Literature Survey:
Non-Negative Matrix Factorization

Joel A. Tropp

Institute for Computational Engineering and Sciences, 1 University Station, C0200, The University of Texas at Austin, Austin, TX 78712

E-mail address: `jtropp@ticam.utexas.edu`

Abstract. This article surveys recent research on Non-Negative Matrix Factorization (NNMF), a relatively new technique for dimensionality reduction. It is based on the idea that, in many data-processing tasks, negative numbers are physically meaningless. The NNMF technique addresses this problem by placing non-negativity constraints on the data model. I discuss the applications of NNMF, the proposed algorithms and the qualitative results. Since many of the algorithms suggested for NNMF seem to lack a firm theoretical foundation, this article also surveys techniques for proving that iterative algorithms converge. It concludes with a description of additional investigations which are presently underway.

1. Introduction

A basic technique in dimensionality reduction is Principal Component Analysis (PCA), which calculates a set basis vectors that can be used to approximate high-dimensional data optimally in the least-squares sense. The number of basis vectors is much smaller than the number of dimensions, so encoding the data as linear combinations of the basis vectors transforms it to a lower-dimensional space. This reduction can be used to improve the tractability of data analysis algorithms, to discover features in the data or to find hidden variables. See Jolliffe [**1**] for a survey.

One major problem with PCA is that the basis vectors have both positive and negative components, and the data are represented as linear combinations of these vectors with positive and negative coefficients. The optimality of PCA can be traced back to constructive cancelation of the signs. In many applications, the negative components contradict physical realities. For example, the pixels in a grayscale image have non-negative intensities, so an image with negative intensities cannot be reasonably interpreted.

To address this philosophical problem, several researchers suggested that the search for a representative basis should be confined to non-negative vectors. Formally, this idea can be interpreted as decomposing a non-negative matrix A into two non-negative factors V and H, i.e.

$$\min_{V\geq 0, H\geq 0} \|A - VH\|_{\mathrm{F}}^2 . \tag{1.1}$$

Hence the paradigm is called Non-Negative Matrix Factorization (NNMF). We assume that V has far fewer columns than A has rows, so the approximation can succeed only if it discovers latent structure in the data. Although NNMF is computationally intensive, the last ten years have seen several efforts at implementing it for various purposes.

2. Non-Negative Matrix Factorization

In this section, we shall summarize the development of NNMF in several different research communities. This survey is by no means comprehensive, but it addresses the most important work.

2.1. Work of Paatero. The idea of non-negative matrix factorization can be traced back to a 1994 paper of Paatero and Tapper [**2**]. Their goal was to perform factor analysis on environmental data, which problem involves finding a small number of root causes that explain a large set of measurements. Each factor is a positive combination of some elementary variables. In real circumstances, a factor is present (in which case it has a certain positive effect) or it is absent (in which case it has no effect). Therefore, it often makes sense to constrain the factors and their influences to be non-negative.

The problem can be posed formally. Suppose that the columns of A are the measurements, the columns of V are the factors and the rows of H are the influences of each factor (called scores). Use W to denote the weight associated to each element, which indicates the level of confidence in that measurement. Paatero and Tapper advocate optimizing the functional

$$\|W \cdot (A - VH)\|_{\mathrm{F}}^2 \qquad \text{subject to} \qquad V \geq 0 \text{ and } H \geq 0.$$

Here, $\cdot$ denotes the Hadamard (also known as the Schur or elementwise) product, and the inequalities are also elementwise.

Paatero and Tapper originally proposed using a constrained alternating least squares algorithm (ALS) to solve the problem [**2**]. This method fixes V and solves the optimization with respect to H. Then it reverses the roles of the variables and repeats the process *ad infinitum*. The algorithm is initialized with different random matrices in an effort to obtain a global optimum. de Leeuw argues that ALS algorithms converge since they steadily decrease the objective function [**3**], but this convergence is not in the usual mathematical sense.

Paatero subsequently invented several other algorithms for attacking the optimization. His second algorithm, PMF2, may be viewed as a version of the alternating method with a plethora of ad hoc modifications, which complicate the implementation enormously [**4**]. Later, Paatero presented a more general method, called the Multilinear Engine, for finding multi-factor models with non-negativity constraints [**5**]. In these models, the approximant VH is replaced by a longer product of matrices. The program uses a modified conjugate gradient algorithm to solve the optimization problem.

