Document Classification Using Nonnegative Matrix Factorization and Underapproximation

Michael W. Berry
Dept. of Electrical Engineering and Computer Science
University of Tennessee
203 Claxton Complex
Knoxville, TN 37996-3450
Email: mberry@utk.edu

Nicolas Gillis and François Glineur
Center for Operations Research and Econometrics
Université catholique de Louvain
Voie du Roman Pays, 34
B-1348, Louvain-La-Neuve, Belgium
Email: {nicolas.gillis,francois.glineur}@uclouvain.be

***Abstract*— In this study, we use nonnegative matrix factorization (NMF) and nonnegative matrix underapproximation (NMU) approaches to generate feature vectors that can be used to cluster Aviation Safety Reporting System (ASRS) documents obtained from the Distributed National ASAP Archive (DNAA). By preserving nonnegativity, both the NMF and NMU facilitate a sum-of-parts representation of the underlying term usage patterns in the ASRS document collection. Both the training and test sets of ASRS documents are parsed and then factored by both algorithms to produce a reduced-rank representations of the entire document space. The resulting *feature* and *coefficient* matrix factors are used to cluster ASRS documents so that the (known) associated anomalies of training documents are directly mapped to the feature vectors. Dominant features of test documents are then used to generate anomaly relevance scores for those documents. We demonstrate that the approximate solution obtained by NMU using Lagrangrian duality can lead to a better sum-of-parts representation and document classification accuracy.**

I. Introduction

Nonnegative matrix factorization (NMF) has been widely used to approximate high dimensional nonnegative data sets. Lee and Seung [1] demonstrated how NMF techniques can be used to generate basis functions for image data that could facilitate the identification and classification of objects. They also showed how to use NMF for extracting concepts/topics from unstructured text documents. In this study, the so-called *sum-of-parts* representation offered by the NMF and related factorizations is exploited for the classification of documents from the Aviation Safety Reporting System (ASRS) collection.

Although many manuscripts have cited [1], NMF was first introduced by Paatero and Tapper [2]. The NMF problem can be simply stated as follows:

Given a nonnegative matrix $\mathbf{A} \in \Re^{\mathbf{m}\times\mathbf{n}}$ and a positive integer $k < \min\{m,n\}$, find nonnegative matrices $\mathbf{W} \in \Re^{\mathbf{m}\times\mathbf{k}}$ and $\mathbf{H} \in \Re^{\mathbf{k}\times\mathbf{n}}$ to minimize the functional

$$f(\mathbf{W},\mathbf{H}) = \frac{\mathbf{1}}{\mathbf{2}}\|\mathbf{A}-\mathbf{WH}\|_{\mathbf{F}}^{\mathbf{2}}. \quad (1)$$

The product $\mathbf{WH}$ is called a nonnegative matrix factorization of $\mathbf{A}$, although $\mathbf{A}$ is not necessarily *equal* to the product $\mathbf{WH}$. Hence, the product $\mathbf{WH}$ is an approximate factorization of rank at most k. The optimal choice for the rank k is problem dependent and in most cases chosen such that $k \ll \min(m,n)$. Alternatively, the product $\mathbf{WH}$ can be considered a *compressed* form of the data in $\mathbf{A}$.

A key characteristic or property of NMF is the ability of numerical methods that minimize expression (1) to generate underlying features as basis vectors in $\mathbf{W}$ that can used for object identification and classification. Without any negative components in $\mathbf{W}$ and $\mathbf{H}$, the NMF enables a non-subtractive combination of parts to form a whole [1]. Features may be parts of faces in image data, topics or clusters in textual data, or specific absorption characteristics in hyperspectral data [3]. In this paper, we discuss an extension of the classic NMF problem for the primary goal of improving feature extraction and identification in text/document mining.

Important challenges in the numerical minimization of expression (1) include the existence of local minima due to the non-convexity of $f(\mathbf{W},\mathbf{H})$ in both $\mathbf{W}$ and $\mathbf{H}$, and the non-uniqueness of its solution. Clearly any invertible matrix $\mathbf{D}$ such that $\mathbf{WD} \geq \mathbf{0}$ and $\mathbf{D^{-1}H} \geq \mathbf{0}$ generates an equivalent solution (which is the case, for example, when D has exactly one positive entry by row and by column). In practice, the NMF has been shown to be quite useful for text/data mining even with local minima. The resulting features (basis vectors) provide desirable data compression and classification capabilities.

