International Journal of Applied Mathematics

Volume 36 No. 3 2023, 425-436
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
doi: http://dx.doi.org/10.12732/ijam.v36i3.8

SEPARATION OF TWO MUSICAL INSTRUMENTS USING MATRIX FACTORISATION TECHNIQUES

Wen Kai Adrian Tang[1] §, Wei Shean Ng[2], How Hui Liew[3]

[1,2,3] Universiti Tunku Abdul Rahman
Lee Kong Chian Faculty of Engineering & Science
Department of Mathematical and Actuarial Sciences
Sungai Long Campus, Jalan Sungai Long,
Bandar Sungai Long, Cheras 43000
Kajang, Selangor, MALAYSIA

Abstract: Source separation is important in audio processing. This research focuses on musical instrument source separation. The methods used are Interpolative Decomposition (ID), Nonnegative Matrix Factorisation (NMF) and Convolutive Matrix Factorisation (CNMF). These three matrix factorisations are simple algorithms used for extracting features for image processing. The performances of NMF, CNMF and ID are compared when applying them to musical instrument source separation. Signal-to-noise ratio, Similarity Index and Residual Energy are used to measure the performance of each method. Numerically, Nonnegative Matrix Factorisation with Kullback Leibler divergence is found to have performed better. However, in theory, the Itakura-Saito divergence variant of NMF and CNMF is recommended for solving music-related source separation.

AMS Subject Classification: 94A12; 15A23; 00A65
Key Words: audio signal processing; nonnegative matrix factorisation; convolutive nonnegative matrix factorisation; interpolative decomposition

1. Introduction

Signal processing is a sub-field of electrical engineering. In the real world, we

Received: May 18, 2023

§Correspondence author

can represent images, audio, videos and vibrations in signals and this problem is known as audio compression, speech recognition, source separation, etc. In this paper, we investigate the method of performing the signal audio source separation and the musical instrument data is chosen as the signal audio data. Source separation is a method to recover two or more audio from the mixed audio. Hence, the signal audio source separation formulation is shown below

$$y(t) = \sum_{j=1}^{N} x_j(t),$$

where $y(t)$ denotes the mixed audio signal, $x_j(t)$ denotes the j original audio signals in the mixed audio signal where $j = 1, 2, \ldots, N$ and N takes the positive integer values. We also can represent the audio signal in matrix form by using the Short-Time Fourier Transform (STFT) as in [3]. Therefore, we apply matrix factorisation to solve the musical instrument source separation problem. We choose Nonnegative Matrix Factorisation (NMF), Convolutive Nonnegative Matrix Factorisation (CNMF) and Interpolative Decomposition (ID). Then, we apply the matrix factorisation to the musical instrument source separation problem to compare the performance of each method in solving that problem.

2. Related works

Back in 1994, Paatero and Tapper introduced the Positive Matrix Factorisation [20], and later it is a well-known method after Lee and Seung popularised it and it is called Nonnegative Matrix Factorisation (NMF) [6]. NMF splits a matrix A into W and H to compute an approximated matrix A. In order to calculate W and H, multiplicative updates rules are proposed by [6]. This mentioned NMF is categorised as Basic NMF and there are other NMF methods which were reviewed in [26].

NMF has been applied in the field of source separation. NMF was used to solve the real number domain until it was extended to the complex domain where the combination of NMF and Sparse Coding were used [10]. In 2013, positive semidefinite tensor factorisation (PSDTF) was proposed and it was extended from NMF with IS divergence [15]. The proposed method can perform audio separation directly in the time domain but the disadvantage is that the computational time is expensive.

The methods previously mentioned are suitable for single-channel source separation. Then, some researchers expand it to solve multi-channel source

separation. [14] stated that humans and animals have two ears and can attain information on the audio signal direction and location which is the idea for multi-channel source separation. [24] extended the application of NMF in solving the multi-channel source separation. Besides this, [2] proposed two multi-channel models which are Expectation-maximisation (EM) and multiplicative update (MU) rules algorithm by using the convolutive mixing but the computational cost is expensive.