By now, Paatero's community has generated an extensive body of work on non-negative factor analysis, which is impossible to review here. These papers share several shortcomings from the present point of view. First, they concentrate on a specific application of NNMF, so it is not clear how relevant they are to the wider body of applications. Second, the algorithms they use are frequently Byzantine, which precludes straightforward modification for other domains. Third, they do not seem to prove that the algorithms converge or to address the properties of the solutions which their algorithms yield. Indeed, it appears more or less impossible to ascertain the theoretical complexity or convergence of their methods. They make claims based primarily on empirical evidence.

2.2. Conic and Convex Coding. Independent of Paatero, Lee and Seung introduced the concept of NNMF in a 1997 paper on unsupervised learning [**6**]. They begin by considering the following encoding problem. Suppose that the columns of V are fixed feature vectors and that $\boldsymbol{a}$ is an input vector to be encoded. The goal is to minimize the reconstruction error

$$\min_{\boldsymbol{h}} \|\boldsymbol{a} - V\boldsymbol{h}\|_2^2 .$$

Different learning techniques can be obtained from various constraints on the vector $\boldsymbol{h}$. PCA corresponds to an unconstrained minimization, while Vector Quantization (VQ) requires that $\boldsymbol{h}$ equal one of the canonical basis vectors (i.e. a single unit component with the remaining entries zero). Lee and Seung propose two techniques that compromise between PCA and VQ. The first, convex coding, requires the entries of $\boldsymbol{h}$ to be nonnegative numbers which sum to one. So the encoded vector is the best approximation to the input from the convex hull of the feature vectors. The second, conic coding, requires only that the entries of $\boldsymbol{h}$ be nonnegative. Then the encoded vector is the best approximation to the input from the cone generated by the feature vectors.

Next, Lee and Seung consider how to find the best set of feature vectors for their new coding strategies. This leads them to the matrix approximation problem (1.1), where the columns of A contain training examples and V has far fewer columns that A. For both convex and conic coding, they require V and H to be nonnegative. In addition, for convex coding, they force the column sums of V and the row sums of H to equal one.

To solve their minimization problems, they suggest an alternating projected gradient method. In other words, fix V; perform a step of gradient descent with respect to H; then zero all the negative components of H. Reverse the roles of the variables and repeat. For the convex coding problem, they also introduce a penalty function into the minimization to maintain the row and column sums. The algorithms are executed multiple times with random starting points in an effort to locate a global optimum.

Using these algorithms, they found that convex coding discovers locally linear models of the data, while conic coding discovers features in the data. Their paper did not provide any proof of convergence, nor did it consider other types of algorithms which might apply.

2.3. Learning Parts of Objects. Lee and Seung subsequently developed simpler algorithms for computing the factorization [**7**]. One algorithm attempts to minimize the Frobenius norm of the residual, while the second attempts to minimize a modified Kullback-Liebler divergence. Both are based on multiplicative update rules, which amount to optimally-scaled gradient descents. As usual, the procedures are applied multiple times with different starting points. Lee and Seung provide proofs that both multiplicative update rules converge to local minima of the respective objective functions, and they claim that these rules are significantly faster than the algorithms of [**6**]. Nevertheless, each optimization requires several hours of computer time on a Pentium II. Ease of implementation is the only clear advantage over the other algorithms described here.

Using these new algorithms, Lee and Seung performed more extensive experiments [**8**]. When the columns of the input matrix were images of faces, NNMF produced basis vectors which correspond to facial features, such as eyes, ears and mustaches. When the columns of the input matrix were word counts from documents, NNMF produced basis vectors which correspond to semantic categories. Moreover, the factorization was able to distinguish separate meanings of homonyms by placing the same word in multiple categories. In both cases, the feature vectors were highly sparse.

3. Behavior of Iterative Algorithms

Many of the iterative techniques we have discussed lack convergence proofs, or they converge only in a weak sense. I have been interested in demonstrating when alternating minimization algorithms converge in the usual mathematical sense. To that end, I have also surveyed the literature on convergence of optimization algorithms.

3.1. Majorization Algorithms. A function $G(h, h')$ is called an *auxiliary function* for $F(h)$ if $G(h, h) = F(h)$ and $G(h, h') \geq F(h)$. These functions are useful on account of the following lemma.