Alternative formulations of the NMF problem have certainly been documented [3]. For example, an information theoretic formulation in [4] is based on the Kullback-Leibler divergence of $\mathbf{A}$ from $\mathbf{WH}$ and the cost functions proposed in [5] are based on Csiszár's φ-divergence. The formulation in [6] enforces constraints based on the Fisher linear discriminant analysis and [7] uses a diagonal weight matrix $\mathbf{Q}$ in the factorization model, $\mathbf{AQ} \approx \mathbf{WHQ}$, as an attempt to compensate for feature redundancy. See [8] and [9] for other approaches using cost functions.

To speed up convergence of Lee and Seung's (standard) NMF iteration, various alternative minimization strategies for expression (1) have been suggested. In [10], the use of a projected gradient bound-constrained optimization method was shown to have better convergence properties than the standard multiplicative update rule approach. However, the use of

978-1-4244-3828-0/09/$25.00 ©2009 IEEE

certain auxiliary constraints in expression (1) may break down the bound-constrained optimization assumption and thereby limit the use of projected gradient methods. Acceleration using an interior-point gradient method has been suggested in [11], and a quasi-Newton optimization approach for updating $\mathbf{W}$ and $\mathbf{H}$, where negative values are replaced with small positive ϵ parameter to enforce nonnegativity, is discussed in [12]. Another technique, simple and yet efficient, with nice convergence properties and based on a coordinate-descent approach, was introduced in [13] and studied in details in [14] (see also [15]). Finally, an overview of enhancements to improve the convergence of the (standard) NMF algorithm is available in [3].

Typically, $\mathbf{W}$ and $\mathbf{H}$ are initialized with random nonnegative values to start the standard NMF algorithm. Another area of NMF-related research has focused on alternate approaches for initializing or seeding the algorithm. The goal, of course, is to speed up convergence. In [16] spherical k-means clustering is used to initialize $\mathbf{W}$ and in [17] singular vectors of $\mathbf{A}$ are used for initialization and subsequent cost function reduction.

II. NMF Algorithm

As surveyed in [3], there are three general classes of NMF algorithms: multiplicative update algorithms, gradient descent algorithms, and alternating least squares algorithms. For this study, we improve upon the most basic multiplicative update method (first analyzed in [4]). This approach, based on a mean squared error objective function, is illustrated below using MATLAB®array operator notation:

MULTIPLICATIVE UPDATE ALGORITHM FOR NMF

```
W = rand(m,k);     % W initially random
H = rand(k,n);     % H initially random
for i = 1 : maxiter
      H = H .* (W'A) ./ (W'WH + ε);
      W = W .* (AH') ./ (WHH' + ε);
end
```

The parameter $\epsilon = 10^{-9}$ is added to avoid division by zero. If this multiplicative update NMF algorithm does converges to a stationary point, there is no guarantee that the stationary point is a local minimum for the objective function [3]. If the limit point to which the algorithm has converged lies on the boundary of the feasible region, one cannot conclude that it is, in fact, a stationary point. Modifications of the Lee and Seung multiplicative update scheme that resolves some of the convergence issues and guarantees convergence to a stationary point are provided in [18], [15]. Vavasis [19] has shown that NMF is NP-hard (see also [15]). In a recent work [20], Gillis and Glineur consider an NMF-like approximation problem (also NP-hard) whose solutions can be generated using a Lagrangian relaxation technique. This new approximation problem, referred to as nonnegative matrix underapproximation (or NMU, first introduced in [21]) is outlined in the next section and compared with NMF for text classification tasks in Section V.

III. NMU Algorithm

As discussed in [20] for a rank-1 NMF approximation in expression (1), the first (dominant) singular triplets of the matrix $\mathbf{A}$ provide optimal solutions for the nonnegative (vector) factors $\mathbf{W}$ and $\mathbf{H}$. Hence, an alternative approach to produce a rank-k NMF would be to recursively generate optimal rank-1 approximations (W_k,H_k), successively subtracting each factor W_kH_k from $\mathbf{A}$ before determining a new rank-1 approximation for the remainder $\mathbf{A} - \mathbf{W_kH_k}$. One can immediately see that such an approach would generate iterates (W_k,H_k) containing negative values. Therefore, [20] adds an upper bound constraint $\mathbf{WH} \leq \mathbf{A}$ to the NMF, to obtain a problem referred to as a nonnegative matrix underapproximation (or NMU). For example, if ($\mathbf{W_1}$,$\mathbf{H_1}$) is a rank-1 underapproximation to $\mathbf{A}$, that is, $\mathbf{W_1H_1} \lneq \mathbf{A}$, then $\mathbf{R_1} = \mathbf{A} - \mathbf{W_1H_1}$ is nonnegative. $\mathbf{R_1}$ can then be underapproximated by $\mathbf{W_2H_2} \lneq \mathbf{R_2}$, if $\mathbf{R_2} = \mathbf{R_1} - \mathbf{W_2H_2}$, and so on. After k steps, a rank-k underapproximation of A is given by