Another researcher [13] suggested Hermitian positive semidefiniteness of a matrix which is the same as nonnegativity, which is another extension of NMF for solving the multi-channel source separation in the complex domain by setting the EM algorithm as the cost function. [11] and [12] can only separate two audio sources and their methods have difficulty to have good separated audio when dealing with more than two audio sources. Hence, [13] proposed the bottom-up clustering method with conditions where redundant spatial properties are allowed then it can separate the three sources. The authors also mentioned that the multi-channel algorithm and single-channel algorithm are quite similar where both are iterative methods.

The disadvantage of NMF with multiplicative update rules is when solving the high-dimensional matrix the convergence rate reduces. [7] sees this problem and suggested the first-order primal-duel algorithm for NMF with KL divergence by using the Chambolle-Pock algorithm as the base. Thus the result shows us that this method has a faster convergence rate as compared to the traditional update rules [6] and alternating direction methods of multiplier [4].

Besides using the NMF method in source separation, some researchers also use deep learning techniques in source separation. [25] suggested Conv-TasNet in source separation and solving it in the time domain as there are some drawbacks when solving it in the frequency-time domain. Another method that applies in the time domain is Wave-U-Net [5] which is improvised from U-Net [1] which is used in the frequency-time domain. However, Wave-U-Net has a better performance based on their result. [18] suggested the multi-scale multi-band densenets modified from [8]. Normally, densenets are used in image processing and have good performance, hence the authors make some modifications to the method to apply in musical source separation.

3. Matrix factorisation

3.1. Nonnegative Matrix Factorisation (NMF)

Lee and Seung [6] defined NMF as a technique for splitting a matrix A into $W \in \mathbb{R}^{n\times k}$ and $H \in \mathbb{R}^{k\times m}$ where rank, k, is $0 \leq k \leq \min(n, m)$, i.e., $A \approx A' = WH$, where all the entries in W and H are non-negative and A' is an approximated matrix of A calculated by using the smaller dimension matrices W and H.

3.2. Convolutive Nonnegative Matrix Factorisation (CNMF)

CNMF is an extension of NMF using the convolutive structure and can capture the short-term temporal dependencies in the time series data [19]. The structure of the CNMF is shown below [21],

$$A \approx A' = \sum_{t=0}^{T-1} W(t) \cdot \overset{t\rightarrow}{H},$$

where $A \in \mathbb{R}^{n\times m}$ is the input audio, $W(t) \in \mathbb{R}^{n\times k}$ is the basis and $H \in \mathbb{R}^{k\times m}$ is the coefficient matrix of rank k, $0 \leq k \leq \min(n, m)$. All the entries in A, $W(t)$ and H are positive. On the other hand, the operator $\overset{t\rightarrow}{(\cdot)}$ shifts the columns of the matrix t steps to the right.

3.3. Interpolative Decomposition (ID)

ID is a matrix factorisation method using numerical analysis to compute the factorised matrix. ID preserves the structure of the matrix such as sparsity and nonnegativity whereas truncated Singular Value Decomposition is not able to. Thus, Advani and O'Hagan [23] defined it as given $A \in \mathbb{R}^{n\times m}$, then factorised it into $C \in \mathbb{R}^{n\times k}$ whose columns are chosen from the columns of A, and $Z \in \mathbb{R}^{k\times m}$ where the rank, k, is $0 \leq k \leq \min(n, m)$.

4. Methodology, results and discussion

4.1. Musical instrument audio datasets

In this research, we use two different sets of mixture audio datasets to test the algorithm. The first mixture of musical instrument audio consists of a drum

and a guitar extracted from different audio [9]. The mixture is done manually. This dataset is known as a "simple music datasets".

