Optimizing Conflicting Objectives in NMF Using Pareto Simulated Annealing

Kevin Foley
School of Computer Science & Informatics
University College Dublin
kevnfoley@gmail.com

Derek Greene
School of Computer Science & Informatics
University College Dublin
derek.greene@ucd.ie

Pádraig Cunningham
School of Computer Science & Informatics
University College Dublin
padraig.cunningham@ucd.ie

Abstract

Non-Negative matrix factorization (NMF) has emerged as an important technique for simplifying high-dimension data into interpretable factors. NMF has the attractive characteristic that the factor matrices are naturally sparse, thus allowing them to be readily interpreted. However, there is a tension between the accuracy of the factorization and the sparseness – it is the management of the trade-off between these two criteria that is the subject of this paper. We introduce a multi-criteria Simulated annealing framework that produces a Pareto set of solutions, which are non-dominated on both criteria. We show that solutions at one end of the Pareto front of solutions correspond to NMF factorizations produced with conventional optimization techniques, while solutions at the other end exhibit enhanced sparseness. Clustering is no longer to be observed either in the raw-data form of the matrix, or the generated heat-map form.

1 Introduction

Non-negative matrix factorization (NMF) algorithms simplify data represented in matrix form, while simultaneously uncovering the underlying latent structures in the data. For instance, in text analysis, if a matrix $\mathbf{A}$ represents m unique terms contained in n documents, NMF will factorize this matrix into two reduced-order non-negative factors that approximate the original data matrix:

$$\mathbf{A} \approx \mathbf{WH} \quad \text{where} \quad \mathbf{W} \in \mathbb{R}^{m \times k} \, , \mathbf{H} \in \mathbb{R}^{k \times n} \quad \text{and} \quad \mathbf{W} \geq 0 \, , \mathbf{H} \geq 0$$

The factor $\mathbf{H}$ can be considered a projection of the documents into a latent space defined by basis vectors in $\mathbf{W}$. The number of dimensions in the factors is determined by the value k – this value reflects the number of latent groups in the data. Lowering the value of k reduces the order of $\mathbf{W}$ and $\mathbf{H}$, without changing the order of $\mathbf{WH}$. When a matrix representing a large document collection is factorized in this way, patterns within the data begin to be revealed [10, 5].

Due to the intractable nature of the problem of factorizing large matrices, approximations are sought that minimize the difference between the approximation $\mathbf{W} \times \mathbf{H}$ and the original matrix $\mathbf{A}$. How *equivalent* the approximation is to the original $\mathbf{A}$ matrix is measured using some error criterion, most commonly Euclidean distance. This *reconstruction error* is one of the two competing criteria that is the subject of this paper and it is referred to as 'distance' in the evaluation. Orthodox methods for NMF are based on monotonic, multiplicative update rules[8, 10] that move initial, randomly

created matrix factors in an iterative process towards matrix factors with a low reconstruction error (*i.e.* distance).

NMF has the advantage that the factor matrices $\mathbf{W}$ and $\mathbf{H}$ tend to be sparse. This is advantageous in that it improves the interpretability of the resulting clusters. If the factors $\mathbf{W}$ and $\mathbf{H}$ become too sparse then the reconstruction error will increase accordingly. This highlights a trade-off between reconstruction error and sparseness. This suggests that if we are interested in sparseness *and* reconstruction error we have an optimization problem with two conflicting criteria.

In this paper we explore Pareto simulated annealing (PSA) [1] as a potential solution to this problem. Evolutionary algorithms, specifically those based on simulated annealing (SA), move initial random solutions towards optimal solutions in a stochastic manner [15], while making provision for backward steps away from the local optimal solutions in a controlled manner. The reason for these backward moves is to prevent algorithms becoming stuck in these local minima. PSA extends this idea to simultaneously optimize more than one criterion. This is done in a Pareto optimality framework [13] that considers the optimality of a solution based on the dominance or non-dominance of the various criteria - see Section 3.1. The idea of PSA is to find not just one optimal solution, as with an orthodox update algorithm, but instead to find a set of optimal solutions that lie along a frontier created by the two competing criteria - see Figure 3.

The remainder of the paper is structured as follows. Overviews of NMF and multi-criteria optimization are provided in sections 2 and 3 respectively. We then present our novel PSA optimization framework for NMF in section 3.3. An evaluation of this framework on two text corpora is described in section 4.

