Development of Novel Deep Recurrent Sparse NMF for Source Separation

Siva Priyanka S, Vamshi Krishna Krishnamaraju

Department of ECE, Kakatiya Institute of Technology and Science, Warangal, India

Corresponding Author Email: ssp.ece@kitsw.ac.in and m20sp001@kitsw.ac.in

Abstract:
Non–negative matrix factorisation (NMF) which is a technique for reducing the dimensionality of the data matrix considered, into two lower rank matrices. It is a rather popular algorithm for matrix decomposition and it implicitly imposes non-negative constraint on its data sets. This constrains proves to be enhancing the interpretability by obtaining parts based representation. The paper explores NMF definition and its variations, modifications and extensions pursued over the years. A deep recurrent variation of the sparse NMF for source separation is also discussed extensively. The problems of optimisation in sparse NMF are tackled by sequentially iterating a thresholding algorithm. This exhibits more interpretability and a better convergence rate when compared with basic sparse NMF. When small amount of data is available, the deep variation of the NMF gives better performance compared to sparse NMF. Various furthering of the NMF algorithm avenue is explored and addressed.

Keywords:
NMF, SOURCE SEPARATION, SPARSE NMF, DEEP RECURRENT ARCHITECTURE, TYPES OF NMF

1. INTRODUCTION

Of many things, one of the fundamental concepts is to believe that there is something meek and compact is performing the basic parts beneath the obvious complexity of engineering and science, which is deeply embedded into these fields. In the fields of signal processing, data mining, including machine learning, and pattern recognition, this is also true. As there is an exponential increase of raw data quantity, which is a direct consequence of advancements in sensor and computer technology, there is a dire need of obtaining an effective representation with the use of a relevant dimensionality reduction technique which is challenging in the field of multivariate data analysis. The general idea is to satisfy two fundamental properties- dimensionality reduction of original data and effective identification of principal components, prominent features, or hidden concepts in the data, depending on the context of the application has to be done.

In most of the cases, the data matrices are the observations which can be described by linear/multilinear combination models. Hence, formulating the reduction of dimensionality can be considered as decomposition of the original data matrix into a couple of sub-matrices termed as factor matrices observed from an algebraic standpoint. Such low-rank approximations are exemplified by the likes of canonical methods such as LDA (Linear discriminant analysis), PCA (Principal Component Analysis), VQ (Vector Quantization), and ICA (Independent Component Analysis). The methods mentioned vary in statistical features because of the various constraints on the component sub matrices along with their core assemblies. But the thing is that there is no sign constraint i.e., the components can have negative sign or a subtractive combination. So, by imparting the non-negativity constraint which increases the interpretability of the concerned issue, Non-negative Matrix Factorization (NMF) is proposed by Paatero and Tapper in [1] along with Lee and Seung [2].

In chemometrics, in which the vectors are continues rather than discrete, the concept of “self-modelling curve resolution” is an implementation of NMF [3]. In[1], the implementation is more towards a Positive Matrix Factorisation but is flawed when theoretical analysis like algorithm convergence, as well as the applicability of the algorithms into different applications which will be constrained when these are taken into consideration. But, Lee and Seung made NMF popular by overcoming the mentioned limitations by their efficient approach by emphasising the utilisation of parts based representation potentially. One of the advantages is that the underlying fundamental concept behind the NMF is its closeness to perception mechanism because NMF will establish a feasible model for knowledge of object parts, hence the success of NMF in real world tasks and scenarios.

As the non-negative constraint imposed obviously lead to some sparsity [2], it is shown that the NMF is a more effective representation than both the entirely distributed and the exclusively active component description [4]. This lead to the successful implementation of NMF in the fields of data mining [5], machine learning [6], signal processing [7], facial expression recognition [8], gene expression classification [9], blind source separation [10], and in many others too. Due to its usage and elegance, NMF has been since a fruitful field for many researchers. They are striving to improve this concept from different standpoints.

So, one can divide NMF concept into four subcategories. Although the basic classification is preliminarily done in [11], the categorisation is extensively described with the help of Figure.1.

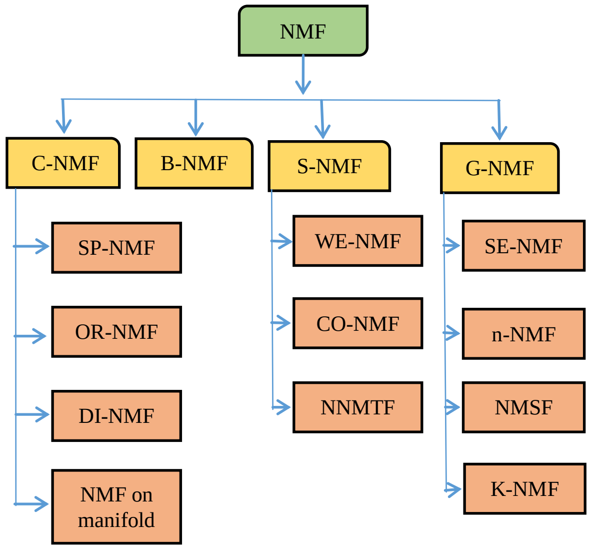


Figure1. Different types of NMF- Algorithms and models

2. BASIC NMF(B-NMF) ALGORITHM

The B-NMF is the basic version of NMF algorithm. Typically, Non-Negative Matrix Factorization (NMF) is a cutting-edge feature extraction approach that falls under the category of unsupervised learning. When there are a lot of attributes and they're vague or unpredictable, NMF comes in handy. NMF can create significant patterns, subjects, or themes by integrating attributes. NMF creates each feature by combining the original attribute set in a linear fashion. Each feature has a set of coefficients that represent the relative importance of each attribute on the feature. Each numerical attribute and each individual value of each categorical attribute have their own coefficient. All of the coefficients are non-negative.

NMF employs multivariate analysis and linear algebra techniques. It uses a nonnegative constraint to replicate a range of naturally occurring signals seen in many situations, such as pixel intensities, occurrence counts, and amplitude spectra. For the task of single-channel source separation, NMF is majorly used. A non-negative matrix **X** being the aaudio signals' fourier spectrogram where $\mathbf{X} \in \mathrm{R}^{P\times Q}{}_{+}$ with *Q* frames and *P* frequency bins in each frame can be used to perform NMF for audio source separation. So, the algorithm decomposes the data into matrix X which would be the product of two lower ranking sub-matrices: B (**p** by **x**), the NMF basis matrix and C(**x** by **q**), consisting the related coefficients or weights. The rank (**x**) of the matrix is chosen appropriately of order (**p** by **q**). To put it another way, this way denotes a stochastic pattern which is high-dimensional with a small number of bases, hence the ideal estimate can only be realised if the intrinsic features in B are discovered. In basic NMF, a simple multiplicative updating rule is employed to quantify the difference between the product of the factor matrices B and C, and the input data matrix X.

By the usage of basis vectors in $\mathbf{B} = [\mathbf{b}_1,...,\mathbf{b}_X]$ and weight vectors in W can alternatively be viewed as a basis-based approximation for reconstruction of **X** mixed signals.

