Available online at www.sciencedirect.com
ScienceDirect
Linear Algebra and its Applications 429 (2008) 1020–1025
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

Non-negative matrix factorization with fixed row and column sums

Ngoc-Diep Ho, Paul Van Dooren*

CESAME, Université catholique de Louvain, Av. Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium

Received 27 October 2006; accepted 23 February 2007
Available online 24 March 2007
Submitted by H. Bart

Abstract

In this short note, we focus on the use of the generalized Kullback–Leibler (KL) divergence in the problem of non-negative matrix factorization (NMF). We will show that when using the generalized KL divergence as cost function for NMF, the row sums and the column sums of the original matrix are preserved in the approximation. We will use this special characteristic in several approximation problems.

Keywords: Non-negative matrix factorization; Generalized Kullback–Leibler divergence; Stochastic matrix approximation; Row sums; Column sums

1. Introduction

It is argued in [6] that non-negative data are quite naturally occurring in the human perception of signals such as light intensities or sound spectra, and that we then typically decompose the complete signal into simpler parts that are non-negative signals as well. One can rewrite this decomposition as a non-negative matrix factorization (NMF), which explains the popularity of such factorizations in the problem of approximating non-negative data in a sum of non-negative parts.

The general approximation problem of a $m \times n$ non-negative matrix A by a linear combination of k diadic $u_i v_i^{\mathrm{T}}$ products ($k < m, n$) reduces to the minimization of the error matrix

* Corresponding author.
E-mail addresses: ho@inma.ucl.ac.be (N.-D. Ho), vdooren@inma.ucl.ac.be (P. Van Dooren).

0024-3795/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2007.02.026

$$A - \sum_{i=1}^{k} \sigma_i u_i v_i^{\mathrm{T}},$$

where the non-negative elements σ_i are the weighting factors of the linear combination. The approximation error is usually determined by a cost function which depends on the application.

If there is no non-negativity constraint on the vectors u_i and v_i, one can obtain an optimal rank k approximation in the Frobenius norm by using the singular value decomposition (SVD) for which efficient algorithms are available [3].

When a non-negativity constraint is imposed, u_i and v_i are limited to *non-negative* vectors of appropriate length. In many applications, this is a crucial property that one wants to preserve. This makes the low-rank approximation problem non-convex and difficult, but one can still look for a local minimum of a particular cost function of the low-rank approximation. This can be obtained in polynomial time, and iterative algorithms for obtaining such local minima have been proposed in e.g. [6].

In general, adding more constraints often makes the approximation problem more difficult to solve. But in this paper, we will show that the constraint under which the row and column sums are preserved is automatically satisfied by using the generalized Kullback–Leibler divergence as the cost function for the NMF problem. In [2], a discussion is made leading to the preservation of the matrix sum when using the generalized Kullback–Leibler divergence in the NMF. But this is just a straightforward result from the preservation of row and column sums, which will be pointed out in this paper.

One can apply these results in approximating large-scale Markov chains. Typical applications of this are PageRank [10] used for ordering the web pages, Markov Clustering [12] used for clustering vertices in a graph and Hidden Markov Models [11] used for learning and predicting sequential data. The number of such applications is growing, but the size of the underlying problems is growing as well. The size of the stochastic matrices representing Markov chains related to the web is e.g. of the order of billions. In order to cope with such large scale stochastic matrices, one could use NMF to approximate a large stochastic matrix by a lower rank one.

2. NMF problem using generalized KL divergence

The *non-negative matrix factorization* (NMF) problem imposes non-negativity conditions on the factors (i.e. $A \approx \sum_{i=1}^{k} u_i v_i^{\mathrm{T}}, u_i, v_i \geqslant 0$) and can be stated as follows:

Given a non-negative $(m \times n)$ matrix A, find two non-negative matrices $U(m \times k)$ and $V(n \times k)$ with $k \ll m, n$ that minimize $F(A, UV^{\mathrm{T}})$, where $F(A, UV^{\mathrm{T}})$ is a cost function defining the “nearness” between matrices A and UV^{T}.