Lemma 3.1. *If G is an auxiliary function for F, then F is non-increasing under the update $h_{k+1} = \arg\min_h G(h, h_k)$.*

Moreover, the sequence formed by iterative application of this update rule converges to a local minimum of F.

The method, such as it is, involves choosing an objective function and finding an auxiliary function for that objective function. The associated majorization algorithm involves repeated minimization of the auxiliary function, as described in the lemma. This technique is used most famously to prove the convergence of the EM algorithm [**9**]; Lee and Seung use it as well [**7**]. The main problem is that finding appropriate auxiliary functions is an art, and it may not be possible to frame an arbitrary algorithm as a majorization algorithm.

3.2. Algorithmic Mappings. A very general approach was developed by Zangwill for analyzing mathematical programming techniques [**10**]. It is based on the concept of a point-to-set mapping, a function that maps a single point to a collection of points. An *algorithm* is an iterative process associated with an collection of point-to-set maps $\{A_k\}$. Given an initial point z_1, the algorithm generates a sequence of points which satisfy $z_{k+1} \in A_k(z_k)$. In words, any point in $A_k(z_k)$ is an appropriate successor to z_k.

Under mild conditions, Zangwill shows that an algorithm whose iterates systematically decrease an auxiliary function will converge in a weak sense. Specifically, any convergent subsequence of iterates converges to a fixed point of the algorithm. This does not imply that the entire sequence of iterates converges in norm, the usual mathematical sense of the word.

Subsequently, Zangwill's work has been extended. Under additional conditions, it can be shown that the sequence of iterates either converges or has a continuum of accumulation points [**3**]. This result is much better, but it still falls short of a full convergence proof.

3.3. Alternating Projections. A surprising number of algorithms can be cast as alternating minimizations. Given a function of two variables, such as a metric, these procedures perform a constrained minimization of that function with respect to one variable while holding the other variable fixed. Then the minimization is performed again with the roles of the variables reversed, etc. These techniques are frequently used when the one-variable optimization problems have closed-form solutions or are otherwise tractable.

The first alternating minimization was introduced by von Neumann in 1933 as a technique for finding the projection onto the intersection (or direct sum) of two subspaces of a Hilbert space. He shows that projecting a point onto each subspace in turn would yield, in the limit, the projection onto the intersection [**11**]. Independently, Diliberto and Straus introduced an identical algorithm in 1951 for approximating bivariate continuous functions by a sum of two univariate continuous functions [**12**]. Cheney and Light present an attractive treatment of alternating projections in the Banach space setting [**13**]. The Hilbert space theory has also been generalized to encompass algorithms which cyclicly project a point onto a finite collection of subspaces. These algorithms generally yield the projection onto the intersection of subspaces, but the rate of convergence is sometimes abysmal. The most recent work on the subject appears in [**14**].

Another direction of generalization is from subspaces to convex sets. Cheney and Goldstein proved than an alternating minimization technique can find a pair of minimally distant points from two closed convex sets of a Hilbert space under mild conditions on the sets [**15**]. Independently, Dykstra and Han invented a corrected alternating minimization algorithm which can be used for arbitrary closed convex sets [**16, 17**]. An analysis of its convergence appears in [**18**]. Csiszár and Tusnády offer results in more general metric settings [**19**]. They show that an alternating minimization produces a global optimum whenever the objective function and the constraint sets have certain geometric properties. They apply their theory to the problem of finding a minimally divergent pair of probability distributions under the Kullback-Leibler divergence [**19**]. For an extensive discussion on using cyclic projection algorithms for finding a point in the intersection of convex sets, see [**20**].

Occasionally, alternating and cyclic minimizations are used in non-convex situations, in which cases it is usually impossible to provide strong guarantees on the sequence of iterates [**21, 22**]. For example, Chu has proposed an alternating minimization algorithm for producing a Hermitian matrix with given eigenvalues and diagonal entries, subject to the majorization condition [**21**]. (Chu claims that his algorithm converges on the grounds that it is a descent algorithm; this argument is specious.) Another specific example is a technique for computing Welch-Bound-Equality sequences [**22**]. More generally, Cadzow has discussed a class of algorithms that he calls composite property mappings [**23**]. Given a finite collection of general sets in a metric space and an initial point, a composite property mapping projects onto each set in turn in the hope of producing a point in the intersection nearest to the initial point. Cadzow uses Zangwill's theory to provide a partial convergence proof. He remarks that his algorithms have never failed to converge in the usual sense.