$$\mathbf{W_1H_1} + \mathbf{W_2H_2} + \cdots + \mathbf{W_kH_k} = [\mathbf{W_1}, \mathbf{W_2}, \ldots, \mathbf{W_k}]\,[\mathbf{H_1}, \mathbf{H_2}, \ldots, \mathbf{H_k}] = \mathbf{WH} \lneq \mathbf{A}.$$

The formal NMU optimization problem is given by

$$\begin{array}{l} \min\limits_{\mathbf{W},\mathbf{H}} \|\mathbf{A} - \mathbf{WH}\|_\mathbf{F}^\mathbf{2}, \text{ where} \\ \mathbf{WH} \leq \mathbf{A} \text{ and } \mathbf{W}, \mathbf{H} \geq \mathbf{0}. \end{array} \tag{2}$$

Using a Lagrangian relaxation approach, Gillis and Glineur [20] have shown that NMU iterates $\mathbf{W}$ and $\mathbf{H}$ can be obtained by minimizing $\|(\mathbf{A} - \mathbf{\Lambda}) - \mathbf{WH}\|_\mathbf{F}^\mathbf{2}$, where $\mathbf{\Lambda} = [\mathbf{\Lambda_{ij}}]$ contains the appropriate nonnegative Lagrange multipliers. This problem is the same as NMF except that the matrix to factorize $(\mathbf{M} - \mathbf{\Lambda})$ is not necessarily nonnegative; it is studied in [15] where the multiplicative updates are generalized as follows

$$\begin{aligned} \mathbf{H} &= \mathbf{H}.*(\mathbf{W^TA})./(\mathbf{W^TWH} + \mathbf{W^T\Lambda} + \epsilon); \\ \mathbf{W} &= \mathbf{W}.*(\mathbf{AH^T})./(\mathbf{WHH^T} + \mathbf{\Lambda H^T} + \epsilon). \end{aligned}$$

Since optimal Lagrange multipliers should satisfy a complementary constraint $\mathbf{\Lambda_{ij}}\,(\mathbf{A} - \mathbf{WH})_\mathbf{ij} = \mathbf{0},\ \forall \mathrm{i}, \mathrm{j}$, the following update rule for Λ between NMU iterates is suggested in [20]

$$\mathbf{\Lambda} = \max\{\mathbf{0}, \mathbf{\Lambda} - \mu_\mathbf{k}(\mathbf{A} - \mathbf{WH})\},\ \mu_\mathbf{k} \to \mathbf{0}, \tag{3}$$

where $\mu_k = \rho^k \mu_0$ for $\rho < 1$ and appropriate initial μ_0.

IV. Document Parsing and Term Weighting

The General Text Parsing (GTP) software environment [22] (written in C++) was used to parse all the Aviation Safety Reporting System (ASRS) documents for this preliminary study. If $\mathbf{A} = [\mathbf{R}|\mathbf{T}] = [\mathbf{a_{ij}}]$ defines the $m \times n$ term-by-document matrix for factorization, then the submatrices $\mathbf{R}$ and $\mathbf{T}$ represent training and test documents, respectively. Each element or component a_{ij} of the matrix A defines a *weighted* frequency at which term i occurs in document j. We define $a_{ij} = l_{ij}g_i$, where l_{ij} is the local weight for term i occurring in document j and g_i is the global weight for

term i in the subcollection. Let f_{ij} be the number of times (frequency) that term i appears in document j, and define $\hat{p}_{ij} = f_{ij}/\sum_j f_{ij}$, i.e., the empirical probability of term i appearing in document j. Using GTP, we deploy a common log-entropy term weighting scheme whereby

$$l_{ij} = \log(1 + f_{ij}) \text{ and } g_i = 1 + (\sum_j \hat{p}_{ij} log(\hat{p}_{ij}))/\log n) \ .$$

By default, GTP requires that the global frequency of any term, i.e., $\sum_{j=1}^{n} f_{ij}$, be greater than 1 and that a term's document frequency (or number of documents containing that term) be greater than 1 as well. No adjustments to these thresholds were made in parsing the ASRS documents. A *stoplist* of 493 words[1] was used by GTP to filter out unimportant terms.