The choice of cost function F of course affects the solution of the minimization problem. In this paper, we will focus on the generalized *Kullback–Leibler* (KL) divergence that is defined as follows:

$$F(A, UV^{\mathrm{T}}) = D(A\|UV^{\mathrm{T}}) := \sum_{ij} A_{ij} \log \frac{A_{ij}}{[UV^{\mathrm{T}}]_{ij}} - A_{ij} + [UV^{\mathrm{T}}]_{ij}. \qquad (1)$$

The problem

$$\min_{U \geqslant 0, V \geqslant 0} \quad D(A\|UV^{\mathrm{T}}), \qquad (2)$$

where $A \geqslant 0$, is called non-negative matrix factorization using the generalized Kullback–Leibler divergence. For this divergence, the gradients are easy to construct (see [7]):

$$\nabla_{U_{ij}} D(A\|UV^{\mathrm{T}}) = -\sum_k \left(\frac{A_{ik}}{[UV^{\mathrm{T}}]_{ik}} V_{kj} - V_{kj} \right), \tag{3}$$

$$\nabla_{V_{ij}} D(A\|UV^{\mathrm{T}}) = -\sum_k \left(\frac{A_{ki}}{[UV^{\mathrm{T}}]_{ki}} U_{kj} - U_{kj} \right). \tag{4}$$

The Karush–Kuhn–Tucker (KKT) optimality conditions are then found to be (see [1])

$$U \geqslant 0, \quad V \geqslant 0, \tag{5}$$

$$\nabla_U D(A\|UV^{\mathrm{T}}) \geqslant 0, \quad \nabla_V D(A\|UV^{\mathrm{T}}) \geqslant 0, \tag{6}$$

$$U \circ \nabla_U D(A\|UV^{\mathrm{T}}) = 0, \quad V \circ \nabla_V D(A\|UV^{\mathrm{T}}) = 0, \tag{7}$$

where $A \circ B$ is the Hadamard product between two matrices having the same size (i.e. $[A \circ B]_{ij} = A_{ij}B_{ij}$).

It is important to note that the cost function $F(A, UV^{\mathrm{T}})$ is convex in each of the factors U and V, but it is not convex in the two factors at the same time, hence the problem can have many local minima. Some iterative search methods that converge to a stationary point of the above problem can be found in [6,2,8,9]. In the next section, we will investigate the stochasticity of the stationary points.

3. Stationary points

In this section, we use the optimality conditions (5) and (7) to show a particular property of the stationary points of NMF using the generalized KL divergence.

Theorem 1. *Let $A_{m\times n}$ a non-negative matrix. Then every stationary point (U, V) of the cost function in (2) preserves the column sums of A i.e. $(\mathbf{1}_{1\times m}A = \mathbf{1}_{1\times m}(UV^{\mathrm{T}}))$, the row sums of A i.e. $(A\mathbf{1}_{n\times 1} = (UV^{\mathrm{T}})\mathbf{1}_{n\times 1})$ and the matrix sum of A i.e. $(\mathbf{1}_{1\times m}A\mathbf{1}_{n\times 1} = \mathbf{1}_{1\times m}(UV^{\mathrm{T}})\mathbf{1}_{n\times 1})$, where $\mathbf{1}_{p\times l}$ is $p \times l$ matrix with all elements equal to* 1.

Proof. At a stationary point, from (7), the matrix V must satisfy the following optimality condition:

$$V_{ij} \sum_k \frac{A_{ki}}{[UV^{\mathrm{T}}]_{ki}} U_{kj} = V_{ij} \sum_k U_{kj} \quad \forall i, j.$$

Calculating the sum over j of the left-hand side matrix gives

$$\sum_j V_{ij} \sum_k \frac{A_{ki}}{[UV^{\mathrm{T}}]_{ki}} U_{kj} = \sum_k \left(\sum_j V_{ij} U_{kj} \right) \frac{A_{ki}}{[UV^{\mathrm{T}}]_{ki}} = \sum_k A_{ki}$$

and the sum over j of the right-hand side matrix gives

$$\sum_j V_{ij} \sum_k U_{kj} = \sum_k \sum_j V_{ij} U_{kj} = \sum_k [UV^{\mathrm{T}}]_{ki}.$$

This implies that $\sum_k A_{ki} = \sum_k [UV^{\mathrm{T}}]_{ki}$ or $\mathbf{1}_{1\times m}A = \mathbf{1}_{1\times m}(UV^{\mathrm{T}})$. For the row sums, one can easily prove the equality by the same development using the optimality condition of V. The matrix sum is preserved as a consequence of the preservation of column sums or row sums. □

Using the above theorem, one obtains the following *standard form* for every stationary point of the KL divergence iteration.

