Atlantis Studies in Uncertainty Modelling, volume 3
Joint Proceedings of the 19th World Congress of the International Fuzzy Systems Association (IFSA), the 12th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), and the 11th International Summer School on Aggregation Operators (AGOP)

Fuzzy Clustering-based Switching Non-negative Matrix Factorization and Its Application to Environmental Data Analysis

*K. Honda[a] and T. Furukawa[a] and S. Ubukata[a] and A. Notsu[b]
[a]Graduate School of Engineering, Osaka Prefecture University, Sakai Osaka 599-8531 Japan, {honda,subukata}@cs.osakafu-u.ac.jp
[b]Graduate School of Humanities and Sustainable System Sciences, Osaka Prefecture University, Sakai Osaka 599-8531 Japan, notsu@cs.osakafu-u.ac.jp

Abstract

Non-negative matrix factorization (NMF) is a basic method for analyzing the intrinsic structure of non-negative matrices but cannot work well when datasets include some subsets drawn from different generative schemes. This paper proposes a novel switching NMF algorithm, which simultaneously estimates multiple NMF models supported by a fuzzy clustering concept. A fuzzy NMF reconstruction measure is modeled by introducing fuzzy memberships of each object and is also utilized as the fuzzy clustering criterion. Object fuzzy partition estimation and cluster-wise local NMF modeling are iteratively performed based on the iterative optimization principle. The characteristics of the proposed algorithm are demonstrated through numerical experiments using an artificial dataset and an environmental observation dataset.

Keywords: Fuzzy clustering, Non-negative matrix factorization, Switching analysis.

1 Introduction

Non-negative matrix factorization (NMF) [5, 6] or positive matrix factorization (PMF) [10, 9] are powerful tools for analyzing the intrinsic factors varied in non-negative matrix products, where a non-negative matrix is decomposed into two matrices having positive elements only. For example, in air pollutant measurement analysis, all observation values of some pollutants are assumed to be products of non-negative quantities of each generation, and NMF or PMF are expected to be useful for estimating the sources of the pollutants [2].

In many applications, however, some basic methods cannot work well if datasets include some subsets drawn from different generative schemes such that air pollutant characteristics can be different due to seasonal effects. Switching analysis such as switching regression [11] and local principal component analysis [12] is a useful approach for handling mixed source datasets and several fuzzy clustering-based methods have also been proposed. Fuzzy c-regression models (FCRM) [3] and linear fuzzy clustering [4] are hybrid models of fuzzy c-means (FCM) clustering [1, 8] and regression analysis or principal component analysis, where the reconstruction measures of regression model or principal component estimation are also utilized as the FCM criterion for simultaneously estimating fuzzy memberships of objects and cluster-wise least square models.

In this paper, a novel switching NMF algorithm is proposed by combining the NMF reconstruction measure with the FCM clustering concept. A fuzzy NMF reconstruction measure is modeled by introducing fuzzy memberships of each object and is also utilized as the fuzzy clustering criterion. Object fuzzy partition estimation and cluster-wise local NMF modeling are iteratively performed based on the iterative optimization principle. The characteristics of the proposed algorithm are demonstrated through numerical experiments using an artificial dataset and an environmental observation dataset.

The remaining parts of this paper are organized as follows: Section 2 briefly reviews NMF and fuzzy clustering-based switching regression and Section 3 proposes a novel approach of switching NMF supported by the FCM clustering concept. Experimental results are presented in Section 4 and a summary of conclusions is given in Section 5.

2 NMF and Fuzzy Clustering-based Switching Regression

2.1 NMF

NMF [5, 6] is designed for approximating an $n \times m$ matrix $X = \{x_{ij}\}$, whose elements are all non-negative, by decomposing into the product of two non-negative matrices of $n \times p$ matrix $W = \{w_{ik}\}$ and $p \times m$ matrix $H = \{h_{kj}\}$ as follows:

$$X \approx WH. \tag{1}$$

The value p for the number of columns of W and rows of H represents the dimension of intrinsic factors and must satisfy $n > p$ and $m > p$. In order to minimize the reconstruction errors, the objective function of NMF to be minimized is given under the least square principle as below:

$$J_{nmf} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \sum_{k=1}^{p} w_{ik}h_{kj}\right)^2. \tag{2}$$

