NONNEGATIVE MATRIX FACTORIZATION AND APPLICATIONS

MOODY CHU and ROBERT PLEMMONS
Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205.
Departments of Computer Science and Mathematics, Wake Forest University, Winston-Salem, NC 27109.

1 Introduction

Data analysis is pervasive throughout science, engineering and business applications. Very often the data to be analyzed is nonnegative, and it is very often preferable to take this constraint into account in the analysis process. In this paper we provide a survey of some aspects of nonnegative matrix factorization and its applications to nonnegative matrix data analysis. In general the problem is the following: given a nonnegative data matrix Y find reduced rank nonnegative matrices U and V so that

$$Y \approx UV.$$

Here, U is often thought of as the source matrix and V as the mixing matrix associated with the data in Y. A more formal definition of the problem is given below. This approximate factorization process is an active area of research in several disciplines (a Google search on this topic recently provided over 250 references to papers involving nonnegative matrix factorization and applications written in the past ten years), and the subject is certainly a fertile area of research for linear algebraists.

An indispensable task in almost every discipline is to analyze a certain data to search for relationships between a set of exogenous and endogenous variables. There are two special concerns in data analysis. First, most of the information gathering devices or methods at present have only finite bandwidth. One thus cannot avoid the fact that the data collected often are not exact. For example, signals received by antenna arrays often are contaminated by instrumental noises; astronomical images acquired by telescopes often are blurred by atmospheric turbulence; database prepared by document indexing often are biased by subjective judgment; and even empirical data obtained in laboratories often do not satisfy intrinsic physical constraints. Before any deductive sciences can further be applied, it is important to first reconstruct or represent the data so that the inexactness is reduced while certain feasibility conditions are satisfied. Secondly, in many situations the data observed from complex phenomena represent the integrated result of several interrelated variables acting together. When these variables are less precisely defined, the actual information contained in the original data might be overlapping and ambiguous. A reduced system model could provide a fidelity near the level of the original system. One common ground in the various approaches for noise removal, model reduction, feasibility reconstruction, and so on, is to replace the original data by a lower dimensional representation obtained via subspace approximation. The notion of low rank approximations therefore arises in a wide range of important applications. Factor analysis and principal component analysis are two of the many classical methods used to accomplish the goal of reducing the number of variables and detecting structures among the variables.

However, as indicated above, often the data to be analyzed is nonnegative, and the low rank data are further required to be comprised of nonnegative values only in order to avoid contradicting physical realities. Classical tools cannot guarantee to maintain the nonnegativity. The approach of low-rank **nonnegative matrix factorization** (NNMF) thus becomes particularly appealing. The NNMF problem, probably due originally to Paatero and Tapper [21], can be stated in generic form as follows:

(NNMF) *Given a nonnegative matrix $Y \in R^{m\times n}$ and a positive integer $p < \min\{m, n\}$, find nonnegative matrices $U \in R^{m\times p}$ and $V \in R^{p\times n}$ so as to minimize the functional*

$$f(U, V) := \frac{1}{2}\|Y - UV\|_F^2. \tag{1}$$

The product UV of the least squares solution is called a nonnegative matrix factorization of Y, although Y is not necessarily equal to the product UV. Clearly the product UV is of rank at most p. An appropriate decision on the value of p is critical in practice, but the choice of p is very often problem dependent. The objective function (1) can be modified in several ways to reflect the application need. For example, penalty

terms can be added to $f(U,V)$ in order to enforce sparsity or to enhance smoothness in the solution U and V [13, 24]. Also, because $UV = (UD)(D^{-1}V)$ for any invertible matrix $D \in R^{p\times p}$, sometimes it is desirable to "normalize" columns of U The question of uniqueness of the nonnegative factors U and V also arises, which is easily seen by considering case where the matrices D and D^{-1} are nonnegative. For simplicity, we shall concentrate on (1) only in this essay, but the metric to be minimized in the NNMF problem can certainly be generalized and constraints beyond nonnegativity are sometimes imposed for specific situations, e.g., [5, 13, 14, 15, 18, 19, 24, 25, 26, 27]. In many applications, we will see that the p factors, interpreted as either sources, basis elements, or concepts, play a vital role in data analysis. In practice, there is a need to determine as few factors as possible and, hence the need for a low rank NNMF of the data matrix Y arises.