Corollary 2. *Let $A_{m\times n}$ be a non-negative matrix. Every stationary point $(U_{m\times k}, V_{n\times k})$ of the KL minimization problem has the form:*

$$UV^{\mathrm{T}} = P_{m\times k} D_{k\times k} Q^{\mathrm{T}}_{n\times k},$$

where P, Q are column stochastic, D is diagonal non-negative, and $\sum_i D_{ii} = \sum_{ij} A_{ij}$. Furthermore, if A is column stochastic (or row stochastic) then the matrix DQ^{T}(or PD) are also column stochastic (or row stochastic).

Proof. Define the normalization factors D_U and D_V as the column sums of U and V, respectively. Then there exist column stochastic matrices P and Q such that $PD_U = U$, $QD_V = V$. These matrices are obtained by dividing the respective columns by their non-zero column sums, and by choosing an arbitrary stochastic column in P or Q if the corresponding column sum in U or V was zero. Define $D = D_U D_V$, then $\sum_i D_{ii} = \sum_{ij} A_{ij}$ follows since P and Q are column stochastic. Moreover, PDQ^{T} is easily shown to preserve the matrix sum of A.

It is straightforward then that the column sums of DQ^{T} and row sums of PD are those of A, which also proves the last assertion for stochastic matrices. □

Furthermore, if there exists a zero column in U (or V), one can remove that zero column in U (or V) and the corresponding column in V (or U) without changing the product UV^{T}. This amounts to saying that one can also obtain a reduced rank factorization of the same type, in which the diagonal elements of D are restricted to be all strictly positive. By writing the stationary point in this form, one can compare this with the singular value decomposition (SVD) of a matrix $A = U\Sigma V^{\mathrm{T}}$ where the orthogonal matrices U and V are replaced by column stochastic matrices P and Q.

In particular, if the reduced rank k is 1, it follows from Theorem 1 that we can have a unique global minimizer:

$$\widehat{A} = \sigma u v^t, \tag{8}$$

where $\sigma = \sum_{i,j} A_{ij}$, $u_j = \sum_j A_{ij}/\sigma$ and $v_i = \sum_i A_{ij}/\sigma$. And if the rank k is equal to $\min(m, n)$ we have a trivial solution which is $(U = A, V = I_n)$ or $(U = I_m, V = A^{\mathrm{T}})$.

If we consider Problem (2) for a non-negative matrix with unit element sum (i.e. $\sum_{i,j} A_{ij} = 1$) then, the stationary points are in fact solutions of the Probabilistic Latent Semantic Analysis (pLSA) [5] which is used in document classification. The link between pLSA and Problem (2) was first pointed out in [4]. The pLSA problem is then to find a low-rank joint-probability matrix that approximates a full rank or higher rank joint-probability matrix $A/(\sum_{i,j} A_{ij})$ using the generalized Kullback–Leibler divergence.

4. Application: stochastic matrix approximation

If the input matrix A is stochastic, the stochastic matrix UV^{T} that minimizes the generalized KL divergence is called a low-rank stochastic approximation of A. Theorem 1 shows that if the input matrix A is column stochastic (or row stochastic or doubly stochastic), the stationary points UV^{T} are actually column stochastic (or row stochastic or doubly stochastic). In other words, the stochasticity of the original matrix is naturally preserved in the approximation.

Using the iterative algorithm in [6], one can numerically obtain a solution for the stochastic matrix approximation problem. Here are some examples of stationary points that are candidates for a solution:

- column stochastic matrix

$$A = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 \\ 0 & 1 & \frac{1}{2} \end{bmatrix} \approx PDQ^{\mathrm{T}} = \begin{bmatrix} 0 & \frac{2}{3} \\ 0 & \frac{1}{3} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 0 & \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & 0 & \frac{1}{3} \end{bmatrix}, \tag{9}$$

- row stochastic matrix

$$A = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{2}{3} & \frac{1}{3} & 0 \end{bmatrix} \approx PDQ^{\mathrm{T}} = \begin{bmatrix} 0 & \frac{3}{5} \\ \frac{3}{4} & 0 \\ \frac{1}{4} & \frac{2}{5} \end{bmatrix} \begin{bmatrix} \frac{4}{3} & 0 \\ 0 & \frac{5}{3} \end{bmatrix} \begin{bmatrix} 0 & \frac{5}{8} & \frac{3}{8} \\ \frac{7}{10} & 0 & \frac{3}{10} \end{bmatrix}, \tag{10}$$