The goal of matrix decomposition is to decompose multi-dimensional data into lower-order matrices, which represent some intrinsic factors. The lower-dimensional factors are available for making it easier for humans to understand the intrinsic features of multi-dimensional data. Then, NMF has been used in various fields such as air pollutant measurement analysis, where the observed data matrices are assumed to be the products of non-negative elements. The algorithm is based on the alternating optimization of the elements of matrices W and H, where the following updating formulas are used:

$$h_{kj} = h_{kj}\frac{(W^\top X)_{kj}}{(W^\top WH)_{kj}}, \tag{3}$$

$$w_{ik} = w_{ik}\frac{(XH^\top)_{ik}}{(WHH^\top)_{ik}}. \tag{4}$$

2.2 Switching Regression Based on FCM Clustering Concept

Fuzzy c-means (FCM) [1, 8] is one of the famous clustering methods, which is designed for partitioning data objects into a pre-determined number of fuzzy clusters. In FCM, each object is assigned to clusters with fuzzy memberships, which are more useful in handling ambiguous cluster boundaries rather than such crisp clustering models as k-means [7].

When the goal is to divide n objects with m-dimensional vectors x_i $(i = 1,\dots,n)$ into C clusters, the FCM objective function to be minimized is given as:

$$L_{fcm} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}||x_i - b_c||^2, \tag{5}$$

where u_{ci} is the fuzzy membership of object i to cluster c, and b_c is the representative centroid of cluster c. θ ($\theta > 1$) is the fuzzifier weight for adjusting the degree of fuzziness such that θ makes the objective function nonlinear with respect to u_{ci}. The larger the value of θ, the more ambiguous the cluster boundary becomes. $\theta \to 1$ reduces to k-means and the most common setting is $\theta = 2$. u_{ci} is often identified with the probability of belongingness of object i to cluster c under the probabilistic constraint:

$$u_{ci} \in [0,1], \qquad \sum_{c=1}^{C} u_{ci} = 1. \tag{6}$$

The FCM clustering concept was extended to switching regression for analyzing the dependency among pairs of outcome variable y_i and predictor x_i, where the goal is to estimate multiple regression models $y_c = f_c(x)$ in conjunction with data partitioning. Fuzzy c-regression models (FCRM) [3] achieves switching regression by replacing the FCM criterion with the regression error measure in regression analysis as follows:

$$L_{fcrm} = \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}e_{ci}^{2}, \tag{7}$$

where $e_{ci} = |y_i - f_c(x_i)|$ is the regression error measure between the outcome variable y_i of object i and its prediction $f_c(x_i)$ in local regression model c. Considering the fuzzy memberships u_{ci}, local fuzzy regression models $y_c = f_c(x)$ are estimated in each cluster.

3 Switching Non-negative Matrix Factorization Based on FCM Clustering Concept

Although NMF is a useful tool for analyzing the intrinsic structure of non-negative matrices, it cannot work well when datasets include some subsets drawn from different generative schemes. For example, real world environmental observation data can be drawn from multiple generative models due to seasonal effects and the NMF estimation can be improved by focusing only on a portion of whole observations [13]. In this paper, the previous noise robust model [13] is further enhanced into switching NMF by introducing the FCM clustering concept.

The novel switching NMF is constructed based on a hybrid approach of NMF and FCM such that the FCM criterion is replaced with the NMF reconstruction measure. Assume that we want to estimate C separate NMF models

$$X \approx W_cH_c, \qquad (c = 1,\dots,C), \tag{8}$$

by partitioning n objects into C clusters, where $W_c = \{w_{cik}\}$ and $H_c = \{h_{ckj}\}$ are cluster-wise component matrices. By introducing u_{ci} of fuzzy membership of object i to cluster c, the NMF objective function is enhanced as follows:

$$J_{fcnmfm} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ci}^{\theta}\left(x_{ij} - \sum_{k=1}^{p} w_{cik}h_{ckj}\right)^2, \quad (9)$$

where fuzzy memberships u_{ci} follow the same probabilistic constraint with FCM as Eq.(6).

