NONNEGATIVE MATRIX FACTORICATION (NMF)

JUMAN AHMED

1. Introduction

Development of computer technology has increased the quantities of row data collection. This continuous collection of data led to the question of what should be done with the data. Thus, the importance of appropriate dimensional reduction technique in multivariate data analysis to the represent these raw data became essential. In general, there are two basic properties must be satisfied: firstly, the dimension of the original raw data should be reduced to appropriate factors that are necessary to analyze, and secondly, the principal components, hidden concepts, prominent features, or latent variables of the data, depending on the application context, should be identified efficaciously. There are many acceptable methods with low rank approximation such as Principal Component Analysis (PCA), Discriminant Analysis (DA), or Vector Quantization (VQ) can be used to decompose the data matrix into two factor matrices for easier computation and analyses, however, the properties of nonnegative matrix factorization (NMF) has significant advantage in deep exploration of row data for higher statistical outcomes. NFM also became an imperative multivariate data analytical tool that has been used in the field of mathematics, optimization, neural computing, pattern recognition, machine learning, data mining, signal processing, image restoration, image segmentation, audio pattern separation, face hallucination, face recognition and many others applications. In this paper, we will discuss the principles, basic models, properties, and algorithms of NMF including various modifications, extensions, generalizations, and how it can be useful in decomposition of the original data matrix to low rank two factor matrices.

2. Concept and Properties of NMF

Definition 2.1. Given an M dimensional random vector x with nonnegative elements, whose N observations are denoted as $x_j = 1, 2, ..., N$. Let data matrix be $X = [x_1, x_2, ..., x_N] \in \mathbb{R}^{M\times N}_{\geq 0}$, then NMF seeks to decompose X into nonnegative $M \times L$ basis matrix $U = [u_1, u_2, ..., u_L] \in \mathbb{R}^{M\times L}_{\geq 0}$ and nonnegative $L \times N$ coefficent matrix $V = [v_1, v_2, ..., v_L] \in \mathbb{R}^{L\times N}_{\geq 0}$, such that $X \approx UV$. This can be written as the equivalent vector formula:

$$x_j \approx \sum_{i=1}^{L} u_i V_{ij}$$

In most cases, an alternative model of NMF: $X = UV + E$ is widely use where $E \in \mathbb{R}^{M\times N}$ is the residue or noise matrix representing the approximation error.

In the above definition, v_j is the weighted coefficient of the observation x_j on the columns of U, the basis vectors of X. Hence, NMF decomposes each data set into the linear combination of

the basis vectors. The basis vectors are incomplete over the original vector space because of the initial condition $L << min(M, N)$. But in general, L can be smaller, equal or larger then M while having fundamental differences in the decomposition for $L > M$ and $L < M$. Spare coding and compressed sensing with over complete basis is the cause of the variation in dimension of L. Hence, the dimensionality of L is not limited of the data, which is useful for some applications such as classification.

Example 2.2. Let us consider the analysis of a newspaper data matrix. The data matrix consists of the word counts of the newspaper articles. In Figure 1, colored column vectors correspond to articles, and a number in a column vector represents the number of times a word appears in an article. In this case, the pattern matrix consists of column vectors, where each vector corresponds to a topic discussed in articles, such as sports, economics, and politics. Each pattern has its own word frequency distribution. Hence, it is possible to identify the contents of a pattern. Moreover, each article (data-set) has its own pattern strengths, as summarized in a strength matrix. Using this strength matrix, we can easily summarize the contents of many articles. In summary, NMF extracts patterns and pattern strengths at the same time by decomposing the original data matrix into two sets of nonnegative matrices.