Initial testing of NMF with ASRS documents with NMF used as many as $n = 21,519$ documents (see [23]). In comparing the classification performance of NMF and NMU, we use only the first $n = 100$ documents for this study. GTP extracted exactly $m = 733$ terms from these documents and all results were obtained using MATLAB® Version 7.7.

V. NMF/NMU Classification

The classification of ASRS documents using NMF and NMU follows the strategy first discussed in [23]. Let $\mathbf{H_i}$ represent the i-th column of matrix $\mathbf{H}$ and define α, as a the threshold on the relevance score or (target value) t_{ij} for document i and anomaly/label j. Let δ, be a lower bound (threshold) on the column elements of $\mathbf{H}$ such that all accepted (non-filtered) elements in $\mathbf{H_i}$ are greater than or equal to $(1 - \delta) \times \max(\mathbf{H_i})$. This threshold will filter out the association of features with both training ($\mathbf{R}$) and test ($\mathbf{T}$) documents. Let σ denote the percentage of documents used to define the training set (or number of columns of $\mathbf{R}$). Table I briefly summarizes the steps needed (see [23] for more details) to classify ASRS documents using matrix factors $(\mathbf{W}, \mathbf{H})$ generated by NMF or NMU. For all NMU-based classifications, the choices $\mu_0 = 2$ and $\rho = 0.35$ (see Equation (3)) yielded the best results.

A. Testing Methodology

The rank or number of columns of the feature matrix factor $\mathbf{W}$ used to test our NMF and NMU models was $k = 10$. Hence, the $\mathbf{W}$ and $\mathbf{H}$ matrix factors were 773×10 and 10×100, respectively. The percentage of ASRS documents used for training (subset $\mathbf{R}$) in our testing was 70% (i.e., $\sigma = .70$). Hence, a random selection of 70 documents was used as the training set ($\mathbf{R}$) and the remaining 30 documents were used for testing ($\mathbf{T}$) our classifiers. In Step 1 of Table I we chose $\delta = .30$ for the columnwise pruning of the elements in the coefficient matrix $\mathbf{H}$. This parameter effectively determines the number of features (among the $k = 10$ possible) that any document (training or test) can be associated with. As δ decreases, so does the sparsity of $\mathbf{H}$ [3].

The α parameter mentioned above is the prediction control parameter that ultimately determines whether or not document i will be given label (anomaly) j. We note that the initial matrix factors $\mathbf{W}$ and $\mathbf{H}$ (for NMF and NMU) are randomly generated and will produce slightly different features (columns of $\mathbf{W}$) and coefficients (columns of $\mathbf{H}$) per iteration[2]. After 5 iterations of the NMU multiplicative update rules mentioned in Section III, the residual ($\|\mathbf{A} - \mathbf{WH}\|_\mathbf{F}$ from Equation (2)) was reduced by two orders of magnitude (from 32.5 to 0.7).

TABLE I
NMF- and NMU-based Classifier for ASRS Documents

Step	Description
1	Filter elements of $\mathbf{H}$ given $\mathbf{A} \approx \mathbf{WH}$; for $i = 1, \ldots, n$, determine $\eta_i = \max(\mathbf{H_i})$ and zero out all values in $\mathbf{H_i}$ less than $\eta_i \times (1 - \delta)$.
2	Normalize the (new) filtered matrix $\mathbf{H}$ so that all column sums are 1.
3	Generate a set of indices (integers) that will partition the documents into the training ($\mathbf{R}$) and test ($\mathbf{T}$) subsets based on the σ parameter.
4	Cluster the columns of $\mathbf{H}$ corresponding to documents in the training set $\mathbf{R}$ by known anomalies (labels).
5	Sum the number of documents associated with each anomaly per NMF/NMU feature (k of them), and determine the number of anomaly j documents associated with feature i.
6	For each document in subset $\mathbf{T}$, produce a score (or probability) that the document is relevant to each anomaly.
7	Using α, produce the relevance score t_{ij} for (document i, anomaly j) pairs; the score will yield a positive prediction if $t_{ij} > \rho_i \times (1 - \alpha)$, where $\rho_i = max(\mathbf{H_i^T})$.