The algorithm is based on the alternating optimization scheme in a similar manner to FCM, which is composed of estimation of fuzzy memberships u_{ci} and cluster-wise NMF estimation of matrix W_c and H_c.

First, the derivation process of updating formulas of NMF components is given below. For matrix H_c, the objective function is reformulated in an attribute-wise manner as:

$$\begin{aligned} J'_{fcnmfm} &= \sum_{c=1}^{C}\sum_{j=1}^{m}\left(\tilde{x}_j - W_c\tilde{h}_{cj}\right)^{\top} U_c^{\theta}\left(\tilde{x}_j - W_c\tilde{h}_{cj}\right) \\ &= \sum_{c=1}^{C}\sum_{j=1}^{m}\Big(\tilde{x}_j^{\top}U_c^{\theta}\tilde{x}_j - 2\tilde{x}_j^{\top}W_cU_c^{\theta}\tilde{h}_{cj} \\ &\qquad + \tilde{h}_{cj}^{\top}W_c^{\top}U_c^{\theta}W_c\tilde{h}_{cj}\Big), \end{aligned} \quad (10)$$

where $\tilde{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^{\top}$, $\tilde{h}_{cj} = (h_{c1j}, h_{c2j}, \ldots, h_{cpj})^{\top}$ and U_c is the orthogonal matrix, whose i-th diagonal element is u_{ci}. Here, comparing with Eq.(3), we have the modifications of $W^{\top}X \to W_c^{\top}U_c^{\theta}X$ and $W^{\top}WH \to W_c^{\top}U_c^{\theta}W_cH_c$. Then, the updating rule for elements of matrix H_c is given as:

$$h_{ckj} = h_{ckj}\frac{(W_c^{\top}U_c^{\theta}X)_{kj}}{(W_c^{\top}U_c^{\theta}W_cH_c)_{kj}}. \quad (11)$$

In a similar way, for matrix W_c, the objective function is reformulated in an object-wise manner as:

$$\begin{aligned} L''_{fcnmfm} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}\left(x_i - H_c^{\top}w_{cj}\right)^{\top}\left(x_i - H_c^{\top}w_{cj}\right) \\ &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}^{\theta}\Big(x_i^{\top}x_i - 2w_{ci}^{\top}H_cx_i \\ &\qquad + w_{ci}^{\top}H_cH_c^{\top}w_{ci}\Big), \end{aligned} \quad (12)$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^{\top}$ and $w_{ci} = (w_{ci1}, w_{ci2}, \ldots, w_{cip})^{\top}$. Here, comparing with Eq.(4), the updating formula for elements of matrix W_c is free from u_{ci} and is given as:

$$w_{cik} = w_{cik}\frac{(XH_c^{\top})_{ik}}{(W_cH_cH_c^{\top})_{ik}}. \quad (13)$$

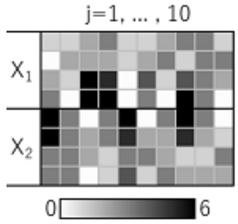


Figure 1: The artificial data matrix X, whose elements are depicted in grayscale.

Second, the updating formula of fuzzy memberships u_{ci} is given by replacing the FCM criterion with cluster-wise NMF reconstruction measure as:

$$u_{ci} = \frac{d_{ci}^{\frac{1}{1-\theta}}}{\sum_{\ell=1}^{C} d_{\ell i}^{\frac{1}{1-\theta}}}, \quad (14)$$

where

$$d_{ci} = \sum_{j=1}^{m}\left(x_{ij} - \sum_{k=1}^{p} w_{cik}h_{ckj}\right)^2. \quad (15)$$

Following the above derivations, the proposed algorithm is summarized as follows:

[Fuzzy c-Non-negative Matrix Factorization Models (FCNMFM)]

Step 1. Initialize elements of matrices W_c and H_c with random non-negative values and calculate fuzzy memberships u_{ci} by Eq.(14).

Step 2. Update elements of matrix H_c by Eq.(11).

Step 3. Update elements of matrix W_c by Eq.(13).

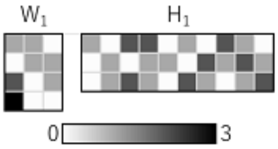
Step 4. Update fuzzy memberships u_{ci} by Eq.(14).