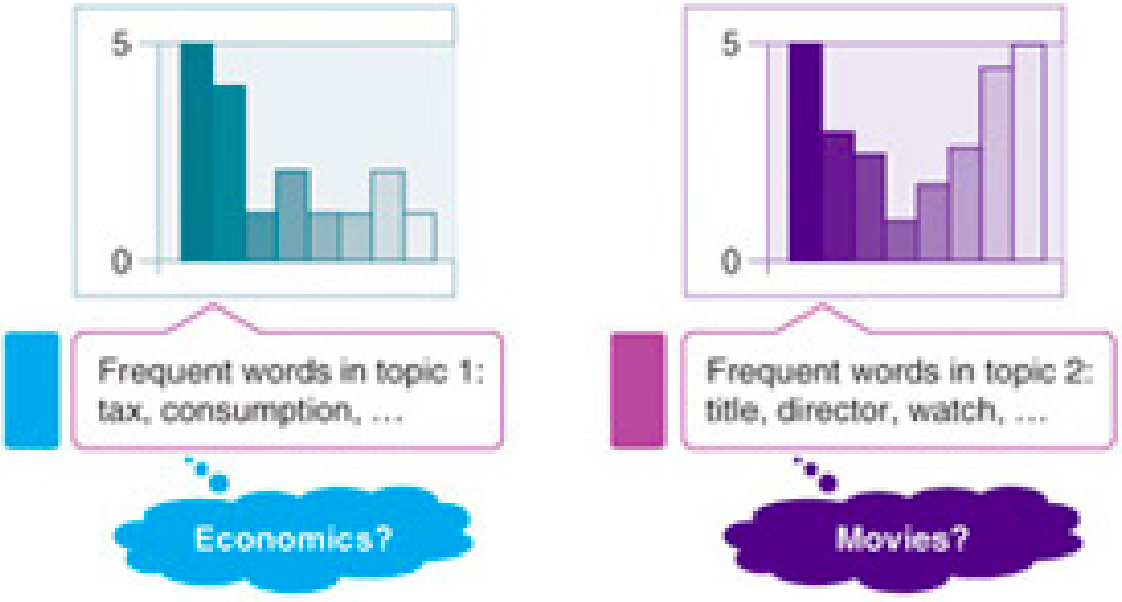

Figure 1: The appearance frequencies of words in an summary

As a kind of matrix factorization model, there are three important questions need answering for NMF: 1) existence of a non trivial solutions; 2) under what assumptions the solution for NMF is unique; 3) under what assumptions the solutions for NMF is valid or "Correct". A complete NMF $X = UV$ is considered first for the analysis of existence, convexity and computational complexity. The trivial solution always exists as $U = X$ and $V = I_N$.

The existing NMF algorithms are divided into four categories illustrated in Figure 2, following some unified criteria:

1. Basic NMF imposes the non-negative constraint.

2. Constrained NMF imposes additional constraints as regularization.
3. Structured NMF modifies the the standard factorization formulations.
4. Generalized NMF break through the conventional data type or factorization modes in a broad sense.

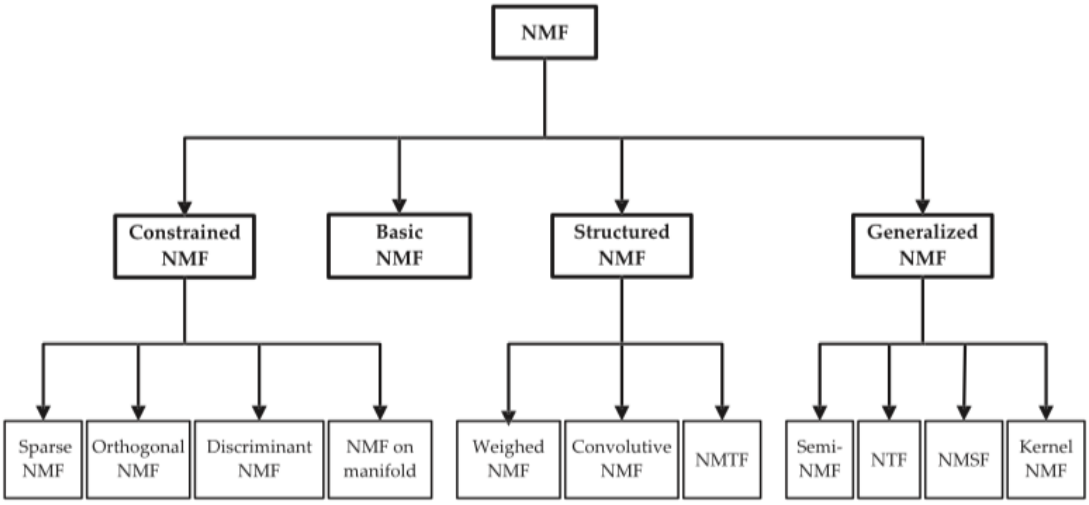

Figure 2: The categorization of NMF models and algorithms.