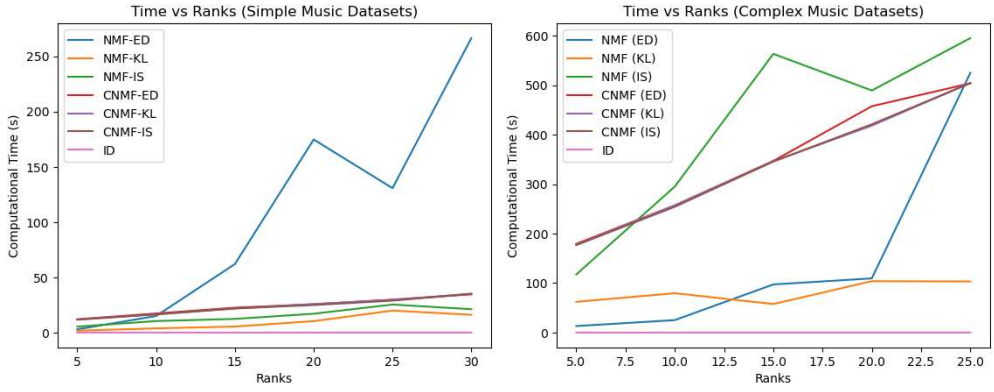
The second mixture of musical instrument datasets also consists of a drum and a guitar extracted from the song "Beat it" [16] by using MIDI software [27]. This dataset is known as a "complex music datasets".

4.2. Experimental steps

First, we import the signal data using librosa [3]. Then we perform Short-Time Fourier Transform (STFT) via librosa [3] to transform the imported signal data into a matrix representation. Next, we find the rank, k, using principal component analysis (PCA) and perform the matrix factorisation to get $A \approx WH$. Then, we can extract the individual component l by taking the outer product of the l columns of the W and l rows of the H and perform the inverse STFT to listen to the audio sound for all $l = 1, 2, \ldots, k$. Next, we categorise the l individual components into drum or guitar. Finally, we measure the performance using signal-to-noise (SNR) [17], similarity index (SI) [22] and residual energy (RE) [22].

4.3. Computational time

Time is important to determine which algorithms are fast, as society needs fast algorithms to solve real-world problems. Hence, we plot the rank vs time graph.



(a) Simple Music Datasets. (b) Complex Music Datasets.

Figure 1: Computational time of each method.

Figure 1 shows the computational time needed for each rank in both simple and complex music datasets. Simple music datasets have a shorter music audio hence it has a smaller matrix as compared to complex music datasets. Due to the difference in size, we observe that the y-axis in Figure 1(a) is larger than the y-axis in Figure 1(b).

Next, we look into Figure 1(a) and we categorise the methods into multiplicative updates-based methods and numerical-based methods. Only ID is the numerical-based method while others are the multiplicative updates-based methods. We can clearly see that those multiplicative updates-based methods require more computational time compared to the numerical-based methods. We also observe that NMF-ED has higher computational time as compared to other methods. On the other hand, the CNMF methods' computational time is similar for CNMF-ED, CNMF-KL and CNMF-IS, slightly higher than NMF-KL and NMF-IS. We also observe that ID computational time does not increase when the rank increases.

Figure 1(b) shows that NMF-IS computational time is the highest when rank > 7 followed by the CNMF-ED, CNMF-KL and CNMF-IS. On the other hand, NMF-ED has a low computational time but there are sudden increases at rank $= 25$ because of the initialise values of W and H is far from the convergence point, thus it required more computational time and iterations. While NMF-KL has a stable computational time. Then, we observe that ID computational time does not affect by the increase in ranks. Hence, this shows the advantages of ID in factorising a larger dimension matrix.

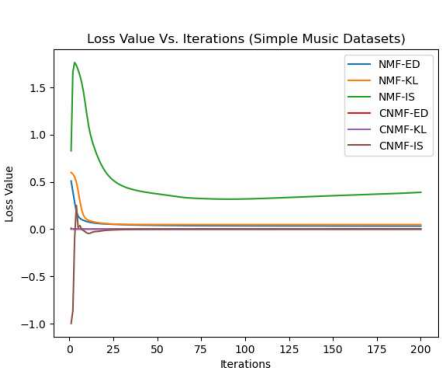
5. Convergence rate

The purpose to study the convergence rate is to see how fast the method converges to a value and the patterns of the convergence line. In NMF, we use the equation below to calculate the loss value between the original audio and the separation audio:

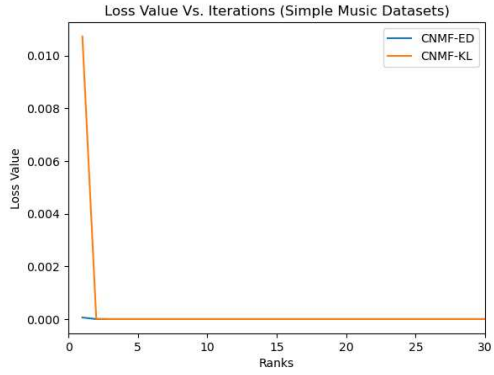
$$\text{Error} = \frac{(X_o - X_{rd})^2}{(X_o)^2},$$

where X_o denotes the original signal spectrogram and X_rd denotes the separated audio. On the other hand, the loss value of CNMF is directly obtainable from the algorithm.

From Figure 2 and Figure 3, we observe that they have almost similar convergence patterns where they converge to a value except for CNMF-IS. From both Figure 2(a) and Figure 3(a), we observe that NMF-ED, NMF-KL and

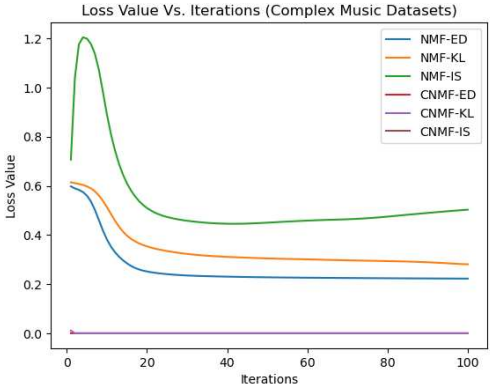


(a) Convergence Rate.

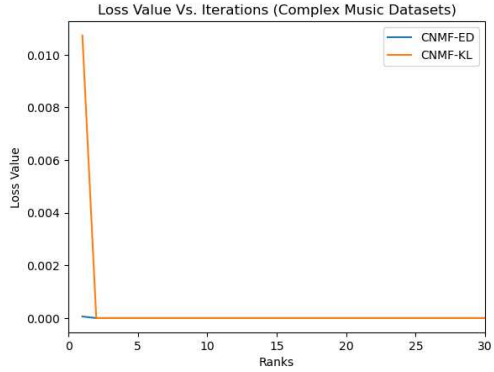


(b) Convergence Rate zoom in for CNMF-ED and CNMF-KL.

Figure 2: Convergence Rate for Simple Music Datasets.



(a) Convergence Rate.



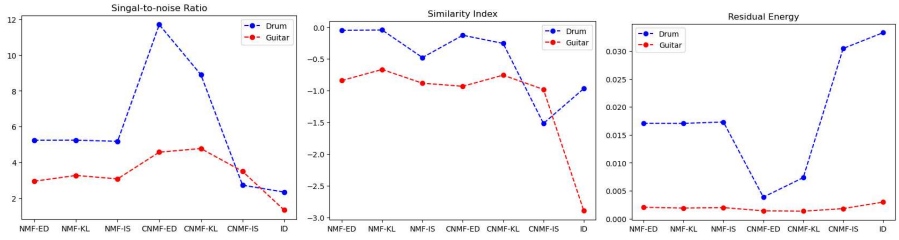
(b) Convergence Rate zoom in for CNMF-ED and CNMF-KL.

Figure 3: Convergence Rate for Complex Music Datasets.

NMF-IS continue to converge after 100 iterations. Thus, a stopping condition is needed. Another observation is that NMF-IS has a minimum point that the other does not have. When we enlarge the convergence rate of CNMF-ED and CNMF-KL in Figure 2(b) and Figure 3(b), we observe that they have the same convergence pattern and rate of convergence is faster than the NMF method. As for the case CNMF-IS in Figure 2(a) they have a similar pattern as NMF-IS but in Figure 3(a) they return the "nan" value in Python for CNMF-IS.

5.1. Performance of different methods in musical instrument audio source separation

In this section, we discuss the performance of NMF, CNMF and ID applied in both simple audio datasets and complex audio datasets. We start with simple audio datasets.