2 Some Applications

The basic idea behind the NNMF is the linear model. The matrix $Y = [y_{ij}] \in R^{m\times n}$ in the NNMF formulation denotes the "observed" data whereas each entry y_{ij} represents, in a broad sense, the *score* obtained by entity j on variable i. One way to characterize the interrelationships among multiple variables that contribute to the observed data Y is to assume that y_{ij} is a linearly weighted score by entity j based on several "factors". We shall temporarily assume that there are p factors, but often it is precisely the point that the factors are to be retrieved in the mining process. A linear model, therefore, assumes the relationship

$$Y = AF, \tag{2}$$

where $A = [a_{ik}] \in R^{m\times p}$ is a matrix with a_{ik} denoting the *loading* of variable i to factor k or, equivalently, the *influence* of factor k on variable i, and $F = [f_{kj}] \in R^{p\times n}$ with f_{kj} denoting the *score* on factor k by entity j or the *response* of entity j to factor k. Depending on the applications, there are many ways to interpret the meaning of the linear model. We briefly describe a few applications below.

2.1 Air Emission Quality

In the air pollution research community, one observational technique makes use of the ambient data and source profile data to apportion sources or source categories [12, 15]. The fundamental principle in this model is that mass conservation can be assumed and a mass balance analysis can be used to identify and apportion sources of airborne particulate matter in the atmosphere. For example, it might be desirable to determine a large number of chemical constituents such as elemental concentrations in a number of samples. The relationships between p sources which contribute m chemical species to n samples, therefore, lead to a *mass balance equation*,

$$y_{ij} = \sum_{k=1}^{p} a_{ik} f_{kj}, \tag{3}$$

where y_{ij} is the elemental concentration of the ith chemical measured in the jth sample, a_{ik} is the gravimetric concentration of the ith chemical in the kth source, and f_{kj} is the airborne mass concentration that the kth source has contributed to the jth sample. In a typical scenario, only values of y_{ij} are observable whereas neither the sources are known nor the compositions of the local particulate emissions are measured. Thus, a critical question is to estimate the number p, the compositions a_{ik}, and the contributions f_{kj} of the sources.

Tools that have been employed to analyze the linear model include principal component analysis, factor analysis, cluster analysis, and other multivariate statistical techniques. In this receptor model, however, there is a physical constraint imposed upon the data. That is, the source compositions a_{ik} and the source contributions f_{kj} must all be nonnegative. The identification and apportionment, therefore, becomes a nonnegative matrix factorization problem of Y.

2.2 Image and Spectral Data Processing

Digital images are represented as nonnegative matrix arrays, since pixel intensity values are nonnegative. It is sometimes desirable to process data sets of images represented by column vectors as composite objects in

many articulations and poses, and sometimes as separated parts for in, for example, biometric identification applications such as face or iris recognition. It is suggested that the factorization in the linear model would enable the identification and classification of intrinsic "parts" that make up the object being imaged by multiple observations [7, 16, 26]. More specifically, each column $\mathbf{y}_j$ of a nonnegative matrix Y now represents m pixel values of one image. The columns $\mathbf{a}_k$ of A are basis elements in R^m. The columns of F, belonging to R^p, can be thought of as coefficient sequences representing the n images in the basis elements. In other words, the relationship,

$$\mathbf{y}_j = \sum_{k=1}^{p} \mathbf{a}_k f_{kj}, \tag{4}$$

can be thought of as that there are *standard parts* $\mathbf{a}_k$ in a variety of positions and that each image represented as a vector $\mathbf{y}_j$, making up the factor U of basis elements is made by superposing these parts together in specific ways by a mixing matrix represented by V in (1). Those parts, being images themselves, are necessarily nonnegative. The superposition coefficients, each part being present or absent, are also necessarily nonnegative. A related application to the identification of object materials from spectral reflectance data at different optical wavelengths has been investigated in [25].

2.3 Text Mining

Assume that the textual documents are collected in an *indexing matrix* $Y = [y_{ij}] \in R^{m\times n}$. Each document is represented by one column in Y. The entry y_{ij} represents the *weight* of one particular *term* i in document j whereas each term could be defined by just one single word or a string of phrases. To enhance discrimination between various documents and to improve retrieval effectiveness, a term-weighting scheme of the form,

$$y_{ij} = t_{ij} g_i d_j, \tag{5}$$

is usually used to define Y [2], where t_{ij} captures the relative importance of term i in document j, g_i weights the overall importance of term i in the entire set of documents, and $d_j = (\sum_{i=1}^{m} t_{ij} g_i)^{-1/2}$ is the scaling factor for normalization. The normalization by d_j per document is necessary because, otherwise, one could artificially inflate the prominence of document j by padding it with repeated pages or volumes. After the normalization, the columns of Y are of unit length and usually nonnegative.