3. Basic NMF Algorithm

The core of Basis NMF is to find an efficient and effective solution to NMF problem under the sole nonnegative constraint. Due to lack of appropriate convex formulations, the nonconvex formulations with relatively easy computation are general adopted and only local minima are achievable in a reasonable computational time. Hence, the most practical approach is to perform alternating minimization of a suitable cost function as the similarity measures between X and the product UV. In this section, we will discuss the objective functions followed by the details about the Basic NMF optimization framework and the paragon algorithms. We will also summarize the some new vision of NMF, such as geometric formulations of NMF and the pragmatic issues of NMF, such as large scale data sets, online processing, parallel computing.

3.1. **Objective Functions.** The objective function or a similarity measures of the optimization model, $D(X||UV)$, quantifies the difference between the original data X and the approximation UV. These similarity measures can be either distances or divergences, and corresponding objective function can a cost function with the same global minima to be minimized sequentially. The most commonly used objective functions are:
Stream Editor (SED)

$$D_F(X||UV) = \frac{1}{2}||X - UV||_F^2 = \frac{1}{2}\sum_{ij}\left(X_{ij} - [UV]_{ij}\right)^2$$

Kullback-Leibler divergence (KLD)

$$D_{KL}(X||UV) = \sum_{ij} \left(X_{ij} ln \frac{X_{ij}}{[UV]_{ij}} - X_{ij} + [UV]_{ij} \right)$$

The gradients needed in optimization heavily depends on the scales of factorizing matrices. Thus, the original KLD was manipulated more efficiently for NMF to normalize the input data. Most of the cost functions are element-wise measured, however, some are more robust with respect to noise and out-liers such as hyper-surface cost function. Based on statistical knowledge about the probability distribution of the noise that reflects the statistical structure of the signals and the disclosed components can be utilize to differentiate the objective functions. For example, the SED minimization can be seen as a maximum likelihood estimator where the difference is due to additive Gaussian noise, whereas KLD can be shown to be equivalent to the Expectation Maximization algorithm and maximum likelihood for Poisson distribution.

3.2. **Classic Basic NMF Optimization Framework.** Using SED as the objective function, the multiplicative and gradient descent algorithms are special cases of general framework called Alternating Nonnegative Least Squares (ANLS). It is a modification of Alternating Least squares under nonnegative constraint. Taking into a count of seperate convexity, a two variable optimization problems is converted into the Nonnegative Least Squares optimization mini-problems. Given the SED function of two factor matrices, the conversion procedure performed a constrained minimization with respect to one matrix while unchanging the other matrix. Then minimization is performed again with the role of the matrices reversed as follows:

$$min_{U \geq 0} \; D_F(X||UV) = min_{U \geq 0} \frac{1}{2} ||X - UV||_F^2$$

$$min_{V \geq 0} \; D_F(X||UV) = min_{V \geq 0} \frac{1}{2} ||X - UV||_F^2$$

This approach called "Block Coordinate Descent", where two blocks can be partitioned into several more basic separable cells, i.e., find a global minimum of each mini-problems for each block of variables. This procedure can also be applicable for the KLD objective function.

3.3. **Paragon Algorithms.** An interior-point gradient method was proposed using a search direction equivalent to the Least Squares algorithms. The objective function is minimized if the updated values is positive. If not, then a certain proportion of the longest step is chosen as the update step to ensure the positive constraint. This operation decreases the objective function as its minimum while guaranteeing the nonnegativity. To expedite the decreasing speed of the objective function, Gonzales ad Zhang introduced a multiplicative regulatory factor beside the original adaptive learning rate. In most cases, the first order optimization scheme is proper for approximate NMF with noise (error) added. For more accurate solution, a second order optimization on objective function using Taylor Series expansion can be utilize.