2 Non-negative Matrix Factorization

When factorizing large matrices, the factors created ($\mathbf{W}$ and $\mathbf{H}$) must be evaluated to determine how accurate the reconstructed matrix ($\mathbf{WH}$) is when compared to the original large matrix ($\mathbf{A}$). In this section we outline how the quality of matrix factors is determined as well as showing why the property of sparsity within these matrix factors is so important from the perspective of cluster identification. Evaluation of solutions to NMF problems from both of these perspectives in necessary if multi-criteria optimization is be achieved.

2.1 NMF Formulation & Applications

A disadvantage of common dimensionality reduction techniques, such as principle component analysis and related spectral clustering methods, stems from the presence of negative entries in the eigenvectors or singular vectors used to construct the reduced space. As a result, the features of a reduced representation do not have any intuitive meaning. Therefore, the application of a post-processing technique is generally required to produce a final partition of the data. This is an issue that is inherent in many feature extraction techniques.

To address this problem, Lee and Seung [8] proposed an alternative unsupervised approach for reducing the dimensionality of non-negative matrices, referred to as Non-negative matrix factorization. Unlike spectral methods, NMF algorithms seek to decompose the data into factors that are constrained so that they will not contain negative values. By modeling each object as the additive combination of a set of non-negative basis vectors, a readily interpretable clustering of the data can be produced without the requirement for further post-processing. These basis vectors are generally not required to be orthogonal, which facilitates the discovery of overlapping groups.

In recent years a large number of variants of Lee and Seung's original approach have been proposed [3]. In the standard formulation, given a rectangular non-negative matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ such that each column represents a data object, NMF generates a reduced rank-k approximation in the form of the product of two non-negative factors $\mathbf{A} \approx \mathbf{WH}$. The rows of the factor $\mathbf{W}$ represent a set of k basis vectors, and the columns of $\mathbf{H}$ are linear combinations of the basis vectors. Typically the number of basis vectors k is a user-defined parameter, which is chosen so that $k << min(m, n)$.

In addition to applications in bioinformatics and image processing, NMF has also been successfully used to cluster document collections [16]. After applying NMF to the term-document representation $\mathbf{A}$ of a corpus, the resulting factor $\mathbf{W}$ can be viewed as a set of semantic variables corresponding to

the topics in the data, while $\mathbf{H}$ describes the contribution of the documents to each topic. This idea of representing each document as the additive combination of several overlapping topics is intuitive. Furthermore, the non-negativity of the factors allows them to be directly interpreted as a soft k-way co-clustering of both documents and terms — the rows of $\mathbf{W}$ provide the term cluster memberships, while the columns of $\mathbf{H}$ provide the document cluster memberships.

2.2 Reconstruction Error & Sparseness

The choice of factors in NMF is determined by some objective function that seeks to minimize the error of the reconstruction of $\mathbf{A}$ by the product $\mathbf{WH}$. In the original formulation proposed by Lee and Seung [8], this involves minimizing squared Euclidean distance as computed by the Frobenius norm:

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

Although other objectives have been proposed, such as information theoretic divergence measures [9], the Euclidean objective has been most widely adopted. An approximate solution to minimize $f(\mathbf{W}, \mathbf{H})$ may be found by employing a diagonally re-scaled gradient descent search strategy, which involves alternating between a pair of multiplicative update rules [8]. At each iteration, $\mathbf{W}$ and $\mathbf{H}$ are updated by multiplying the current factors by a measure of the quality of the current approximation $\mathbf{WH}$. This process continues until the search procedure converges to a local minimum.

NMF has a tendency to produce basis vectors that are somewhat sparse (*i.e.* $\mathbf{W}$ and $\mathbf{H}$ will contain a large proportion of near-zero values). This can potentially lead NMF to perform well on data where structures exist in different low-dimensional subspaces of the original high-dimensional space. It also helps users to identify coherent clusters, based on the presence of prominent non-zero values in the factor matrices. However, in the default formulation of NMF, there exists no explicit means of controlling the degree of sparseness required for a particular application. While the standard NMF formulations described previously do tend to generate reasonably sparse factors, it may be desirable in some situations to increase the level of sparsity in the basis vectors to produce a more localized solution. Several algorithms have been proposed that implement sparsity constraints by including additional penalty terms in the objective function, including local non-negative matrix factorization (LNMF) [11] and sparse non-negative matrix factorization (SNMF) [12]. In both cases, user-defined parameters are used to manually control the trade-off between sparseness and reconstruction error.