The signals which are segregated can be represented as a collection of basis vectors in a matrix $\mathbf{C} = \mathbf{K}^{\mathrm{T}} = [\mathbf{k}_1,..., \mathbf{k}_X] \in \mathrm{R}_+{}^{X\times Q}$, with the accompanying weight parameters. NMF can be described as a bilinear model where summation of these bilinear combinations mentioned of rank-one K nonnegative matrices, but the outer product of two vectors $\mathbf{b}_x$ and $\mathbf{k}_x$is every single matrix which can be represented as $\mathbf{X} \approx \mathbf{BC} = \mathbf{BK}^{\mathrm{T}} = \mathbf{X} \approx \mathbf{BC} = \mathbf{B}\boldsymbol{K}^{T} = \sum_x \boldsymbol{b}_x \circ \boldsymbol{k}_x$.

For a matrix factorization approach, the three most important problems that needs to be addressed are (i) During which of the assumptions NMF can recover the correct answer i.e., effectiveness (ii) During which of the assumptions, NMF will be unique i.e., uniqueness and mainly (iii) whether the NMF solutions exist that are nontrivial. The first two problems were addressed in [12] from a geometric point of view. The solution existence was proved for the first time in [13] with the help of Completely Positive Factorization (CPF). While Principal Component Analysis (PCA) can be through in polynomial time, the optimization issue of NMF is more challenging than its unconstrained version in terms of establishing the rank which is non-negative and calculating the related factorization. The overall NMF complexity for factorization rank which is fixed would be, however, unidentified in general [14]. Hence, representation by parts and sparseness of the NMF algorithm comes with more complexity

Furthermore, the spectrum of SVD (Singular Value Decomposition) or PCA is always less than that of NMF [13]. There will be trade-off between effectiveness and complexity.

3. CONSTRAINT NMF (C-NMF)

It can be seen in the B-NMF, there is no possibility of a unique solution by imposing only a non-negativity constraint. As a result, in order to correct the ill-posedness, terms of regularisation incorporating past information which are secondary constraints on U and/or V and more completely describe the features of presented issues. C-NMF can be divided into SP-NMF, OR-NMF, DI-NMF, and NMF on manifold.

3.1 Sparse NMF (SP-NMF)

The uniqueness of the matrix decomposition by imposing a local-based representation can be done by the sparseness constraint. Considering all the problems of C-NMF, SP-NMF is the most commonly and deeply mined, and it has virtually become a necessity in practise. The question here is whether B or C should be chosen as the candidate on whom the sparseness restriction is placed [15]. This answer entirely depends on the application it is being used. In every observation, just a small part is affected by the basis, if the B's vectors being B's columns are sparse whereas, if C's columns are imposed by sparseness, a rectilinear mixture of basis vectors in small amounts is used to approximate each observation. A restricted amount of data which has to be trained will deduce every single basis vector implementation, where it will be closely connected to clustering. The cost function can be written as

$$min_{B,C}\frac{1}{2}\{\|X-BC\|^2+\alpha\|B\|^2+\beta\|C\|^2\}. \qquad (1)$$

By the usage of Basic NMF techniques, the result to the factor matrix that is not sparse is nonetheless effective generally. Under the sparseness constraint, the sparse candidate matrix's solution is meant to be adjusted. As proposed in [16], NSC (Non-negative Sparse Coding) has the objective function by combining the l_1 norm of **C,** i.e., the sparseness penalty and the SED (Square of Euclidian Distance). By sparseness and non-negative constraints, using dynamic set based technique least squares minimization has to be solved which in turn got named to be the modified least angle regression and selection (LARS) [17], is one of the highlights of NSC. In order to determine both the active set and the non-active set, the improved algorithm merely presents the non-negativity criteria by deleting the features from the active set which are updated to zero, since LARS is dependent on the active set algorithm to solve the stated LASSO (Least Absolute Shrinkage and Selection Operator) concept. Hence, the SP-NMF problems can be dissolved into N separable LASSO sub problems.

Gao and Church [18] choose the l_2 norm of **C** as the penalty component of sparseness in the SED objective function which is regular. But, the quadratic penalty in this case could result in truncated values than providing sparse values. In comparison to the sparse formula based on the l_2 norm, the l_1 norm formula would be more useful in managing sparsity [19]. Furthermore, high value suppresses the l_2 norm of C, resulting in a higher value of the l_2 norm of B as there is a scaling impact of the sparse regulatory factor. So, there is a need for normalizing the columns of B during iterations. In [9], sparsing the addition of l_1 norm squares of vectors in cadindate matrix, non-sparsing the Frobenius norm of another matrix by linear combination of SED is a complex objective function. As is customary, the second term would be the sparseness penalty term. The third term suppresses the related matrix, lowering the values of its elements and mitigating the scrambling impact discussed before, being a notable distinction from former Sparse NMF algorithms. For controlling the amount of sparsity, whole another technique has to be integrated. A variation of SP-NMF i.e., graph regulasied SP-NMF is presented by [20] where the sparsity constraint is integrated along with graph regularization constraint for source separation. It is demonstrated by figure 2.

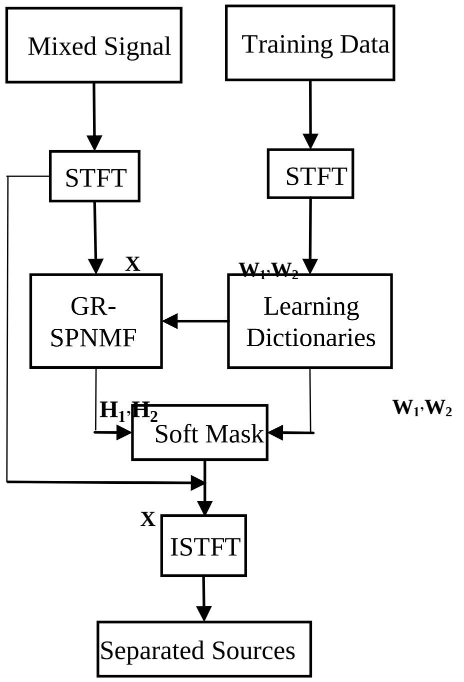


Figure2. Source separation by graph regularisation SP-NMF

3.2 Orthogonal NMF (OR-NMF)

The OR-NMF imposes the constraint of orthogonality on either the feature B or C. Li et al. is the first to use the orthogonality principle in NMF [21] for redundancy minimization of the bases, thereafter Ding et al. had explicitly used the concept of OR-NMF[22]. The orthogonality in NMF will definitely create the sparseness constraint without overtly imposing it. Hence, OR-NMF would be considered a distinct case of SP-NMF where the optimization models vastly differs between them. OR-NMF will have a solution region with unique sparsity learning the most diverse parts. Hence, OR-NMF is pursued separately from SP-NMF. The constraints would be $B^TB = I$ resulting in more diverse parts or $C^TC = I$ resulting in more clustering accuracy. If both the cases are imposed, it is termed bi-orthogonality which has poor performance in terms of approximation. OR-NMF is similar to clustering of data matrix either of rows or columns where one matrix is mapped to cluster centres and other to cluster indicator vectors. OR-NMF is taken for clustering problems because of its interpretation to clustering mentioned above.