- and doubly stochastic

$$A = \begin{bmatrix} \frac{3}{8} & \frac{1}{4} & \frac{3}{8} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{3}{8} & \frac{1}{4} & \frac{3}{8} \end{bmatrix} = PDQ^{\mathrm{T}} = PDP^{\mathrm{T}} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{bmatrix}. \tag{11}$$

In the above examples, the approximations are written in the form presented in Corollary 2. Especially, in the third example of a doubly stochastic matrix, we have an exact and symmetric factorization of the form PDP^{T}. In general, it is not easy to find such a symmetric approximation.

Furthermore, instead of considering the ordinary sum of a vector, we can consider the *weighted sum* of a vector x, defined as

$$s_w(x) = \sum_i w_i x_i = x^{\mathrm{T}} w, \tag{12}$$

where w is a positive weight vector. One can find an approximation that preserves the weighted column sums and row sums of the original matrix. In fact, suppose w_r and w_c are weight vectors with respect to which we want to find a low-rank approximation $\widetilde{U}\widetilde{V}^{\mathrm{T}}$ of A that preserves the weighted row sums and the weighted column sums respectively, i.e.

$$\widetilde{U}\widetilde{V}^{\mathrm{T}} w_r = A w_r, \quad w_c^{\mathrm{T}} \widetilde{U}\widetilde{V}^{\mathrm{T}} = w_c^{\mathrm{T}} A, \tag{13}$$

we can use the following procedure:

(1) create $\widehat{A} = D_{w_r} A D_{w_c}$, where D_a is the diagonal matrix having the vector a on the main diagonal,
(2) find a low rank non-negative approximation UV^{T} of $\widehat{A}$ by using NMF algorithm for the generalized KL divergence,
(3) and create the desired approximation $\widetilde{U}\widetilde{V}^{\mathrm{T}}$ using $\widetilde{U} = D_{w_c}^{-1} U$ and $\widetilde{V} = D_{w_r}^{-1} V$.

Applying Theorem 1, one can easily check that (13) does hold for the newly created matrix $\widetilde{U}\widetilde{V}^{\mathrm{T}}$.

The preservation of weighted column sums and row sums implies that we can use the same procedure to construct a low rank non-negative approximation that preserve the left and right principal eigenvectors of a square non-negative matrix. The trick is then simply to use the left and right principal eigenvectors of the original matrix as the weight vectors described above to construct the desired approximation.

Acknowledgements

This paper presents research supported by the Concerted Research Action (ARC) "Large Graphs and Network" of the French Community of Belgium, and by the Belgian Programme on Inter-university Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. The scientific responsibility rests with the authors. Ngoc-Diep Ho is a FRIA fellow.

References

[1] M. Catral, L. Han, M. Neumann, R. Plemmons, On reduced rank nonnegative matrix factorizations for symmetric matrices, Linear Algebra Appl. 393 (2004) 107–126.

[2] L. Finesso, P. Spreij, Nonnegative matrix factorization and I—Divergence alternating minimization, Linear Algebra Appl. 416 (2006) 270–287.

[3] G. Golub, C. Van Loan, Matrix Computations, third ed., The Johns Hopkins University Press, Baltimore, 1996.

[4] E. Gaussier, C. Goutte, Relation between PLSA and NMF and implications, in: Proceedings of the 28th Annual International ACM SIGIR, Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15–19, 2005.

[5] T. Hofmann, Probabilistic latent semantic analysis, in: Proceedings of the 15th Conference on Uncertainty in AI, Morgan Kaufman Publishers, 1999, pp. 289–296.

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

[7] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Adv. Neural Inf. Process. 13 (2001) 556–562.

[8] C.J. Lin, On the convergence of multiplicative update algorithms for non-negative matrix factorization, Technical Report, Department of Computer Science, National Taiwan University, 2005.

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

[10] S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine, in: Proceedings of the Seventh International World Wide Web Conference, Elsevier, 1998, pp. 107–117.

[11] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–285.

[12] S. van Dongen, Graph clustering by flow simulation, PhD thesis, University of Utrecht, May 2000.