Step 5. If $\max_{c,i} \mid u_{ci}^{NEW} - u_{ci}^{OLD} \mid < \varepsilon$, then stop. Otherwise, return to Step 2.

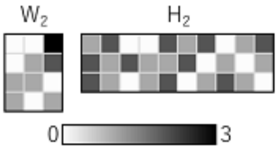
4 Experimental Results

4.1 Artificial Data Analysis

In this section, the characteristics of the proposed method is demonstrated through numerical experiments using an artificially generated dataset. The dataset shown in Fig. 1 is composed of 8×10 matrix $X = \{x_{ij}\}$ ($n = 8$, $m = 10$), whose elements were drawn from two different generative models. The first 4 objects were generated from the product of two component matrices $X_1 = \hat{W}_1\hat{H}_1$, whose elements were depicted in Fig. 2(a), while the remaining 4 objects



(a) class 1 (first half objects)



(b) class 2 (second half objects)

Figure 2: The generative component matrices in each class: 4×3 matrix W and 3×10 matrix H.

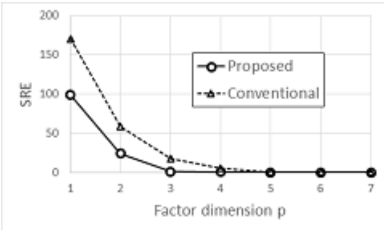


Figure 3: Comparison of squared reconstruction errors with various source dimensions.

were the product of different component matrices $X_2 = \hat{W}_2\hat{H}_2$, whose elements were depicted in Fig. 2(b). Here, both data subsets are generated from three intrinsic sources ($p = 3$) but should be analyzed with different generative models.

In general, we have no a priori knowledge on the number of sources. Therefore, in this experiment, the conventional NMF and the proposed FCNMFM were performed with all possible source dimensions as $p \in \{1, 2, \ldots, 7\}$ while the cluster number was fixed as $C = 2$ in FCNMFM. Fuzziness weight was set as $\theta = 2$ like the standard of FCM. Figure 3 compares the squared reconstruction errors:

$$SRE = \sum_{c=1}^{C} \sum_{i=1}^{n} \sum_{j=1}^{m} u_{ci} \left(x_{ij} - \sum_{k=1}^{p} w_{cik} h_{ckj} \right)^2 , \qquad (16)$$

where $C = 1$ (single cluster model) corresponds to the conventional NMF. Because the dataset is a mixture of two subsets, each of which is drawn from three-dimensional sources, the conventional NMF achieved perfect reconstruction having almost zero errors with $p = 6$, i.e., double of the source dimension, while the proposed method successfully reconstructed with $p = 3$. Then, in the followings, the results of the conventional NMF with $C = 6$ and the proposed FCNMFM with $p = 3$ are discussed.

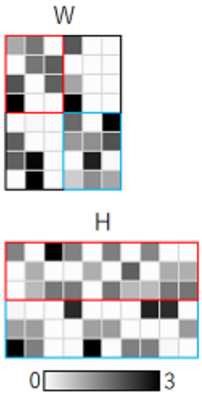


Figure 4: Result of conventional NMF with $p = 6$: 8×6 matrix W and 6×10 matrix H.

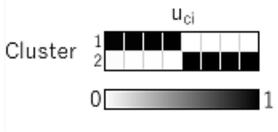
Next, the decomposed matrices are compared. Figure 4 shows W and H derived by the conventional NMF with $p = 6$ and indicates that, in matrix W, the left upper 4×3 diagonal block (enclosed by red rectangle) and the right bottom 4×3 diagonal block (enclosed by blue rectangle) reconstructed the generative matrices of $\hat{W}_1$ and $\hat{W}_2$ in some degree, respectively, such that the two separate generative models were partially reconstructed. However, non-diagonal blocks of W still have some values and then, H was polluted such that neither of the upper and lower half of H do not properly reconstruct the generative matrices of $\hat{H}_1$ and $\hat{H}_2$.