There are two common orthogonality penalty terms included in the SED or GKLD (Generalized Kullback-Leibler Divergence) objective functions. The first order quantity is the trace of the resultant matrix is the difference between the feature and the identity matrices [22]. The second order quantity [23] would be the SP-NMF's optimization model which has l_2 norm of the resultant matrix is the difference between the feature and the identity matrices. The first one would be the orthogonally constrained optimization problem which can be simplified using Lagrange multiplier method, as if there is any constraint imposed on it resulting in the increase of computational load. So, the multiplicative update rule corresponding to regular algorithms is acquired. In the second formulation, complexity is reduced by a single complexity which controls orthogonality. The modified multiplicative rule for updating can be used in the result similarly [23]. Under the orthogonality constraint, an innate portrayal of the following updated multiplicative rules for updating is to trade specific terms in the actual multiplicative rules for updating with a fresh one, suggesting the property of orthogonality. The multiplicative update will substitute in [24] $\mathbf{X}^T\mathbf{B}$ for $\mathbf{C}^T$ in the denominator as there is inclination of coefficients alongside of orthogonal projection.

3.3 Discriminant NMF (DI-NMF)

As mentioned above, the NMF comes under unsupervised learning of machine learning algorithms with respect to pattern recognition. Extension of the B-NMF under supervised learning with the combination of discriminant information and the decomposition too, resulting in DI-NMF or F-NMF (Fisher-NMF) which merges classification and generative model into a joint framework. This DI-NMF has applications in the fields relating to classifications tasks like facial recognition. The result of the difference between the class scatter matrices of within and between would be in the Generalized Kullback-Leibler Divergence (GKLD) as the term of penalty, for the construction of the objective function; this is the Fisher discriminant constraint which is first suggested by Wang et al. [25] which gained acceptance as the basic structure for DI-NMF. The meaning of between class and within-class in [25],[26] are exclusively based of the coefficient matrix C, with no reference to X or B. As there is no guarantee for the DI-NMF convergence, an algorithm of projected gradient DI-NMF) which integrates the projected gradient in B-NMF is proposed in [8]. Lin's proposed projected gradient method [27] is used to ensure that the maximum point would be the stationary point. This proposal is the combination of LDA classifier with NMF. By incorporating SVM (Support Vector Machine) which is a maximum boundary classifier in the NMF model [28] which are proved to be effective in evaluating the classification tasks. For linearly inseparable situations, it will also profit from SVM's kernels with nonlinear nature.

3.4 NMF on manifold

Real-life numbers is frequently appraised from a low-dimensional sub manifold with nonlinear nature, which is a high-dimensional topography resembling Euclidian space locally. If the fundamental geometric structure is identified and stored, there is a significant improvement of learning performance in many cases. Many variations of manifold based learning algorithms have been proposed like Laplacian Eigenmaps, Isomap and so on. These algorithms are differed based on the sole point being their consideration of relationship between the local points which is a topological property. Local invariant properties along with the equivalent learning methods of manifold are integrated with NMF in NMF on manifold which has significant improvement in performance observed in image and document clustering [29] and where the Euclidian space is considered when B-NMF is taken. Basically this type of NMF has an objective function where it will be used as a supplementary regularisation term in this case and combines the geometrical information.

Cai et al. [30], [29] introduced a graph regularised NMF (GRNMF) that characterised the structure of manifold by generating an adjoining neighbour graph where the data points are scattered. Here, the assumption of invariance which would be in turn local, is applied which states that if the points are neighbours in high dimensional topography, then in the low dimensional space, they are supposed to be close enough. The Square of Euclidian Distance objective function's term of penalty is the weighted squares present in the distances in terms of Euclidian form of the corresponding data points, also the revised multiplicative rules for updating are used to access and determine the local variance assumption. It is the NMF integration to the Laplacian

Eigenmaps. The assumption is pursued in a different direction Zhang et al. [31]. As the norm of mapping **C** from low to original high dimensional space will give the degree of separation between the **C**'s mapping of nearest points, formulation of a gradient distance minimization problem is done keeping in the mind to find the map which preserves local topology the best. This NMF which preserves topology has an alternating gradient descent algorithm invented to tag along with it. Furthermore, as compared to the SED function, this model for optimisation is equivalent to minimising the whole norm of variation square amongst X and BC imposing the non-negativity constraint, which maintains improved scale basis.

One more significant topological feature exhibited which is the assumption of LLE (Locally Linear Embedding) stating that a data point spawned on a specific manifold in the actual space as a linear combination of several neighbouring points should be reconstructed from its neighbours in the compact low dimensional subspace in an analogous way or with the same coefficients of reconstruction. Deriving the modified multiplicative update rules consequently and using the above property, Gu and Zhou et al. proposed the neighbourhood preserving NMF for combining LLE [32]. Expansion of this work is done by Shen and Si [33] by expanding the relationship of locally linear in nature from an only manifold to many manifolds, where approximation of a data point is done using nearby samples' linear combination solely on the equivalent manifold. Although, the above two algorithms differ in the aspect of determining the neighbourhood sample points. When NMF usage is approaching, the imposing of sparseness on NMF is deemed necessary practically. Various examples like Discriminant Sparse NMF [34], Manifold-concerning Discriminant NMF [35], and Manifold Normalized Discriminative NMF [36] show the incorporation of diverse constraints for the improvement of decomposition quality and reflect the problems' multilateral properties.

4. STRUCTURED NMF ALGORITHMS

In the NMF problem, the structured NMF will enforce external structures or characteristics into the solution by directly altering the formulation of regular factorisation alternative to introducing the additional constraints substituting the penalty terms as opposed to C-NMF. It is written as **X** ≈ F (**UV**). It can be classified into WE-NMF, CO-NMF, and NMTF.

4.1 Weighted NMF (WE-NMF)

Generally, algorithms involving weighted inventions are nothing but learning algorithms which are slightly revised versions utilizing the emphasis of various components' relative importance. The WE-NMF can be stated with weight matrix **V** as

$$\mathbf{V} \otimes X \approx \mathbf{V} \otimes (UV) \qquad (2)$$

The algorithm can be viewed as WLRA (Weighed Low-Rank Approximation) where it searches for a solution matrix of lower rank which will be close and similar to that of the input matrix, but in line with predefined weights. The decomposition or low-rank matrix completion involving noise is used for construction of recommendation systems which comes under collaborative filtering; it is expected to forecast the missing entries when there is an incomplete data matrix (original) with some missing entries or unseen. Here we can give binary weights to the original matrix so that the binary one for the observed elements and zero to the unknown elements, and creating the weight matrix **V** to solve the problem listed. Although Mao and Saul [37] introduced the WE-NMF in the multiplicative update rules, it resulted in slow convergence. So Zhang et al. [38] proposed another way of EM algorithm approach where the model estimates substitute the missing entries in the E-phase and M- step has the un-weighted multiplicative rules which follows the filled-in matrix. But the EM variation of the WE-NMF does not have above par convergence and increased computational complexity as the original matrix is very dense. So for convergence enhancement, a better way is to use a generalised EM model that combines step of E and a partial step of M. The two parts involving are that, at the M-step, ANLS (Alternating Non-negative Least Squares) is converted to optimise the difficulties initially. Then M-step is partially chosen as there will be termination of iterations due to significant improvement as opposed to identification of optimal solution, hence this M-step solution entirely is not a desirable thing. Modifying the EM WE-NMF, by improving accuracy and convergence rate is done. Furthermore, while the above-mentioned Weighted NMF methods are centred on the singular issue of dealing with inadequate data matrices, they are pertinent to all Weighted NMF models. In [39], non-intrusive load monitoring problem is optimised for various feature utilization using multivariate signals which is a variation of weighted NMF. This could be observed from figure.3.

