac{c(V)}{\| c(V) \|_2} \mathbf{1}$$

where $\| \cdot \|_2$ and $\mathbf{1}$ represent $L_2$-norm and a vector with all elements equal to one, respectively. It is easy to find $H^1 = W^1^T V$.

2. 1) Let $V_n^*$ be the training vector showing the maximum Euclidean distance from the reconstructed vector. $V_n^*$ is decided as

$$V_n^* = \arg\max_{V_n} \| V_n - W^r H^r(n) \|_2$$

where $H^r(n)$ denotes the $n$-th column of $H^r$.

2. 2) A new basis vector $W^{new} \in \mathbb{R}_+^{M \times 1}$ and the corresponding encoding vector $H^{new} \in \mathbb{R}_+^{1 \times N}$ are determined as

$$W^{new} = \frac{V_{n^*}}{\| V_{n^*} \|_2} \mathbf{1}, \quad H^{new} = \| V_{n^*} \|_2 e_{n^*}^T$$

where $e_{n^*}$ denotes the standard vector having all its elements zero except for the $n^*$-th element which is set to 1.

2. 3) Increase the number of bases by one for which

$$W^{r+1} = [W^r, W^{new}] \in \mathbb{R}_+^{M \times (r+1)}$$

$$H^{r+1} = [H^r^T, H^{new^T}]^T \in \mathbb{R}_+^{(r+1) \times N}$$

where $H^r$ is the same as $H^r$ except the $n^*$-th column is replaced by a zero vector, as shown in Fig. 4. $W^{r+1}$ and $H^{r+1}$ will perfectly reconstruct $V_{n^*}$ as a result.

3. Iteratively update the $H^{r+1}$ and $L$ rightmost columns of $W^{r+1}$ according to (3) and (4) for a fixed number of times to approximate $V$ as $W^{r+1} H^{r+1}$. $(r + 1 - L)$ leftmost columns of $W^{r+1}$ are kept intact to prevent too many updates of each basis.

4. Increase $r$ by one and repeat 2 and 3 if $r < R$. 8
Figure 4: Construction of $H^{r+1}$ for the perfect reconstruction of $V_n$. 

4. Experimental setup

To evaluate the performance of the proposed algorithms, experiments on audio source separation were performed in a variety of noisy conditions with the target source being speech and violin signals. A 512-point discrete Fourier transform with 75% overlap was used to form the spectrogram with Hamming window. The sampling rate was 16kHz. The magnitude spectra were used as data vectors for the NMF analysis. Speech and interfering signals were selected from TIMIT [40] and NOISEX-92 [41] DBs, respectively. The basis matrix for each noise type was obtained from about 120-second long noise signal, and the training speech DB was 130-second long spoken by 56 different speakers. The test speech data set consisted of 32 sentences from 32 different speakers. We tested 4 different types of noises including F-16, factory1, babble, and white noises, while the signal-to-noise ratio (SNR) was 0 dB. For violin signals, we used two music files for the training and separation phases, as shown in Table. 1. There was no overlap between the training and test data.

The performance of the proposed approach was evaluated in terms of the perceptual evaluation of speech quality (PESQ) score [38] and the signal-to-
Table 1: Information on the violin signals used for the bases estimation and source separation (resampled to 16kHz/s).

<table>
<thead>
<tr>
<th>Title</th>
<th>Artist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partita No.1 (BWV 1002)-Double</td>
<td>Ida Haendel</td>
</tr>
<tr>
<td>Sonata No.2 (BWV 1003)-Allegro</td>
<td></td>
</tr>
</tbody>
</table>

distortion ratio (SDR) [39]. To demonstrate the performance improvement achieved by the proposed method, seven versions of the NMF-based source separation algorithm described in Section II for which only the training methods of the basis matrix differed were compared:

- **Rand**: NMF basis estimation with random initialization [1]
- **SNMF**: Sparse NMF [8] with random initialization
- **PCA**: NMF basis estimation that utilizes principal component analysis [30]
- **SVD**: NMF basis estimation that employs singular value decomposition [31]
- **Cent**: NMF basis estimation based on the centroids of the k-means clustering [29]
- **LBG**: incremental NMF basis estimation based on Linde-Buzo-Gray algorithm [9]
- **Prop**: proposed incremental approach to the NMF basis estimation based on GKM algorithm (Fig. 3)

**SNMF** is the NMF with a constraint on the $L_1$ norm of $H$, while the $L_2$ norm normalized version of $W$ is adopted in the objective function [8]. It is included in the experiments as a performance benchmark. In **Cent**, the number of clusters was kept the same as the number of the NMF bases [8]. The number of iterations during the training phase was decided to maximize the separation performance for each algorithm as in [12], [37], which turned out to be 100 for **Rand**, **SNMF**, **PCA**, **SVD**, and **Cent** and 10 for each step of **LBG**. For each step of **Prop**, the number of iterations was 1 for violin signal and 5 for speech and noises. The number of iterations in the
source separations was 20. The number of basis vectors that are updated at each step of Prop, \( L \), was set to 10. The separation performances of Prop were not sensitive to \( L \) in our experiments. For Rand and SNMF, the performances were averaged over 10 trials with different random seeds. The performance deviations depending on the random seeds were less than 0.09 dB in SDR and 0.008 in PESQ scores.