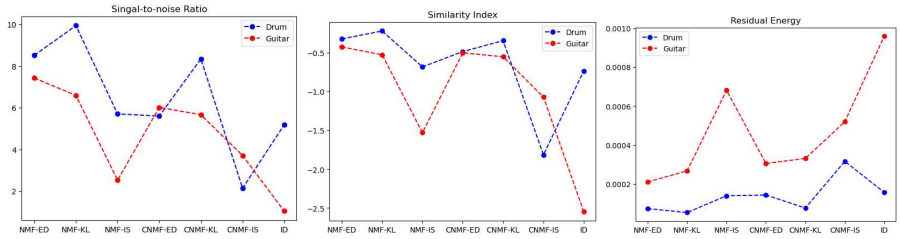


(a) SNR (b) SI (c) RE

Figure 4: Performance measurement of methods apply in simple audio datasets.

From Figure 4(a) and Figure 4(b), it is hard to determine which methods have a better performance. Thus, we first look at Figure 4(c) and we observe that there are small RE values on separated guitar audio among all the methods. On the other hand, as for separated drum audio, CNMF-ED has a lower RE value followed by CNMF-KL and all the NMF methods. Figure 4(b) and we can exclude the NMF-IS as it has a lower SI value compared to CNMF-ED, CNMF-KL, NMF-ED and NMF-KL. From Figure 4(a), we can say that CNMF-ED has a better performance but in Figure 4(b) the SI value of CNMF-ED is lower than NMF-ED and NMF-KL. Thus, this statement also applied to CNMF-KL. From this, we can conclude NMF-KL has better performance as compared to NMF-ED as it has a higher SI value in the separated guitar audio and both of them have similar values in both SNR and RE values.

Next, we look into the matrix factorisation applied in complex audio datasets.



(a) SNR (b) SI (c) RE

Figure 5: Performance measurement of methods apply in complex audio datasets.

Figure 5(c), we notice that both NMF-KL and CNMF-KL have a lower RE value even though the separated guitar audio RE value is a little bit high but not the highest. We also can consider NMF-ED as the top 3 best performance as it has a low RE value in separated guitar audio. Figure 5(b) we observe that NMF-KL have a higher SI value for separated drum compared to NMF-ED and CNMF-KL but the second highest SI value in separated guitar. Hence, we observe the same pattern in 5(a). Numerically, NMF-KL has better performance.

There is some information which the numerical value is not able to tell us and we need to hear the audio. In complex datasets, there is the cymbal sound in the drum audio, but NMF-ED and CNMF-ED are not able to reproduce the sound. On the other hand, NMF-KL and CNMF-KL are not able to when the rank is 11 but when the rank is increased to 19 the sound can be reproduced. As for NMF-IS and CNMF-IS, they possess the scale invariance property which can detect the cymbal sound.

Next, we look into the Interpolative Decomposition (ID) method in Figures 4 and 5, which show that the ID does not perform well in the musical instrument source separation. It may be the cause of choosing the columns of the original matrix randomly, A to be the columns of C, and $A \approx CZ$ is the structure of ID. Sometimes, the chosen columns contain duplicated information and unwanted information to be the columns of C.

6. Conclusion

In this paper, we compare the performance of NMF, CNMF and ID numerically instead of listening to identify which has better performance. As the human ears are not sensitive in determining the performance when there are high similarities between the separated audio and original audio. The numerical result shows us that NMF-KL performs better compared to other methods in both simple audio datasets and complex audio datasets. On the other hand, it is recommended to use the Itakura-Saito divergence variant of NMF and CNMF to solve the music-related source separation problem as it processes the scale invariance properties. The ID performs the worst in solving the problem and some modification is needed to solve musical instrument source separation. On the other hand, the computational time of ID is fast when computing large matrix dimensions compared to other methods.

In future, we can increase the number of musical instruments to three or more to test that algorithm using matrix factorisation. There are some limitations of the algorithm where the original audio sound must be known, but in real life, the original audio may not be known.

Acknowledgement

The research was supported by Universiti Tunku Abdul Rahman (UTAR) through Univerisiti Tunku Abdul Rahman Research Fund IPSR/RMC/UTARRF/2021-C1/N03.

References

[1] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, Singing voice separation with deep U-Net convolutional networks, In: *ISMIR*, Suzhou, China (2017).