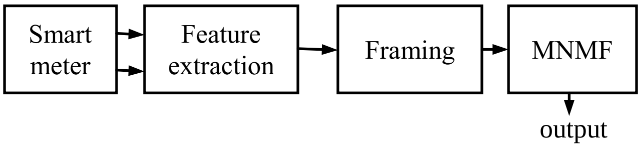


Figure 3. Weighted NMF with multivariate data

4.2 Convolutive NMF (CO-NMF)

CO-NMF is mainly used in the field of source separation because it is decomposed instantaneously with basis matrix B describing every object in the spectrum; coefficient matrix C describes the activation in time. As there is a likely dependence in between the column vectors of original X matrix which states formation in the time domain, there should be accounting of time dependant characteristics of the considered scale. Typically, a convolutive generative model is used to characterise the temporal relationship between several observations spanning close time intervals. As a result, from a computational standpoint, Convolutive NMF can be broken down into a series of Basic NMF issues.

CO-NMF is an example of an overfitted demonstration to some extent. This type of CO-NMF is implemented by Smaragdis [40], where its solutions are mapped to basis matrix B_t and coefficient matrix C. Compared to GKLD, the SED objective function improves by decreasing the computational load with better separation. Generating the CO-NMF is not enough for convergence to occur. The sparseness constraint is integrated into CO-NMF regarding to separating attributes to sparsity by O'Grady and Pearlmutter [41]. Owing to the lack of a suitable reorganizing step size similar to the one in the conventional approach, C is updated using standard multiplicative principles, whereas B_t is updated using the typical additive gradient descent method without a guarantee of algorithm convergence.

It can be interpreted that, the B-NMF would be under the frequency domain wing as opposed to the CO-NMF under the time-frequency domain analysis. As there is a lack of a suitable reorganising step size similar to the one in the conventional approach, C is updated using standard multiplicative principles, whereas B_t is restructured using the typical additive gradient descent method. But the usage of CO-NMF is limited to the field of audio analysis and the multiplicative rules taken up by the CO-NMF model result in the non-convergence of the algorithm.

Nevertheless, the extent of this adaptation form's use is confined to audio data processing. Furthermore, the much more Convolutive NMF model's simple adjustment of the multiplicative updating rules may not converge.

4.3 Non negative Matrix Tri-Factorisation (NMTF)

NMTF (Non negative matrix tri-factorisation) extends B-NMF [42] by forming three product matrices i.e., $X \approx B\,TC$. As this tri-factor NMF without any constrints providing an extra degree of freedom with additional features and can be merged into a two factor NMF. As the bi-orthogonality constraint cannot produce desired results in OR-NMF, extra factor of T is substituted additionally for absorption of various levels of B and C, so that the lower level of matrix illustration would be precise while the orthogonality imposition is in line with it. As a result, the X's rows and columns can be grouped at the same time, which is important in script processing and clustering. If there is amalgamation of a very smooth factor of T which makes B and C sparse thereby solving the disagreement between the sparseness and approximation, stated in [43] as Non-smooth NMF, which would be similar to NMTF. Other variations like LPBNMF(Linear Projection-Based Non-negative Matrix Factorization) [44] and Convex NMF [45] would also be taken into consideration as extensions of NMTF, but the NMTF consideration as a distinct circumstance of Multi-layer NMF [46] would be more accurate.

5. GENERALIZED NMF (G-NMF)

Broad sense extension of B-NMF is the G-NMF where it has advanced the model of decomposition in a deeper manner identical to S-NMF as opposed to the introduction of some penalty constrains in C-NMF. It partially violates the vital constraint of non-negativity, along with variations in the data types, as well as the modifications in the factorization pattern, and many more. The field of G-NMF looks promising with its extensive and emerging research compared to B-NMF, CO-NMF and S-NMF. The G-NMF can be categorised into Semi-NMF, Non-negative Tensor Factorisation, Non-negative Matrix Set-Factorisation and Kernel NMF.

5.1 Semi NMF (SE-NMF)

Main constraint imposed by NMF is the non-negative one, but when the constraint is removed, the original data matrix has mixed signs. SE-NMF is first proposed by Ding et al. [45] where the constraint of non-negativity is imposed just on C, where the restrictions are removed on B suggesting some kernel concept implementation of B-NMF.

As the candidate data received won't always be non-negative, the principal components or latent features could have negative elements regarding the phase information. But this could be interpreted same as B-NMF method, but it still involves the non-subtractive combinations.

The corresponding status of B and C in B-NMF can be compromised here. So, Ding et al. solved the optimisation problem by the separation of positive and negative parts where the parts are combined in the mixed matrix. The convergence is proved using the updating of C by multiplicative rules where B is fixed, and then the inverse is done to obtain analytical local optimal solution. SE-NMF is also equally applicable for the earlier Convex NMF [45].

5.2 n-dimensional Non-negative Tensor Factorisation (n-NTF)

Normally, the multi-way data pre-processing is done by organising the same data into a matrix form, but it can lose the structure of multi way data. Here, naturally generalising the matrix factorisation is the tensor factorisation. The NMF comes under a special instance of n-dimensional Non-negative Tensor Factorisation (n-NTF) with n = 2. Immediately after the first proposal of NMF the concept of PTF (Positive Tensor Factorisation) has been put forward by Welling and Weber [47]. This NTF has gained more acceptances among all the G-NMF methods. The NTF differs from NMF in various ways like the type of data is vectors here. But if there is forced vectorization on some data like images, it could result in problems regarding structural and spatial information. The uniqueness issue of NMF is addressed by NTF as it would be distinctive merely in certain particular weak environments where the uniqueness is directly proposed to tensor order. NTD (Nonnegative Tucker Decomposition) [48] and more restricted NTD [49] are the two types of NTF models with variation in core factor tensor. Some other variations also exist in line with the B-NMF conclusions. The main thing is that, just by generation of matrix to tensor forms, convergence is not promised. These NTF concepts can be integrated with NTF, like SP-NTF, manifold NTF and more [50], [51].