The indexing matrix contains lot of information for retrieval. In the context of latent semantic indexing (LSI) application [2, 10], for example, suppose a query represented by a row vector $\mathbf{q}^\top = [q_1, \ldots, q_m] \in R^m$, where q_i denotes the weight of term i in the query $\mathbf{q}$, is submitted. One way to measure how the query $\mathbf{q}$ matches the documents is to calculate the row vector $\mathbf{s}^\top = \mathbf{q}^\top Y$ and rank the relevance of documents to $\mathbf{q}$ according to the *scores* in $\mathbf{s}$.

The computation in the LSI application seems to be merely the vector-matrix multiplication. This is so only if Y is a "reasonable" representation of the relationship between documents and terms. In practice, however, the matrix Y is never exact. A major challenge in the field has been to represent the indexing matrix and the queries in a more compact form so as to facilitate the computation of the scores [6, 23]. The idea of representing Y by its NNMF approximation seems plausible. In this context, the standard parts $\mathbf{a}_k$ indicated in (4) may be interpreted as subcollections of some "general concepts" contained in these documents. Like images, each document can be thought of as a linear composition of these general concepts. The column-normalized matrix A itself is a term-concept indexing matrix.

Nonnegative matrix factorization has many other applications, including linear sparse coding [13, 29], chemometric [11, 21], image classification [9], neural learning process [20], sound recognition [14], remote sensing and object characterization [25, 30]. We stress that, in addition to low-rank and nonnegativity, there are applications where other conditions need to be imposed on U and V. Some of these constraints include sparsity, smoothness, specific structures, and so on. The NNMF formulation and resulting computational methods need to be modified accordingly, but it will be too involved to include that discussion in this brief survey.

3 Optimality

Quite a few numerical algorithms have been developed for solving the NNMF. The methodologies adapted are following more or less the principles of alternating direction iterations, the projected Newton, the reduced quadratic approximation, and the descent search. Specific implementations generally can be categorized into alternating least squares algorithms [21], multiplicative update algorithms [16, 17, 13], gradient descent algorithm, and hybrid algorithm [24, 25]. Some general assessments of these methods can be found in [5, 18, 28]. It appears that there is much room for improvement of numerical methods. Although schemes and approaches are different, any numerical method is essentially centered around satisfying the first order optimality conditions derived from the Kuhn-Tucker theory. Recall that the computed factors U and V may only be local minimizers of (1).

Theorem 3.1 *Necessary conditions for* $(U, V) \in R_+^{m\times p} \times R_+^{p\times n}$ *to solve the nonnegative matrix factorization problem (1) are*

$$\begin{aligned} U. * \big((Y - UV)V^\top\big) &= 0 \in R^{m\times p}, && (6)\\ V. * \big(U^\top(Y - UV)\big) &= 0 \in R^{p\times n}, && (7)\\ (Y - UV)V^\top &\le 0, && (8)\\ U^\top(Y - UV) &\le 0, && (9) \end{aligned}$$

where .∗ denotes the Hadamard product.

4 Conclusions and Some Open Problems

We have attempted to outline some of the major concepts related to nonnegative matrix factorization and to briefly discuss a few of the many practical applications. Several open problems remain, and we list just a few of them.

- Preprocessing the data matrix Y. It has been observed, e.g. [25, 27], that noise removal or a particular basis representation for Y can improve the effectiveness of algorithms for solving (1). This is an active area of research and is unexplored for many applications.
- Initializing the factors. Methods for choosing, or seeding, the initial matrices U and V for various algorithms (see, e.g., [30]) is a topic in need of further research.
- Uniqueness. Sufficient conditions for uniqueness of solutions to the NNMF problem can be considered in terms of simplicial cones [1], and have been studied in [7]. Algorithms for computing the factors U and V generally produce local minimizers of $f(U, V)$, even when constraints are imposed. It would thus be interesting to apply global optimization algorithms to the NNMF problem.
- Updating the factors. Devising efficient and effective updating methods when columns are added to the data matrix Y in (1) appears to be a difficult problem and one in need of further research.

Our survey in this short essay is of necessity incomplete, and we apologize for resulting omission of other material or references. Comments by readers to the authors on the material are welcome.

References

[1] A. Berman and R. J. Plemmons, Nonnegative Matrices in the Mathematical Sciences, SIAM, Philadelphia, 1994.

[2] M. W. Berry, Computational Information Retrieval, SIAM, Philadelphia, 2000.

[3] M. Catral, L. Han, M. Neumann and R. J. Plemmons, On reduced rank nonnegative matrix factorizations for symmetric matrices, Lin. Alg. and Appl., Special Issue on Positivity in Linear Algebra, 393(2004), 107-126.

[4] M. T. Chu, On the statistical meaning of the truncated singular decomposition, preprint, North Carolina State University, November, 2000.