On the other hand, the proposed method estimated the two decompositions with $p = 3$ as shown in Fig. 5. The 8 objects were partitioned into two clusters as Fig. 5(a) such that the first and second 4 objects form almost completely separate subgroups. Then, the cluster-wise decomposed matrices were given as Figs. 5(b) and (c), where $U_c W_c = \{u_{ci} w_{cij}\}$ are depicted for focusing only on the cluster components. The results imply that each generative models were successfully revealed in each cluster by rejecting the influences of objects, which were assigned to another cluster.

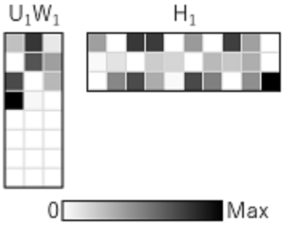
In this way, the proposed method is useful in analyzing the mixed datasets drawn from multiple sub-generative models while the conventional NMF is not necessarily suitable for handling such datasets.

4.2 Application to Environmental Observation Data

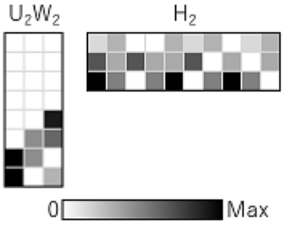
As the second experiment, the proposed method was applied to a real world data on the observations of air pollutants. The 1373×10 observed data matrix X was generated from envi-



(a) fuzzy memberships of 8 objects to 2 clusters

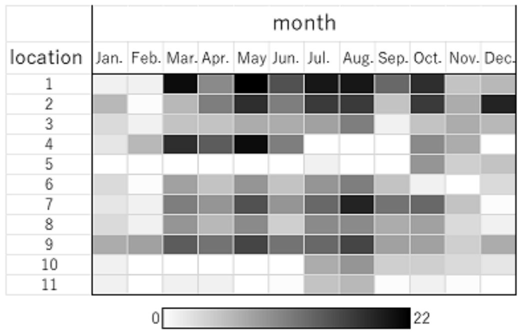


(b) decomposed matrices in cluster 1 (first half objects)

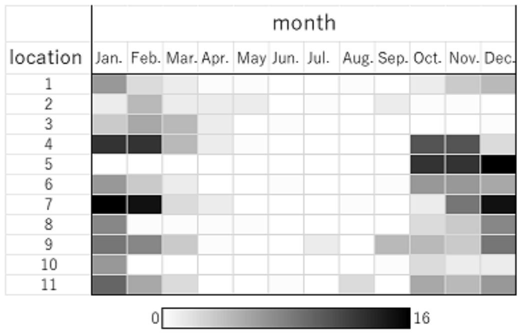


(c) decomposed matrices in cluster 2 (second half objects)

Figure 5: Result of proposed FCNMFM with $C = 2$ and $p = 6$: 2×8 fuzzy memberships u_{ci}, 8×3 matrix UW and 3×10 matrix H.



(a) cluster 1 (*summer*)



(b) cluster 2 (*winter*)

Figure 6: Comparison of cross-tabulation among 11 locations and 12 months in each cluster, which are depicted in grayscale.

ronmental observation data, which was downloaded from the web site of Osaka Prefecture in Japan (http://taiki.kankyo.pref.osaka.jp/taikikanshi/). The dataset is composed of the daily amount of pollutants observed at 11 locations in Osaka Prefecture, Japan from Jan. to Dec. in 2019, where the daily amounts of the following 10 air pollutants are stored:

SO_2, NO, NO_2, NOx, CH_4,

NMHC, OX, SPM, PM2.5, THC.

Before analysis, all incomplete instances were removed and 1373 instances composed of 11 locations $\times$ 124.8 days in average were used, where each attribute was normalized such that the maximum value is 1.

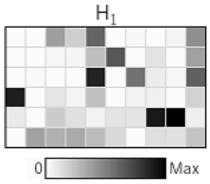
The goal of this analysis is to reveal the intrinsic generative model characteristics through the proposed FCNMFM by estimating multiple decomposition matrices. The source dimension was set as $p = 6$, which was implied in the previous study [2], and the cluster number was set as $C = 2$. Then, two generative models were estimated by partitioning 1373 instances into two clusters.