3 Multi-Criteria Optimization and Pareto Simulated Annealing

Standard update-based methods for NMF search for optimal solutions with regard to one criterion, such as Euclidean distance. However, sparseness is also desirable when it comes to interpreting the factors. As outlined previously, when interpreting a matrix factor, clustering is identified by the most prominent values in the factor matrices. Optimizing both these criteria can be achieved in a number of ways, one of which is aggregation of all the criteria being optimized using some form of weighted sum being one. However, this requires commitment to a weighting of the criteria in order to produce a single optimal solution according to that weighting, such as in LNMF [11]. The aim of this paper is to show how a set of optimal solutions, not just a single optimal solution, may be produced in a Pareto optimality framework [13].

3.1 Pareto Optimality

When criteria compete within a problem space, as solutions progress toward a global optimum a frontier is formed beyond which no viable solutions exist. All the solutions along this frontier are said to be *Pareto optimal* and hence this efficient frontier is known as the Pareto front. The area beyond the frontier is known as the *Utopian space* [2] where no solutions exist due to the constraints of the problem. The solutions along the efficient frontier are said to be *non-dominated* with respect to each other, i.e. no solution in the frontier dominates another on all criteria. A solution is said to be *Pareto optimal* if no part of the solution can be made better without making some other criterion worse [14].

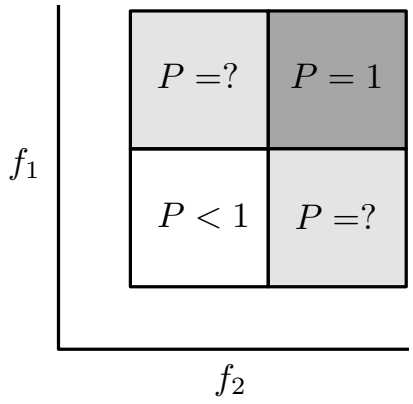

Figure 1: Multi-objective probability rules for the case of two criteria.

3.2 Simulated Annealing

Simulated annealing is a stochastic strategy for exploring very large search spaces. It is particularly appropriate for exponential search problems where an exhaustive exploration of the search space is impossible [7]. It is akin to a hill-climbing search strategy with the added provision that reversals (inferior solutions) may be accepted. The acceptance of reversals is controlled by an artificial temperature variable T so that the acceptance of reversals becomes increasingly unlikely as the system 'cools'.

In single-criterion SA, if a solution x' improves on x, then it is always accepted. If x' is inferior to x, it may still be accepted with probability:

$$P(x, x', T) = exp\left((f(x) - f(x')/T\right) \tag{2}$$

where $f(x)$ is the criterion being optimized, and T is a variable representing the current annealing temperature.

3.3 Pareto Simulated Annealing

When simulated annealing is required to address more than one optimization criterion two new requirements are introduced. Instead of considering a single 'best' candidate solution at any time the system must manage a Pareto front of solutions. Second, the decision being made in Eqn. 2 is made more complicated by the fact that there is more than one criterion to be considered.

In Czyzak and Jaszkievicz's work on Pareto simulated annealing [1] the problem of accounting for more than one criterion is addressed. In the case of a single-criterion problem, a solution that is better than the current solution will be accepted with a probability of 1. If an inferior solution is created, it will be accepted with a probability of < 1, based on Eqn. 2. However, if we have two criteria, there are three possible scenarios to be considered:

- Solution x' dominates solution x $(P = 1)$
- Solution x' is dominated by solution x $(P < 1)$
- Solution x' is non-dominated with respect to solution x $(P =?)$

These scenarios are illustrated in Figure 1. It is clear that the second and third scenarios require a more complex treatment of the probability calculation. If $f1$ and $f2$ are the two criteria to be optimized, then the two segments shaded a lighter gray in Figure 1 exhibit an improvement in one criterion and a dis-improvement in the other. The probability calculation therefore takes both criteria into account and is formally defined by Czyzak and Jaszkievicz as follows:

$$P(x, x', T, \Lambda) = min\left\{1, exp\left(\sum_{j=1}^{J} \lambda_j \left(f_j(x) - f_j(x')\right)/T\right)\right\} \tag{3}$$

where Λ represents the vector of weights λ_j for the various criteria.

It may be desirable to bias the search in favor of one objective function over another so that different portions of the efficient frontier may be explored. There is a requirement to normalize the functions so that the movement of a function in one solution when compared with the function in the original solution is treated as a percentage improvement, rather than a non-normalized value. In this way the probability calculation is independent of the criteria used to measure the quality of a solution.