5.3 Non-negative Matrix-Set Factorization (NMSF)

If vectorization of the original matrix type namely image, audio, lead to poor genrality, or below par approximation and high computational load the learning problem will be a notorious sample problem, this is pursued by Li and Zhang where they proposed Non-negative Matrix-Set Factorization (NMSF) [52]. It is accomplished by processing the candidates which are set of sample matrices directly on the matrix set. The sample matrices are decomposed into K factor matrices' product with factor matrices of K-1 in number representing the acquired knowledge of the features that generalise the matrix of features in non-negative matrix factorisation algorithm to a matrix set of features, and left out factor matrix will differ from model matrix describing the patterns of activation that generalise the NMF's coefficient vector to a matrix of coefficients.

For solving the optimisation problem occurred earlier, Li and Zhang have proposed another variation namely Bilinear Form-based NMSF algorithm (BFNMSF) [53]. Till date this avenue is not comprehended extensively but the NTF and NMSF has provided a basic framework to work on based on the previously stated basic NMF algorithm variations where it can be extended as needed. This NMSF will be focusing solely on three dimensional cases which is better than the NTF considered in the same 3D cases involving much broader ones.

5.4 Kernel NMF (KNMF)

All the variations discussed till now are based and work on linear models that cannot take out nay non-linear information or relations which is hidden away in the data thereby limiting the extent where the NMF can be applied. So, by charting the data to a contained feature space with the usage of non-linear functions similar to the kernel versions of ICA, LDA and PCA, Kernel based NMF can be achieved. It gives rise to many potential appications in fields involving the negative value data processing using specific Kernel methods allowing basis vectors which depend on higher order.

The NMF model will only depend on the Kernel matrix. In polynomial feature space, Buciu et al. suggested the aforementioned model of kernel NMF and used transformed multiplicative rules as the method of updating is merely valid for kernels of polynomial degree [54]. When PG method is involved, this can be generalized for any form of kernel method. There are many variations of NMF which uses Kernel function. Nevertheless, because the current kernel NMF outcomes are earliest and objective function dependent, a systematic kernel NMF creation and assessment methodology will be required in the future.

6. A PROPOSED DEEP RECURRENT BASED SPARSE NMF

As NMF can be integrated into the neural networks as discussed above, many models involving the variations of neural network NMF came into existence. A deep recurrent based NMF or DR-SNMF is a variation of SP-NMF which embeds the deep recurrent technology which is used predominantly in source separation. Here the source considered for separation is speech. The work done can be conveniently described in the form of a flow chart as shown in figure.2. This consists of four stages namely - Statistical Model, Inference Algorithm, Unfolded deep network and finally Speech separation. The statistical model is nothing but the sparse NMF, the inference algorithm is the sequential iterative soft-thresholding algorithm (ISTA), and the unfolded deep network is the DRSNMF recurrent network [55].

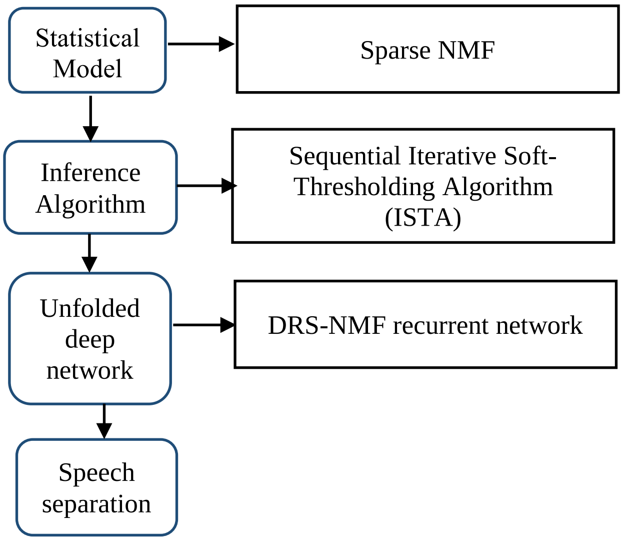


Figure 4. Flowchart of the proposed DRS-NMF algorithm

Initially, statistical modelling is done where sparse NMF is applied. But to avoid problems in the dictionary training of B, an inference algorithm is used. By iterating the sequences continually and by maintaining the connections between the weights and the feed forward deep network, inference algorithm is applied. Then unfolding of the deep network is done as this step is required to achieve the deep recurrent level and the desired network is formed [55]. This DRS-NMF can be assumed as shown in the figure-5.

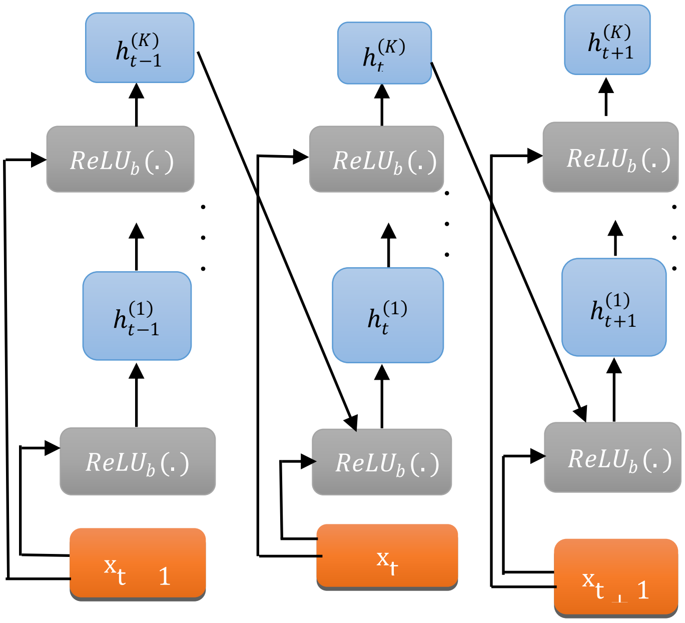


Figure 5. Structure of proposed DRS-NMF algorithm

7. RESULTS

Data set of the CHiME2 corpus [56] where the binaural room impulse responses (RIRs) and wall street journal utterances are convolved and mixed with natural non-stationary noise is considered for comparison. The six SNRs spaced by 3db of these utterances vary in between -6dB to 9dB. The noise as well as RIRs considered is collected from various sources like children, music, television radio and appliances. This data set has 7134 utterances in total, where the test set has 1970 utterances. The validation loss from BSS Eval Matlab toolbox [57, 58] is considered for the measurement of separation performance between sparse NMF and DRSNMF. The depth of sparse NMF stacks and number of hidden nodes are denoted by K and N respectively.

After applying the algorithm, sparse NMF and DRSNMF are compared with respect to learning curves with number of iterations versus validation loss and training loss. Here, when only 10% of data set i.e., right side of the data set is considered there will be no ovefitting, whereas the total training set is considered, there will be overfitting. But, when 100% of data set is cosidered, the results may not be effective. The training loss in the dotted lines in the form of dotted lines decreases as validation loss in solid line increases. So, it can be said that DRNMF gets best results when only a part of data is available to train compared to normal SNMF. This can be followed from the figures mentioned below. K refers to the total number of layers, N refers to the number of NMF basis vectors.