5. Results and discussion

5.1. Separation of speech from noise

Fig. 5 shows the source separation performance with various number of the bases \((R = 64, 128, 256, 512)\) averaged over all noises. Only 5 results are shown for \( R = 512 \), because PCA and SVD cannot be applied when \( R > M \). Fig. 6 summarizes the performance when \( R \) for each method was selected to maximize the performance. PCA, SVD, and Cent did not show improvement of source separation performance, although they might describe each source well. SNMF performed well as expected. The incremental approach adopting LBG algorithm showed similar performance when \( R \) was 256 and better PESQ scores when it was 512 compared with SNMF, which demonstrated the potential of the incremental approach. The proposed algorithm using the GKM-based incremental approach outperformed all the competitors for all value of \( R \). Prop outperformed SNMF about 0.68 dB and 0.07 in terms of SDR and PESQ score, respectively.

5.2. Separation of violin signal from noise

Fig. 7 shows the source separation performance when the target source was violin. Like Fig. 5 and 6, Fig. 7 (a) shows the SDRs for various \( R \) and Fig. 7 (b) shows the performances with the values of \( R \) that produced the highest SDRs. The tendency was similar to the previous experiments in which the target signal was speech. Prop outperformed SNMF and LBG about 1.58 and 1.51 dB in terms of SDR.

5.3. Convergence curve and processing time

Additionally, we have tested the convergence curve of the reconstruction error for the speech signal. Fig. 8 shows the reconstruction errors, \( D(V|WH) \) in (2), according to the number of iterations with various basis training methods when \( R = 64 \). As for LBG and Prop, the numbers of iterations are those in the last stages in which \( R = 64 \), i.e., the iteration numbers after
Table 2: The processing time. ($R = 64$)

<table>
<thead>
<tr>
<th>methods</th>
<th>Rand</th>
<th>SNMF</th>
<th>PCA</th>
<th>SVD</th>
<th>Cent</th>
<th>LBG</th>
<th>Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>time (s)</td>
<td>11.98</td>
<td>17.77</td>
<td>12.24</td>
<td>17.99</td>
<td>18.24</td>
<td>14.45</td>
<td>27.46</td>
</tr>
</tbody>
</table>

we have 64 initial bases. At the same number of iterations of 100, **PCA** showed the lowest reconstruction error, and **LBG** and **Prop** showed similar values to **Rand**. However, at the numbers of iterations that resulted in the best source separation performances, the reconstruction error for **LBG** and **Prop** were even higher, which indicates that lower reconstruction error does not lead to better source separation. The reconstruction error for **SNMF** was high, as **SNMF** tries to minimized not only the reconstruction error but also the $L_1$ norm penalty term.

The total processing time to train the NMF bases using various methods are shown in Table 2, for which $R = 64$ and the parameters were set to maximize the source separation performances. All the algorithms were implemented using Matlab. **Prop** requires the longest processing time as it is composed of 64 stages. It is noted that the processing time in the source separation phase is the same for each method.

6. Conclusion

This paper proposed novel basis estimation method for NMF based on the incremental strategy. The proposed method estimate the bases incrementally while considering the reconstruction error of each data vector. Inspired by the global k-means algorithm, the data vector with the largest reconstruction error is added as a new basis vector before performing conventional NMF updates. The experimental results confirm that the incremental approach based on global k-means algorithm is effective for audio source separation.

Acknowledgment

This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (NRF-2015R1A2A1A15054343) and the research fund of Signal Intelligence Research Center supervised by the Defense Acquisition Program Administration and Agency for Defense Development of Korea.


Figure 5: The source separation performance with various basis training methods according to the number of basis vectors. (target: speech, input SNR = 0 dB)
Figure 6: The performance of source separation for various basis training methods with $R$ that showed best performance. (target: speech, input SNR = 0 dB)
Figure 7: The source separation performance with various basis training methods. (target: violin, input SNR = 0 dB)
Figure 8: The convergence curve with various basis training methods. (speech signals)