Figure 6 compares the cross-tabulation (cluster-wise frequencies of appearance) among 11 locations and 12 months in each cluster after maximum membership classification, which implies that the instances were partitioned into *summer season* (cluster 1) and *winter season* (cluster 2). That is, we have different generative models between the summer season around March to October and the winter season around October to March. However, such locations as *location 2* and *3* are not so sensitive to the seasonal differences and are mainly assigned to cluster 1 only. Additionally, Fig. 7 compares the estimated 6-dimensional sources given in the two clusters, which reveals that the generative sources of air pollutants are different under seasonal effects.

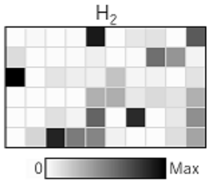
These results are quite suggestive such that environmental pollution countermeasures should be considered by switching the generative models under seasonal effects rather than the conventional single-model analysis.

5 Conclusions

This paper proposed a novel switching NMF algorithm called FCNMFM, which is a hybrid of NMF and FCM



(a) cluster 1 (*summer*)



(b) cluster 2 (*winter*)

Figure 7: Comparison of 6-dimensional air pollutant sources in each cluster: 6×10 matrix H.

clustering. The new objective function was defined by introducing FCM-like fuzzy memberships into the NMF measure, and the updating formulas were derived under the iterative optimization principle.

The characteristics of the proposed method were demonstrated through numerical experiments using an artificially generated dataset and a real world data on the observations of air pollutants. Although the conventional NMF cannot work well in handling the mixed datasets drawn from multiple generative models, the proposed method successfully analyzed cluster-wise models.

Possible future work includes development of the mechanism for selecting the optimal cluster numbers, which is heuristically selected in the present model. In FCM-type fuzzy clustering, several cluster validity measures are utilized in cluster number selection [14]. How to improve such validity measures for switching NMF analysis can be a challenging issue from the theoretical aspect.

Acknowledgement

This work was supported in part by Artificial Intelligence Research Promotion Foundation through 2019 research grant and JSPS KAKENHI Grant Number JP18K11474.

References

[1] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[2] K. Funasaka, M. Sakai, K. Katahira, D. Asakawa, Y. Furuichi, Y. Oku, Source apportionment of pm2.5 at osaka city by positive matrix factorization (pmf) method, Annual report of Osaka City Institute of Public Health and Environmental Sciences 77 (2015) 65–70.

[3] R. J. Hathaway, J. C. Bezdek, Switching regression models and fuzzy clustering, IEEE Trans. Fuzzy Systems 1 (3) (1993) 195–204.

[4] K. Honda, H. Ichihashi, Regularized linear fuzzy clustering and probabilistic pca mixture models, IEEE Trans. Fuzzy Systems 13 (4) (2005) 508–516.

[5] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.

[6] D. D. Lee, H. S. Seung, Algorithms for nonnegative matrix factorization, Advances in Neural Information Processing Systems 13 (2000) 556–562.

[7] J. B. MacQueen, Some methods of classification and analysis of multivariate observations, in: Proc. of 5th Berkeley Symposium on Math. Stat. and Prob., 1967, pp. 281–297.

[8] S. Miyamoto, H. Ichihashi, K. Honda, Algorithms for Fuzzy Clustering: Methods in c-Means Clustering with Applications, Springer-Verlag, Berlin Heidelberg, 2008.

[9] P. Paatero, Least squares formulation of robust non-negative factor analysis, Chemometrics and Intelligent Laboratory Systems 37 (1997) 23–35.

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

[11] R. E. Quandt, A new approach to estimating switching regressions, Journal of the American Statistical Association 67 (1972) 306–310.

[12] M. E. Tipping, C. M. Bishop, Mixtures of probabilistic principal component analysers, Neural Computation 11 (2) (1999) 443–482.

[13] M. Ueno, K. Honda, S. Ubukata, A. Notsu, Robust non-negative matrix factorization based on noise fuzzy clustering mechanism, in: Proc. of 2019 2nd Artificial Intelligence and Cloud Computing Conference and 2019 Asia Digital Image Processing Conference, 2019, pp. 1–5.

[14] W. Wang, Y. Zhang, On fuzzy cluster validity indices, Fuzzy Sets and Systems 158 (19) (2007) 2095–2117.