B. Classification Results

Figure 1 contains the best[3] Receiver Operating Characteristic (ROC) curves (true positive rate versus false positive rate) for the NMF and NMU classifiers, when applied to test ASRS documents (30 out of a 100). Among the 14 anomaly categories spanned by the first 100 ASRS documents, we see that the rank-10 NMU classifier achieved better classification accuracies than the rank-10 NMF classifier for 9 of the categories (see red entries of Table II), which was already obtaining very competitive results on this dataset. The fourteen (of the twenty-two) event types (or anomaly descriptions) listed in Table II were obtained from the Distributed National ASAP Archive (DNAA) maintained by the University of Texas Human Factors Research Project[4]. As the specificity of some topics in the ASRS collection can widely vary [23], it is not surprising to observe poor performance for both classifiers with a few anomaly categories (e.g., 2, 6, 7, and 22). Additional experiments with a larger numbers of features ($k > 10$) and documents ($n > 100$) should produce NMF and NMU models that better capture the diversity of contexts described by those events.

VI. Summary and Future Work

Whereas nonnegative matrix factorization (NMF) has been previously shown to be a viable alternative for automated

[1] See SMART's english stoplist at `ftp://ftp.cs.cornell.edu/pub/smart/english.stop`.

[2] Only five iterations were used in our preliminary study.

[3] After running each classifier ten times with different (random) training and test document sets $\mathbf{R}$ and $\mathbf{T}$, respectively.

[4] See `http://homepage.psy.utexas.edu/HomePage/Group/HelmreichLAB`.

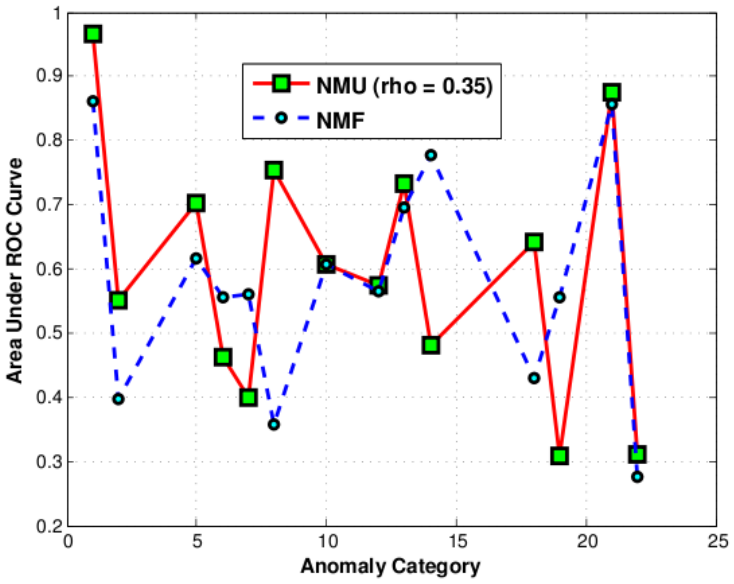

Fig. 1. NMF and NMU classification accuracies (areas under ROC curve) for 14 of the 22 DNAA anomaly categories.

TABLE II
ROC AREAS VERSUS DNAA EVENT TYPES FOR SELECTED ANOMALIES

Anomaly	DNAA Event Type	ROC Area NMF	ROC Area NMU
1	Airworthiness Issue	.8621	.9655
2	Noncompliance (policy/proc.)	.3971	.5502
5	Incursion (collision hazard)	.6173	.7037
6	Departure Problem	.5566	.4615
7	Altitude Deviation	.5600	.4000
8	Course Deviation	.3580	.7531
10	Uncommanded (loss of control)	.6071	.6071
12	Traffic Proximity Event	.5650	.5750
13	Weather Issue	.6964	.7321
14	Airspace Deviation	.7778	.4815
18	Aircraft Damage/Encounter	.4286	.6249
19	Aircraft Malfunction Event	.5556	.3086
21	Illness/Injury Event	.8571	.8750
22	Security Concern/Threat	.2759	.3103

document classification problems, the prospects for nonnegative matrix underapproximation (NMU) are even better. This study demonstrated how NMU can be used to both learn and assign (anomaly) labels for documents from the Aviation Safety Reporting System (ASRS). Of course, there is room for improvement in both the performance and interpretability of NMF- and NMU-based text classifiers. In particular, the summarization of anomalies (document classes) using k NMF/NMU features needs further work. Alternatives to the filtering of elements of the coefficient matrix $\mathbf{H}$ (based on the parameter δ) could be the use of sparsity or smoothing constraints (see [3]) on either (or both) factors $\mathbf{W}$ and $\mathbf{H}$.