The cynosure in ANLS or Block Coordinate descent consists in the partition of variables with convexity preserved. From definition of NMF, it is clear that if every block does not contain an element of a column of U and an element of the corresponding row of V, then the optimization problem under partition is convex. Further more, given a subset of indices $K \subseteq R = 1, 2, ..., L$,

NMF is convex for the two following subsets of variables:

$$P_k = \{U_{\bullet i} | i \in K\} \cup \{V_{j\bullet} | j \in K\}$$

and its complement

$$Q_K = \{U_{\bullet i} | i \in R \setminus K\} \cup \{V_{j\bullet} | j \in K\}$$

In the standard NLS approach, the minimization using separate convexity is further separated into M independent NLS mini-problems in L variables corresponding to each row of U.

3.4. **New Vision of Basic NMF.** The Basic NMF algorithms are all algebraic iterative optimization models and the solutions are not guaranteed to be unique due to the sensitiveness in initialization. Thus, Klingenberg proposed a seminal formulation coined the Extreme Vector Algorithm (EVA), on the basis of the geometric interpretation of NMF. In the reduced space or non-singular condition, EVA searches for smallest cone containing overall data points, which is the most informative with respect to where the data are located in the positive orthant. This is identical to selecting the vertex vectors of the projected boundary polygon of the original data points on the unit hypersphere as the basis vector. This manipulate decouples the functions of reducing dimensionality and identifying latent data structure into two independent stages. Thus it might yield in better performance and solution. The EVA can be regarded as conic coding under certain constraints imposed on the basis vectors. Correspondingly, convex NMF can be viewed as convex coding. More specifically, the data set is generated by the basis vectors u_i; reversely, the basis vectors u_i are also generated by the data set.

3.5. **Pragmatic Issue.** For large scale NMF where $L << M$ and $L << N$, X is usually low rank which implies that the problem $X \approx UV$ become highly redundant. Hence, There is no need to process all elements of X to estimate U and V precisely.

Because of the local rather than global minimization characteristic, it is obvious that the initialization of U and V will directly influence the convergence rate and the solution quality. Poorly initialize conditions might converge slowly to incorrect or even irrelevant solutions. Another ignored issue is the choice of the objective functions. Although, SED is the mostly used and deeply investigated objective function, different objective functions correspond to varied probability distribution assumptions.

4. Constrained NMF Algorithms

The various Constrained NMF models can be unified under similar extended objective function:

$$D_C(X||UV) = D(X||UV) + \alpha J_1(U) + \beta J_2(V)$$

where $J_1(U)$ and $J_2(V)$ are the penalty terms to enforce certain application dependent constraints. Also, α and β are small regularization parameters balancing the trade off between the fitting goodness and the constraints. Based upon $J_1(U)$ and $J_2(V)$, Constrained NMF algorithm are categorize into four classes:

Sparse NMF

Orthogonal NMF

Discriminant NMF
NMF on manifold

4.1. **Sparse NMF.** In Sparse NMF algorithm, factor matrix, U or V, is selected as the candidate to impose the sparseness constraint. If the basis vector, column of U, are sparse, then every basis influence only a small part of each observation. On the other hand, If column of V are sparse, each observation is approximated by a linear combination of a limited number of basis vectors. If rows of V are sparse, then each basis vector is used to approximate a limited number of training data used to infer each basis vectors.

4.2. **Orthogonal NMF.** Orthogonal NMF is NMF with Orthogonal constraint on either the factor U or V. The orthogonal principle is use to minimize the redundancy between different bases. Orthogonality results in sparseness in nonnegative conditions, however, there is notable difference in the optimization models between the Sparse NMF and Orthogonal NMF. The result of Orthogonal NMF corresponds to a unique sparse area in the solution region by deeply sorting the distinct parts. If the basis vectors, column of U, are orthogonal, namely $U^TU = I$, it obtains the most distinct parts. If row of V are orthogonal, namely $V^TV = I$, it imporves the clustering accuracy. Biorthogonality imposes the orthogonality in both U and V with lack in approximation performance. Orthogonal NMF on U or V is identical to clustering the row or column of an input data matrix, where one matrix corresponds to the cluster centers and the other is associated with the cluster indicator vector. Orthogonal NMF is preferable in clustering due to the close interpretation and relationship.