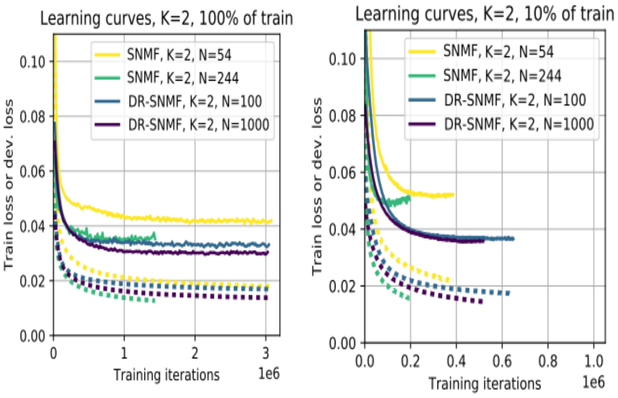


Figure 6. Number of Iterations versus validation loss and training loss for complete data set and 10% of data set for K=2.

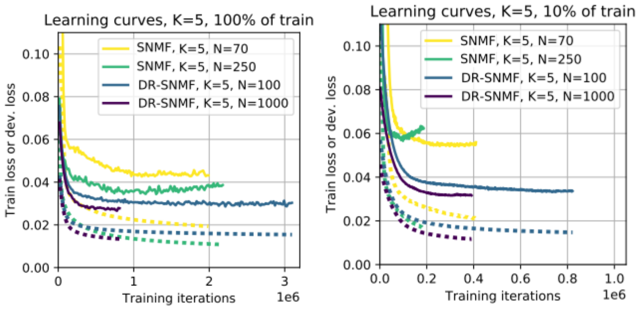
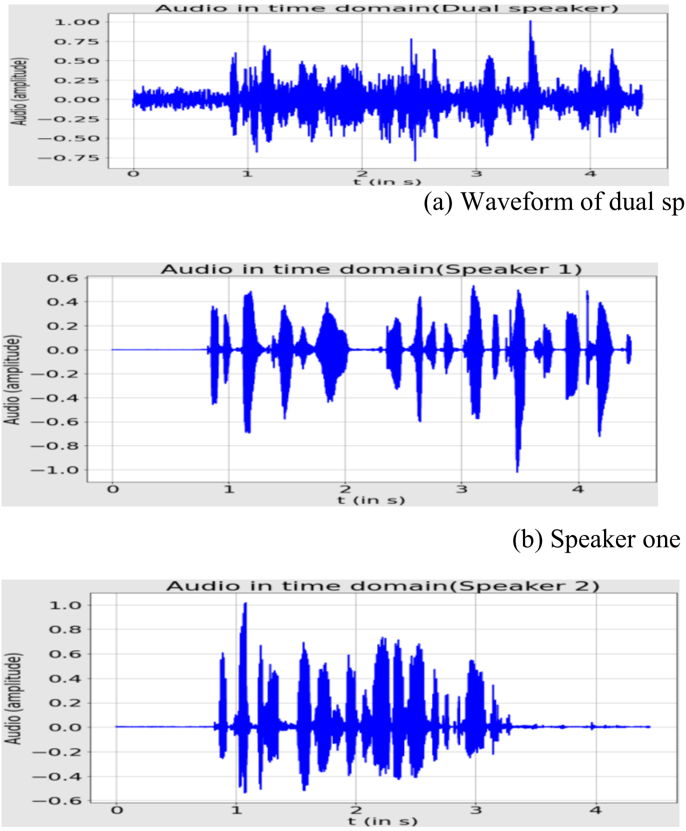


Figure 7. Number of Iterations versus validation loss and training loss for complete data set and 10% of data set for K=5.

A chunk of speech signal is taken from the data set and is considered for source separation. A speech signal consisting of two speakers is taken as mixture signal and given as input to the network. The mixed signal consists of a female and male speakers speaking simultaneously. The female speaker‟s words are “For the first time in the year, the republicans also captured both houses”. The male speaker‟s words are “The Company previously traded over the counter”. The output is separated female and male speakers‟ words and can be viewed in time domain as well as frequency domain. The time domain representations are shown by figure 8 and magnitude spectrograms are shown by figure 9.

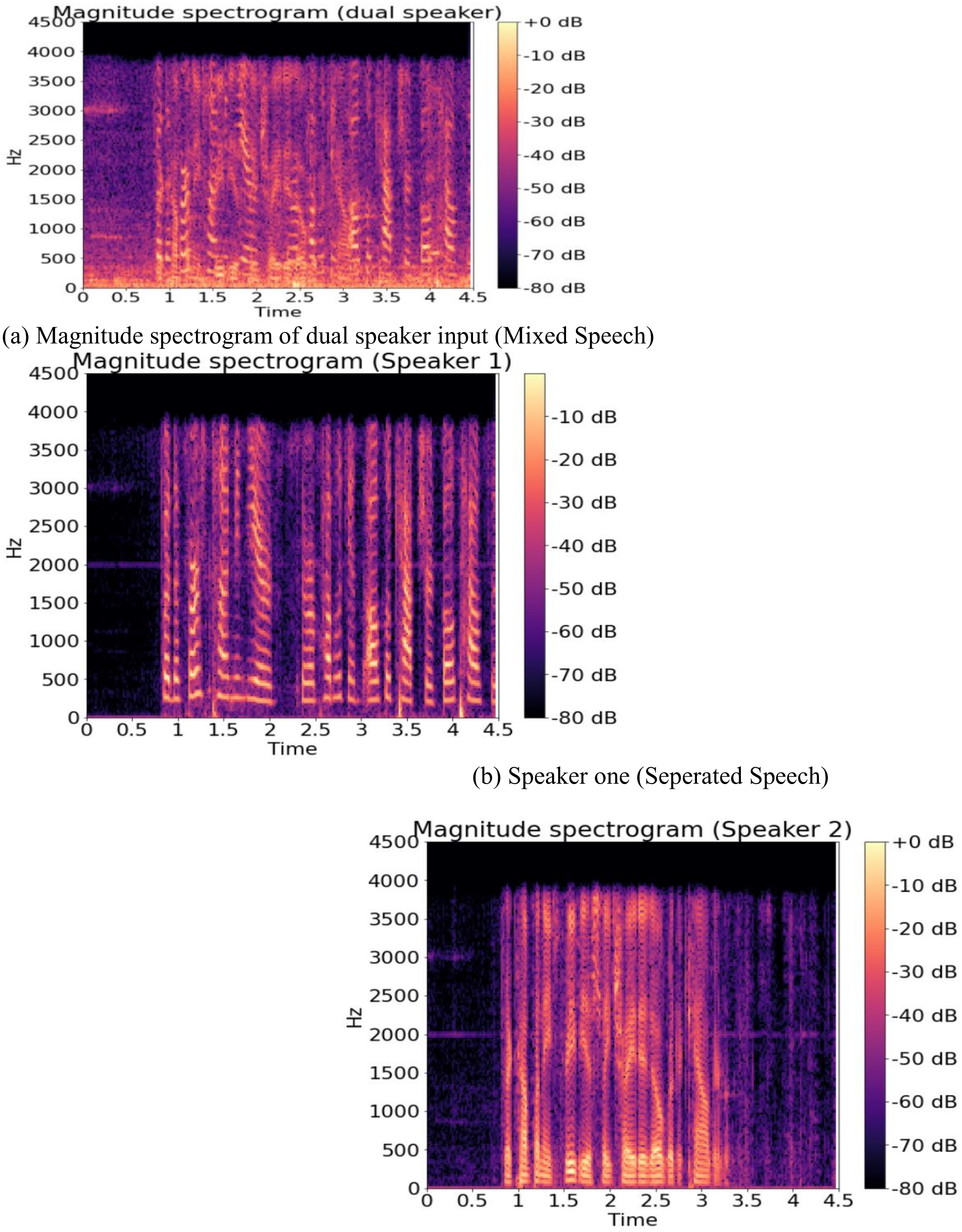


(a) Waveform of dual speaker input (Mixed Speech)