ACKNOWLEDGMENTS

This research was sponsored by the National Aeronautics and Space Administration (NASA) Ames Research Center under contract No. 07024004. Nicolas Gillis is a research fellow of the Fonds de la Recherche Scientifique (F.R.S.-FNRS). This text presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. The scientific responsibility is assumed by the authors.

REFERENCES

[1] D. Lee and H. Seung, "Learning the Parts of Objects by Non-Negative Matrix Factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[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] 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.

[4] D. Lee and H. Seung, "Algorithms for Non-Negative Matrix Factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[5] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms," in *Proc. 6th Int. Conf. on ICA and Blind Signal Separation*, Charleston, SC, March 5-8 2006.

[6] Y. Wang, Y. Jiar, C. Hu, and M. Turk, "Fisher non-negative matrix factorization for learning local features," in *Asian Conference on Computer Vision*, Korea, January 27-30 2004.

[7] D. Guillamet, M. Bressan, and J. Vitria, "A Weighted Non-negative Matrix Factorization for Local Representations," in *Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, Kavai, HI, 2001, pp. 942–947.

[8] A. Hamza and D. Brady, "Reconstruction of Reflectance Spectra Using Robust Non-Negative Matrix Factorization," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3637–3642, 2006.

[9] I. Dhillon and S. Sra, "Generalized Nonnegative Matrix Approximations with Bregman Divergences," in *Proceeding of the Neural Information Processing Systems (NIPS) Conference*, Vancouver, B.C., 2005.

[10] C.-J. Lin, "Projected Gradient Methods for Nonnegative Matrix Factorization," *Neural Computation*, vol. 19, pp. 2756–2779, 2007, MIT press.

[11] E. Gonzalez and Y. Zhang, "Accelerating the Lee-Seung Algorithm for Nonnegative Matrix Factorization," Rice University, Tech. Rep. TR-05-02, March 2005.

[12] R. Zdunek and A. Cichocki, "Non-Negative Matrix Factorization with Quasi-Newton Optimization," in *Proc. 8th Int. Conf. on Artificial Intelligence and Soft Comp., ICAISC*, Zakopane, Poland, June 25-29 2006.

[13] C. Cichocki, R. Zdunek, and S. Amari, "Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization," in *ICA07, London, Lecture Notes in Comp. Sc., Vol. 4666, Springer, pp. 169-176*, 2007.

[14] N.-D. Ho, "Nonnegative matrix factorization - algorithms and applications," Ph.D. dissertation, Université catholique de Louvain, 2008.

[15] N. Gillis and F. Glineur, "Nonnegative Factorization and The Maximum Edge Biclique Problem," *CORE Discussion paper*, no. 64, 2008.

[16] S. Wild, J. Curry, and A. Dougherty, "Motivating Non-Negative Matrix Factorizations," in *Proceedings of the Eighth SIAM Conference on Applied Linear Algebra, July 15-19*. Williamsburg, VA: SIAM, 2003.

[17] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Journal of Pattern Recognition*, vol. 41, pp. 1350–1362, 2008.

[18] C.-J. Lin, "On the Convergence of Multiplicative Update Algorithms for Nonnegative Matrix Factorization," in *IEEE Transactions on Neural Networks*, 2007.

[19] S. Vavasis, "On the Complexity of Nonnegative Matrix Factorization," 2007, preprint.

[20] N. Gillis and F. Glineur, "Using Underapproximations for Sparse Nonnegative Matrix Factorization," *CORE Discussion paper*, no. 2009/6, 2009.

[21] N. Gillis, "Approximation et sous-approximation de matrices par factorisation positive: algorithmes, complexité et applications," Master's thesis, Université catholique de Louvain, 2007, in French.

[22] J. Giles, L. Wo, and M. Berry, "GTP (General Text Parser) Software for Text Mining," in *Software for Text Mining, in Statistical Data Mining and Knowledge Discovery*, H. Bozdogan, Ed. Boca Raton, FL: CRC Press, 2003, pp. 455–471.

[23] E. Allan, M. Horvath, C. Kopek, B. Lamb, T. Whaples, and M. Berry, "Anomaly Detection Using Nonnegative Matrix Factorization," in *Survey of Text Mining II: Clustering, Classification, and Retrieval*, M. Berry and M. Castellanos, Eds. London: Springer-Verlag, 2008, pp. 203–217.