[2] A. Ozerov and C. Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation, *IEEE Trans. Audio, Speech, Language Process.*, **18** (2010), 550–563.

[3] B. McFee et al., Librosa: Audio and music signal analysis in Python, In: *Proc. of the 14th Python in Science Conf.*, Austin, Texas (2015), 18–24.

[4] D. L. Sun and C. Févotte, Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence, In: *ICASSP*, Florence, Italy (2014), 6201–6205.

[5] D. Stoller, S. Ewert, and S. Dixon, Wave-U-Net: A multi-scale neural network for end-to-end audio source separation, In: *ISMIR*, Paris, France (2018), 334–340.

[6] D.D. Lee and H.S. Seung, Algorithms for non-negative matrix factorization, In: *NuerIPS*, Denver, Colorado (2000).

[7] F. Yanez and F. Bach, Primal-dual algorithms for non-negative matrix factorization with the Kullback-Leibler divergence, In: *ICASSP*, New Orleans, Louisiana (2017), 2257–2261.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, Densely connected convolutional networks, In: *CVPR*, Honolulu, Hawaii (2017), 2261–2269.

[9] *GitHub*. https://github.com/stevetjoa/musicinformationretrieval.com/tree/gh-pages/audio.

[10] H. Kameoka, Nobutaka Ono, Kunio Kashino, and Shigeki Sagayama, Complex NMF: A new sparse representation for acoustic signals, In: *ICASSP*, Taipei, Taiwan (2009), 3437 – 3440.

[11] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, New formulations and efficient algorithms for multichannel NMF, In: *WASPAA*, New Paltz, New York (2011), 153–156.

[12] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, Efficient algorithms for multichannel extensions of Itakura-Saito nonnegative matrix factorization, In: *ICASSP*, Kyoto Japan (2012), 261–264.

[13] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, Multichannel Extensions of non-negative matrix factorization with complex-valued data, *IEEE Trans. Audio, Speech, Language Process.*, **21** (2013), 971–982.

[14] J. Blauert, *Spatial Hearing*, The MIT Press, Cambridge (1996).

[15] K. Yoshii, R. Tomioka, D. Mochihashi, and M. Goto, Beyond NMF: Time-domain audio source separation without phase reconstruction, In: *ISMIR*, Curitiba, Brazil (2013).

[16] Michael Jackson - Beat It.mid — Free MIDI, *BitMidi*. https://bitmidi.com/michael-jackson-beat-it-mid.

[17] N. Detlefsen et al., TorchMetrics - Measuring Reproducibility in PyTorch, *J. Open Source Softw*, **7** (2022), 4101.

[18] N. Takahashi and Y. Mitsufuji, Multi-Scale multi-band densenets for audio source separation, In: *WASPAA*, New Paltz, New York (2017), 21–25.

[19] P.D. O'Grady and B. A. Pearlmutter, Convolutive non-negative matrix factorisation with a sparseness constraint, In: *MLSP*, Maynooth, Ireland (2006), 427–432.

[20] P. Paatero and U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics*, **5** (1994), 111–126.

[21] P. Smaragdis, Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs, In: *ICA*, Berlin, Heidelberg (2004), 494–499.

[22] P. Smaragdis, Convolutive speech bases and their application to supervised speech separation, *IEEE Trans Audio Speech Lang Process*, **15** (2007), 1–12.

[23] R. Advani and S. O'Hagan, Efficient algorithms for constructing an interpolative decomposition, *arXiv:2105.07076 [cs, math]* (2022).

[24] R. M. Parry and I. Essa, Estimating the spatial position of spectral components in audio, In: *ICA*, South Caroline, USA (2006), 666–673.

[25] Y. Luo and N. Mesgarani, Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation, *IEEE/ACM Trans. Audio, Speech, Language Process.*, **27** (2019), 1256–1266.

[26] Y.-X. Wang and Y.-J. Zhang, Nonnegative Matrix Factorization: A Comprehensive Review, *IEEE Trans Knowl Data Eng*, **25** (2013), 1336 – 1353.

[27] Signal, *signal.vercel.app*. https://signal.vercel.app/edit.