(b) Speaker one (Seperated Speech)

(c) Speaker two (Seperated Speech)

Figure 8. Time domain representation



(a) Magnitude spectrogram of dual speaker input (Mixed Speech)

(b) Speaker one (Seperated Speech)

(c) Speaker two (Seperated Speech)

Figure 9. Magnitude spectrograms

.

8. CONCLUSION AND FUTURE SCOPE

NMF is a prominent factorisation or dimension reduction technique in multivariate analysis by enhancing the most important feature of any machine learning algorithm i.e., interpretability as it is a parts based representation and has a non-negative constraint. This technique mainly factorises or decomposes a non-negative data matrix into a two lower dimensional matrices comprising basis and coefficients. This technique even has a better performance than classic PCA in some instances by making post-processing task much simpler than in many other techniques. The deep recurrent variation is formed by iterating the algorithm of ISTA with warm start. It can be observed that the optimisation problem faced by sparse NMF is addressed by this variation and it also resulted in a much better performance than SP-NMF when a less amount of data is available with respect to training and validation losses.

Although NMF is an efficient algorithm in many ways, there are some queries that had to be addressed like the completeness of the algorithm where theoretical results are not in bunch that support its decompositions as well as the factorisation point of view, namely the rank of the non-negative matrix. From a statistical perspective, a firm base to use a standard framework has to be developed. This standard framework till now has been the SED objective function for NMF and its variants, which goes on the single viewpoint of optimizing the objective function and focusing on the actual goal to identify the original components of the data set considered. So, there is a need to pursue different objective functions and make first hand NMF algorithms. Generalisation of NMF and opting for optimal solutions globally rather than satisfying to local optimal solutions had to be achieved. As seen in the above sections variations are possible either by applying different constraints or applying these to other networking algorithms will deem open to a whole new universe that need to be explored.

REFERENCES

[1] P. Paatero and U. Tapper, “Positive Matrix Factorization: a Non-negative Factor Model with Optimal Utilization of Error Estimates of Data Values,” Environmetrics, vol. 5, no. 2, pp. 111–126, 1994.

[2] D. Lee and H. Seung, “Learning the Parts of Objects by Nonnegative Matrix Factorization,” Nature, vol. 401, no. 6755, pp. 788–791, 1999.

[3] “Non-negative Matrix Factorization,” Wikipedia, http://en. wikipedia.org/wiki/Non-negative matrix factorization.

[4] D. Field, “What is the goal of sensory coding?” Neural computation, vol. 6, no. 4, pp. 559–601, 1994.

[5] M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons, “Algorithms and Applications for Approximate Nonnegative Matrix Factorization,” Computational Statistics & Data Analysis, vol. 52, no. 1, pp. 155–173, 2007.

[6] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. West Sussex, United Kingdom: John Wiley & Sons, 2009.

[7] I. Buciu, “Non-negative Matrix Factorization, a New Tool for Feature Extraction: Theory and Applications,” Int. J. Computers, Communications and Control, vol. 3, Suppl. S, pp. 67–74, 2008.

[8] I. Kotsia, S. Zafeiriou, and I. Pitas, “A Novel Discriminant Non-negative Matrix Factorization Algorithm with Applications to Facial Image Characterization Problems,” IEEE Trans. Infor. Forensics & Security, vol. 2, no. 3, pp. 588–595, 2007.

[9] H. Kim and H. Park, “Sparse Non-negative Matrix Factorizations via Alternating Non-negativity-Constrained Least Squares for Microarray Data Analysis,” Bioinformatics, vol. 23, no. 12, p. 1495, 2007

[10] R. Zdunek and A. Cichocki, “Non-negative Matrix Factorization with Quasi-Newton Optimization,” in Proc. 8th Int. Conf. Artificial Intelligence and Soft Computing, 2006, pp. 870–879.

[11] Y. -X. Wang and Y. -J. Zhang, "Nonnegative Matrix Factorization: A Comprehensive Review," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336-1353, June 2013, doi: 10.1109/TKDE.2012.51.

[12] D. Donoho and V. Stodden, “When does Non-negative Matrix Factorization Give a Correct Decomposition into Parts?” in Proc. Advances in neural information processing systems 16, 2004, pp. 1141–1148.

[13] N. Vasiloglou, A. Gray, and D. Anderson, “Non-negative Matrix Factorization, Convexity and Isometry,” in Proc. SIAM Data Mining Conf., 2009, pp. 673–684.

[14] N. Gillis and F. Glineur, “Nonnegative Factorization and the Maximum Edge Biclique Problem,” Arxiv preprint arXiv:0810.4225, 2008.

[15] N. Mohammadiha and A. Leijon, “Nonnegative Matrix Factorization Using Projected Gradient Algorithms with Sparseness Constraints,” in Proc. IEEE Int. Symposium on Signal Processing and Information Technology, 2009, pp. 418–423.

[16] P. Hoyer, “Non-negative Sparse Coding,” in Proc. IEEE Workshop on Neur. Networks for Signal Pro., 2002, pp. 557–565.

[17] M. Mørup, K. Madsen, and L. Hansen, “Approximate L0 Constrained Non-negative Matrix and Tensor Factorization,” in Proc. IEEE International Symposium on Circuits and Systems, 2008, pp. 1328–1331.

[18] Y. Gao and G. Church, “Improving Molecular Cancer Class Discovery Through Sparse Non-negative Matrix Factorization,” Bioinformatics, vol. 21, no. 21, pp. 3970–3975, 2005.

[19] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” J. Royal Statistical Society. Series B, vol. 58, no. 1, pp. 267–288, 1996

[20] Pham, T.Q., Lee, Y., Lin, Y., Tai, T., & Wang, J. (2015). Single Channel Source Separation Using Sparse NMF and Graph Regularization. ASE BD&SI.

[21] S. Li, X. Hou, H. Zhang, and Q. Cheng, “Learning Spatially Localized, Parts-based Representation,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2001, pp. 207–212.

[22] C. Ding, T. Li, W. Peng, and H. Park, “Orthogonal Non-negative Matrix Tri-Factorizations for Clustering,” in Proc.12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2006, pp. 126–135.