4.3. **Discriminant NMF.** Basic NMF can be considered as unsupervised learning from the perspective of pattern recognition. However, combining with discriminant analysis with decomposition, Basic NMF is further extended to supervised alternatives, so called Discriminant NMF. Discriminant NMF method unifies the generative model and the classification model into a joint framework, which is being used successfully in many applications such as face recognition. The KLD formula with the penalty terms often accepted as the basic framework. Considering the differences in definition of Within-class and Between-class scatter, both are solely based on the coefficient matrix V and have nothing to do with X or U. Since the actual classification features are closely relevant to the projection matrix U^TX, while only having indirect connection with V, U^TX is impose as the construction of the discriminant constraint. This operation makes the basis vectors somewhat sparse with distinct parts-based, which helpful for classification.

4.4. **NMF on Manifold.** The real world data are often sampled from a nonlinear low-dimensional submanifold embedded in a high-dimensional space, which is appears as euclidean space. The learning performance can be significantly enhanced if the intrinsic geometrical structure is identified and preserved. An important local topological property is the linear embedded assumption, that is, the data point generated as a liner combination of several neighboring points on a specific manifold in the original space. Exploiting the local topological property, the neighborhood preserving NMF can be modified to multiplicative update rules. Approximation of the data

points can be constructed by exploring the locally linear relationship of single manifold with multiple manifold.

5. Structured NMF Algorithms

Structured NMF modifies the regular factorization formulation directly rather than introducing additional constraint as penalty terms shown in Constraint NMF. Formally it can be written as:

$$X \approx F(UV)$$

Structured NMF algorithm categorized into three classes:

Weighed NMF
Convolutive NMF
Nonnegative Matrix Trifactorization (NMTF)

5.1. **Weighed NMF.** Weighed NMF formulation are commonly modified version of learning algorithm, which can be utilized to emphasize the relative importance of different components. The formula for weighed NMF is:

$$W \otimes X \approx W \otimes (UV)$$

where W is a weight matrix. In general, Weighed NMF can be viewed as a case weighted low rank approximation, which seeks for a low rank matrix closest to the input matrix. If the original data matrix is incomplete with some missing entries, W predicts the missing information, which often referred to as low rank matrix completion with noise.

5.2. **Convolutive NMF.** The notion of Convolutive NMF comes from the application of source separation. Conventional Basic NMF decomposes the basis matrix U and the corresponding coefficient matrix V. The potential dependency between the neighboring column vectors of the input data matrix X, it is necessary to take into account the time-varying characteristic. Hence, Basic NMF is extended to Convolutive NMF form, which formulated as following:

$$X \approx \sum_{t=0}^{T-1} U_t(V(t \rightarrow))$$

where U_t basis matrices that varies across time, $(V(t \rightarrow))$ coefficient matrices that satisfies the relationship of right shift and zero padding. Hence, Convolutive NMF can be decomposed into a series of Basis NMF problems.

5.3. **Nonnegative Matrix Trifactorization (NMTF).** NMTF extends conventional NMF to the product of three factor matrices,$X \approx USV$, with constraint, which provides additional degree of freedom. Taking into account of biorthogonality, which has poor low rank approximation in both U and V, an additional factor martix S is being introduce to absorb the different scales of U and V. This easure that the low rank matrix representation remains accurate while satisfying the orthogonality constraint on both U and V. Hence, the rows and columns of X can be clustered simultaneously.