[5] M. T. Chu, F. Diele, R. Plemmons, and S. Ragni, Optimality, computation, and interpretation of nonnegative matrix factorizations, preprint, 2004.

[6] I. S. Dhillon and D. M. Modha, Concept decompositions for large sparse text data using clustering, Machine Learning J., 42(2001), 143-175.

[7] D. Donoho and V. Stodden, When does nonnegative matrix factorization give a correct decomposition into parts, Stanford University, 2003, report, available at http://www-stat.stanford.edu/~donoho.

[8] EPA, National air quality and emissions trends report, Office of Air Quality Planning and Standards, EPA, Research Traingle Park, EPA 454/R-01-004, 2001.

[9] D. Guillamet, B. Schiele, and J. Vitri. Analyzing non-negative matrix factorization for image classification. In Proc. 16th Internat. Conf. Pattern Recognition (ICPR02), Vol. II, 116119. IEEE Computer Society, August 2002.

[10] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, New York, 2001.

[11] P. K. Hopke, Receptor Modeling in Environmental Chemistry, Wiley and Sons, New York, 1985.

[12] P. K. Hopke, Receptor Modeling for Air Quality Management, Elsevier, Amsterdam, Hetherlands, 1991.

[13] P. O. Hoyer, Nonnegative sparse coding, Neural Networks for Signal Processing XII, Proc. IEEE Workshop on Neural Networks for Signal Processing, Martigny, 2002.

[14] T. Kawamoto, K. Hotta, T. Mishima, J. Fujiki, M. Tanaka, and T. Kurita. Estimation of single tones from chord sounds using non-negative matrix factorization, Neural Network World, 3(2000), 429-436.

[15] E. Kim, P. K. Hopke, and E. S. Edgerton, Source identification of Atlanta aerosol by positive matrix factorization, J. Air Waste Manage. Assoc., 53(2003), 731-739.

[16] D. D. Lee and H. S. Seung, Learning the parts of objects by nonnegative matrix factorization, Nature, 401(1999), 788-791.

[17] D. D. Lee and H. S. Seung, Algorithms for nonnegative matrix factorization, in Advances in Neural Information Processing 13, MIT Press, 2001, 556-562.

[18] W. Liu and J. Yi, Existing and new algorithms for nonnegative matrix factorization, University of Texas at Austin, 2003, report, available at http://www.cs.utexas.edu/users/liuwg/383CProject/final_report.pdf.

[19] E. Lee, C. K. Chun, and P. Paatero, Application of positive matrix factorization in source apportionment of particulate pollutants, Atmos. Environ., 33(1999), 3201-3212.

[20] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. Neural Comput., 12:337365, 2000.

[21] 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, pp. 111126, 1994.

[22] P. Paatero and U. Tapper, Least squares formulation of robust nonnegative factor analysis, Chemomet. Intell. Lab. Systems, 37(1997), 23-35.

[23] H. Park, M. Jeon, and J. B. Rosen, Lower dimensional representation of text data in vector space based information retrieval, in Computational Information Retrieval, ed. M. Berry, Proc. Comput. Inform. Retrieval Conf., SIAM, 2001, 3-23.

[24] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons, Text mining using nonnegative matrix factorizations, In Proc. SIAM Inter. Conf. on Data Mining, Orlando, FL, April 2004.

[25] J. Piper, V. P. Pauca, R. J. Plemmons, and M. Giffin, Object characterization from spectral data using non-negative factorization and information theory. In Proc. Amos Technical Conf., Maui, HI, September 2004, see http://www.wfu.edu/~plemmons.

[26] R. J. Plemmons, M. Horvath, E. Leonhardt, V. P. Pauca, S. Prasad, S. Robinson, H. Setty, T. Torgersen, J. van der Gracht, E. Dowski, R. Narayanswamy, and P. Silveira, Computational imaging Systems for iris recognition, In Proc. SPIE 49^{th} Annual Meeting, Denver, CO, 5559(2004), 335-345.

[27] F. Shahnaz, M. Berry, P. Pauca, and R. Plemmons, Document clustering using nonnegative matrix factorization, to appear in the Journal on Information Processing and Management, 2005, see http://www.wfu.edu/~plemmons.

[28] J. Tropp, Literature survey: Nonnegative matrix factorization, University of Texas at Asutin, preprint, 2003.

[29] J. Tropp, Topics in Sparse Approximation, Ph.D. Dissertation, University of Texas at Austin, 2004.

[30] S. Wild, Seeding non-negative matrix factorization with the spherical k-means clustering, M.S. Thesis, University of Colorado, 2002.