[23] Z. Li, X. Wu, and H. Peng, “Nonnegative Matrix Factorization on Orthogonal Subspace,” Pattern Recognition Letters, vol. 31, no. 9, pp. 905–911, 2010.

[24] S. Choi, "Algorithms for Orthogonal Nonnegative Matrix Factorization," in Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN), 2008, pp. 1828–1832.
[25] Y. Wang, Y. Jia, C. Hu, and M. Turk, "Fisher Non-negative Matrix Factorization for Learning Local Features," in Proc. Asian Conf. Comp Vision, 2004, pp. 27–30.
[26] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting Discriminant Information in Nonnegative Matrix Factorization with Application to Frontal Face Verification," IEEE Trans. Neural Network, vol. 17, no. 3, pp. 683–695, 2006.
[27] C.-J. Lin, "Projected Gradient Methods for Nonnegative Matrix Factorization," Neural Computation, vol. 19, no. 10, pp. 2756–2779, 2007.
[28] M. Gupta and J. Xiao, "Non-Negative Matrix Factorization as a Feature Selection Tool for Maximum Margin Classifiers," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2011
[29] D. Cai, X. He, J. Han, and T. Huang, "Graph Regularized Non-negative Matrix Factorization for Data Representation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 8, pp. 1548– 1560, 2011.
[30] D. Cai, X. He, X. Wu, and J. Han, "Non-negative Matrix Factorization on Manifold," in Proc. 8th IEEE Int. Conf. Data Mining (ICDM), 2008, pp. 63–72.
[31] T. Zhang, B. Fang, Y. Tang, G. He, and J. Wen, "Topology Preserving Non-negative Matrix Factorization for Face Recognition," IEEE Trans. Image Processing, vol. 17, no. 4, pp. 574–584, 2008.
[32] Q. Gu and J. Zhou, "Neighbourhood Preserving Nonnegative Matrix Factorization," in Proc. 20th British Machine Vision Conference, 2009.
[33] B. Shen and L. Si, "Nonnegative Matrix Factorization Clustering on Multiple Manifolds," in Proc. 24th AAAI Conf. Artificial Intelligence (AAAI), 2010, pp. 575–580.
[34] R. Zhi and Q. Ruan, "Discriminant Sparse Nonnegative Matrix Factorization," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2009, pp. 570–573.
[35] S. An, J. Yoo, and S. Choi, "Manifold-Respecting Discriminant Nonnegative Matrix Factorization," Pattern Recognition Letters, vol. 32, no. 6, pp. 832–837, 2011.
[36] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Manifold Regularized Discriminative Non-negative Matrix Factorization with Fast Gradient Descent," IEEE trans. Image Processing, vol. 20, no. 2030–2048, 2011.
[37] Y. Mao and L. Saul, "Modeling Distances in Large-Scale Networks by Matrix Factorization," in Proc. 4th ACM SIGCOMM Conf. Internet Measurement, 2004, pp. 278–287.
[38] S. Zhang, W. Wang, J. Ford, and F. Makedon, "Learning from Incomplete Ratings using Non-negative Matrix Factorization," in Proc.6th SIAM Int. Conf. Data Mining (SDM), 2006, pp. 549–553.
[39] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1462 –1469, July 2006.
[40] P. Smaragdis, "Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs," in Proc. 5th Int. Conf. Independent Component Analysis and Blind Signal Separation, 2004, pp. 494–499.
[41] P. O'Grady and B. Pearlmutter, "Convolutive Non-negative Matrix Factorisation with a Sparseness Constraint," in Proc. 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, 2006, pp. 427–432.
[42] J. Yoo and S. Choi, "Orthogonal Nonnegative Matrix TriFactorization for Co-Clustering: Multiplicative Updates on Stiefel Manifolds," Information Processing & Management, vol. 46, no. 5, pp. 559–570, 2010.
[43] A. Pascual-Montano, J. Carazo, K. Kochi, D. Lehmann, and R. Pascual-Marqui, "Nonsmooth Nonnegative Matrix Factorization (nsNMF)," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 3, pp. 403–415, 2006.
[44] L. Li and Y.-J. Zhang, "Linear Projection-Based Non-negative Matrix Factorization," Acta Automatica Sinica, vol. 36, no. 1, pp. 23–39, 2010.
[45] C. Ding, T. Li, and M. Jordan, "Convex and Semi-nonnegative Matrix Factorizations," IEEE Trans. Pattern Analysis and Machine Intelligence, no. 1, pp. 45–55, 2010.
[46] A. Cichocki and R. Zdunek, "Multilayer Nonnegative Matrix Factorization," Elec. Lett., vol. 42, no. 16, pp. 947–948, 2006.
[47] M. Welling and M. Weber, "Positive Tensor Factorization," Pattern Recognition Letters, vol. 22, no. 12, pp. 1255–1261, 2001.
[48] M. Mørup, L. Hansen, and S. Arnfred, "Algorithms for Sparse Nonnegative Tucker Decompositions," Neural computation, vol. 20, no. 8, pp. 2112–2131, 2008.
[49] T. Hazan, S. Polak, and A. Shashua, "Sparse Image Coding Using a 3D Non-negative Tensor Factorization," in Proc. 10th IEEE Int. Conf. Computer Vision (ICCV), vol. 1, 2005, pp. 50–57.
[50] M. Heiler and C. Schnorr, "Controlling Sparseness in Non- negative Tensor Factorization," in Proc. 9th European Conf. Computer Vision (ECCV), Graz, Austria, 2006, pp. 56–67.
[51] S. Zafeiriou, "Discriminant Nonnegative Tensor Factorization Algorithms," IEEE Trans. Neural networks, vol. 20, no. 2, pp. 217–235, 2009.
[52] L. Li and Y.-J. Zhang, "Non-negative Matrix-Set Factorization," Chinese J. Electronics and Information Technology, vol. 31, no. 2, pp. 255–260, 2009.
[53] L. Li and Y.-J. Zhang, "Bilinear Form-Based Non-Negative Matrix Set Factorization," Chinese J. Computers, vol. 32, no. 8, pp. 1536–1549, 2009.
[54] I. Buciu, N. Nikolaidis, and I. Pitas, "Nonnegative Matrix Factorization in Polynomial Feature Space," IEEE Trans. Neural networks, vol. 19, no. 6, pp. 1090–1100, 2008.

[55] Wisdom, Scott & Powers, Thomas & Pitton, James & Atlas, Les. (2017). Deep recurrent NMF for speech separation by unfolding iterative thresholding. 254-258. 10.1109/WASPAA.2017.8170034.

[56] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: An overview of challenge systems and outcomes," in Proc. ASRU, Olomouc, Czech Republic, 2013, pp. 162–167.

[57] E. Vincent, "BSS Eval toolbox version 3.0," " http://bass-db. gforge.inria.fr/bss eval.

[58] Shu, Zhen-qiu & Wu, Xiao-jun & Hu, Cong & You, Cong-zhe & Fan, Hong-hui. (2021). Deep semi-nonnegative matrix factorization with elastic preserving for data representation. Multimedia Tools and Applications. 80. 1-18. 10.1007/s11042-020-09766-w.