Most recently, I have developed some methods which can be used to prove that descent algorithms and alternating minimizations converge in the usual sense [**24**]. The conditions are more demanding than Zangwill's, but the conclusions are commensurately stronger.

4. Intended Contributions

My research group is presently making a comparison of different NNMF algorithms for some traditional data-mining tasks. The tasks include dimensionality reduction for images of faces, images of digits and document collections. The algorithms include Lee and Seung's multiplicative update rules and alternating least squares. We are interested in the quality of the approximations produced rather than the time complexity of the techniques. My role in this project is to provide strong convergence proofs for the algorithms we are considering where these proofs are lacking [**25**].

Bibliography

[1] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., ser. Springer series in statistics. New York: Springer-Verlag, 2002.

[2] 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. 111–126, 1994.

[3] J. de Leeuw, *Information Systems and Data Analysis*. Springer, 1994, ch. Block-relaxation algorithms in statistics.

[4] P. Paatero, "Least-squares formulation of robust non-negative factor analysis," *Chemometrics and Intell. Lab. Sys.*, vol. 37, pp. 23–35, 1997.

[5] ——, "The Multilinear Engine—a table-driven least squares program for solving multilinear problems, including the n-way parallel factor analysis model," *J. Comput. and Graph. Stat.*, vol. 8, no. 4, pp. 854–888, 1999.

[6] D. D. Lee and H. S. Seung, "Unsupervised learning by convex and conic coding," in *Advances in Neural Information Processing Systems*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds., vol. 9. The MIT Press, 1997, pp. 515–521. [Online]. Available: citeseer.nj.nec.com/lee97unsupervised.html

[7] ——, "Algorithms for Non-Negative Matrix Factorization," in *Advances in Neural Information Processing*, 2000.

[8] ——, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, October 1999.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[10] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*. Englewood Cliffs: Prentice-Hall, 1969.

[11] J. von Neumann, *Functional Operators, Vol. II*, ser. Annals of Mathematics Studies. Princeton Univ. Press, 1950, no. 22.

[12] S. P. Diliberto and E. G. Straus, "On the approximation of a function of several variables by the sum of functions of fewer variables," *Pacific J. Math.*, vol. 1, pp. 195–210, 1951.

[13] E. W. Cheney and W. A. Light, *Approximation Theory in Tensor Product Spaces*, ser. Lecture Notes in Mathematics. Berlin: Springer-Verlag, 1985, no. 1169.

[14] J. Xu and L. Zikatanov, "The Method of Alternating Projections and the Method of Subspace Corrections in Hilbert space," *J. Amer. Math. Soc.*, vol. 15, no. 3, pp. 573–597, 2002.

[15] E. W. Cheney and A. A. Goldstein, "Proximity maps for convex sets," *Proc. Amer. Math. Soc.*, vol. 10, no. 3, pp. 448–450, 1959.

[16] R. L. Dykstra, "An algorithm for restricted least squares regression," *J. Amer. Stat. Soc.*, vol. 78, pp. 837–842, 1983.

[17] S.-P. Han, "A successive projection method," *Math. Prog.*, vol. 40, pp. 1–14, 1988.

[18] H. H. Bauschke and J. M. Borwein, "Dykstra's alternating projection algorithm for two sets," *J. Approx. Th.*, vol. 79, pp. 418–443, 1994.

[19] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Stat. and Decisions*, vol. Supplemental Issue 1, pp. 205–237, 1984.

[20] H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Rev.*, vol. 38, no. 3, pp. 367–426, 1996.

[21] M. T. Chu, "Constructing a Hermitian matrix from its diagonal entries and eigenvalues," *SIAM J. Matrix Anal.*, vol. 16, no. 1, pp. 207–217, 1995.

[22] J. A. Tropp, R. Heath, and T. Strohmer, "An alternating projection algorithm for constructing quasi-orthogonal CDMA signature sequences," 2003, accepted to ISIT 2003.

[23] J. A. Cadzow, "Signal enhancement—a composite property mapping algorithm," *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. 36, no. 1, pp. 49–62, Jan. 1988.

[24] J. A. Tropp, "Descent algorithms and stationary points," 2003, in preparation.

[25] I. S. Dhillon, S. Sra, and J. A. Tropp, "Private communications," 2003.