6. Generalized NMF Algorithms

Generalized NMF might be considered as an extension to decompose model itself in depth similar to Structured NMF. It breaches the intrinsic nonnegativity constraint to some extent, or changes the data types, or alters the factorization pattern. Generalized NMF categorized into four classes:

Semi-NMF
Nonnegative Tensor Factorization (NTF)
Nonnegative Matrix-set Factorization (NMSF)
Kernel NMF

6.1. **Semi-NMF.** Conventional NMF restricts every element in data matrix X to be nonnegative. When X is unconstrained, an extended version referred to as Semi-NMF which remains some kernel concept of NMF. It is where V is still restricted to be nonnegative while placing no restriction on the signs of U. An alternating iteration approach to solve the optimization problem was employed where the positive and negative parts are separated from the mixed-sign matrix. The multiplicative rules used in V to updated while holding U fixed, and then the analytical local optimal solution for U is obtained with V fixed.

6.2. **Nonnegative Tensor Factorization (NTF).** A generalization of matrix factorization is tensor factorization, which protects the original structure of the data contrast to Conventional methods of arranging data into matrix. NMF is a particular case of nonnegative n-dimensional tensor factorization (n-NTF) when $n = 2$. Since, NTF possesses many new properties varying from NMF, this generalizarion is not trivial. First, the data processed in NMF are vectors, however, in some applications the original data may not be vectors, and the vectorization might result in some undesirable problems. For instance, the vectorization of image data, which is two dimensional, will lose the local spatial and structural information. Secondly, one of the main concerns in NMF is the uniqueness issue, and tensor factorization will only be unique under some weak conditions while imposing some strong constraints.

6.3. **Nonnegative Matrix-set Factorization (NMSF).** NMSF is imposed directly on the matrix set, whose candidates to be processed are the set of sample matrices. Each sample matrix is decomposed into the product of K factor matrices, where the public $K - 1$ factor matrices represent the learned features which generalize the feature matrix in NMF to a feature matrix set. The remaining factor matrix varying from individual sample matrix describes the activation patterns which generalizes the coefficient vector in NMF to a coefficient matrix. NMSF only concentrates on the 3D situation, where it is more flexible than nonnegative 3D tensor factorization.

6.4. **Kernel NMF.** Variants methods of NMF are essentially linear models. But, when it comes to nonlinear models, it is unable to provide the appropriate structures and relationships hidden in the input data. Thus, a Kernel method is usually implemented by mapping input data into an implicit feature space using nonlinear functions. It is also potential in processing data with negative values by using some specific kernel functions. Given a nonlinear mapping

$\phi : \mathbb{R}^M \to \mathbb{R}$, $x \to \phi(x)$, which which maps the input data space $\mathbb{R}^M$ into the feature space $\mathbb{R}$. Also the original data martix is transformed into $X \to Y = \phi(X) = [\phi(x_1), \phi(x_2), ..., \phi(x_N)]$. Kernel NMF seeks to factor martrices $Z = [\phi(u_1), \phi(u_2), ..., \phi(u_L)]$ and V, such that $Y \approx ZV$, where $u_1, u_2, ..., u_L$ are basis vectors. SED is chosen as the objective function which rewritten using Y as the original data matrix and Z and V are the two factor matrices as followed:

$$D_{KF}(Y||ZV) = \frac{1}{2}||Y - ZV||_F^2 = \frac{1}{2}tr(Y^T Y) - tr(Y^T ZV) + \frac{1}{2}tr(V^T Z^T ZV).$$

Using kernel matrices K^{xx}, K^{uu}, and K^{xu}, where $K_{ij}^{xx} = \langle \phi(x_i), \phi(x_j) \rangle$, $K_{ij}^{uu} = \langle \phi(u_i), \phi(u_j) \rangle$, and $K_{ij}^{xu} = \langle \phi(x_i), \phi(u_j) \rangle$ are denoted as the inner product of each consecutive terms, the above objective function can be modify as:

$$D_{KF}(Y||ZV) = \frac{1}{2}tr(K^{xx}) - tr(K^{xu}V) + \frac{1}{2}tr(V^T K^{uu} V).$$

Thus, the model depends onlyon the kernel matrices.

7. Conclusion

As a multivariate data analysis and dimensionality reduction technique, NMF seeks the compression and interpretability due to its parts-based and sparse representation from the nonnegativity or purely additive constraint. It outperforms classic low-rank approximation approaches such as PCA in some cases, and makes the post-processing much easier.