In any realization of the PSA idea in a new problem domain, the perturbation process that produces x' from x will be domain specific. The details of the perturbation process used here are provided in [4]. In summary, perturbation entails a random mutation of some of the entries in the matrix factors and a provision to insert zero entries in a controlled manner.

4 Evaluation

A comprehensive evaluation of the PSA-based NMF framework is presented in [4]. The main findings arising from the evaluation are summarized here:

- The emergence of a Pareto front in all runs of the framework shows that there is a conflict between distance and sparseness – see Figures 2 and 3.
- The PSA-NMF algorithm produces solutions close in distance terms to those produced by the conventional update method.
- The PSA-NMF algorithm produces an impressive spectrum of solutions with different degrees of sparseness – see Figures 4 to 7.

4.1 Experimental Setup

The results we report here are based on two datasets constructed from a corpus of news articles from the BBC's sports website, which was previously used in document clustering [6]. The larger dataset contains the full set of 737 documents, covering five different topics. The smaller dataset contains a subset of 348 of these documents, represented by 2,660 terms, and relating to articles on athletics, rugby, and tennis. The PSA-NMF algorithm was run over 5000 iterations with a working set size of 25. The mutation rate for the perturbation of the candidate solutions at each iteration of the algorithm ranged from 0 to 0.02. These parameters resulted in a gradual advancement toward the efficient frontier, with the set of Pareto-optimal solutions tracked along the way.

4.2 Spectrum of Solutions

The final set of Pareto solutions for the dataset of 737 documents is shown in Figure 2. The sparseness of the $\mathbf{WH}$ factor matrix is plotted on the x-axis, and the value of the Euclidean distance criterion is on the y-axis. Note that sparseness here is calculated as the fraction of zero values in the matrix. For the distance criterion, a smaller value is better, so the Utopian point, in this case the point that represents a solution with zero distance and full sparsity, is in the bottom left of the graph. A collection of solutions produced by the standard update-based NMF is also shown. While these are better than the solutions on the Pareto front by a small margin on the distance criterion, they do not display anything approaching the range of sparseness scores achieved by the PSA-NMF strategy.

4.3 Clustering Behavior

In this section, we use the smaller dataset of 348 documents to show the range of sparseness in solutions produced by the PSA-NMF method, and examine the degree of clustering in these solutions. The Pareto front for the dataset is shown in Figure 3. In this case sparseness is calculated as the fraction of non-zero values in the $\mathbf{H}$ matrix, which is of particular interest as it contains document membership weights. A selection of the transpose of these $\mathbf{H}$ matrices is shown as heat-maps in Figures 4 to 7. These heat maps correspond to the point circled in Figure 3. As a matrix value moves from 0 to 1, the corresponding shade assigned to its value moves from white to black. This provides a good visualization of how sparse a solution is, and how well the non-zero elements of the

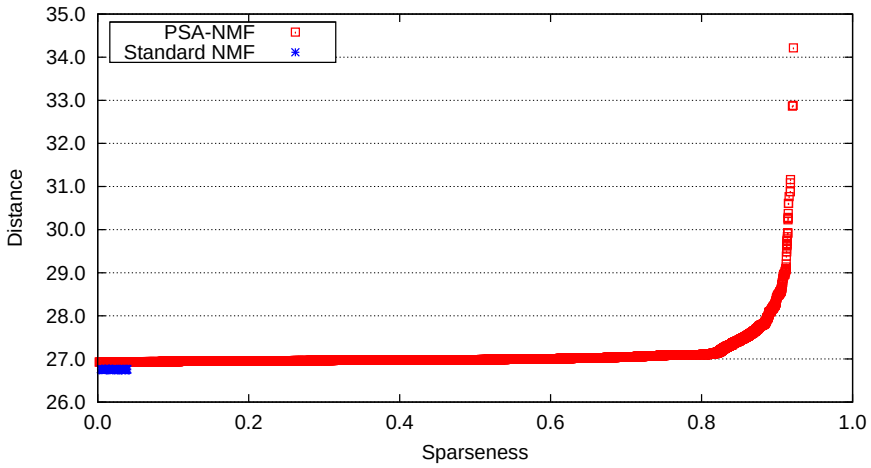

Figure 2: Graph of Pareto-optimal solutions versus standard NMF on the larger 737 document set, showing the sparseness of the factor matrix $\mathbf{H}$ and the corresponding Euclidean distance value of the solution.

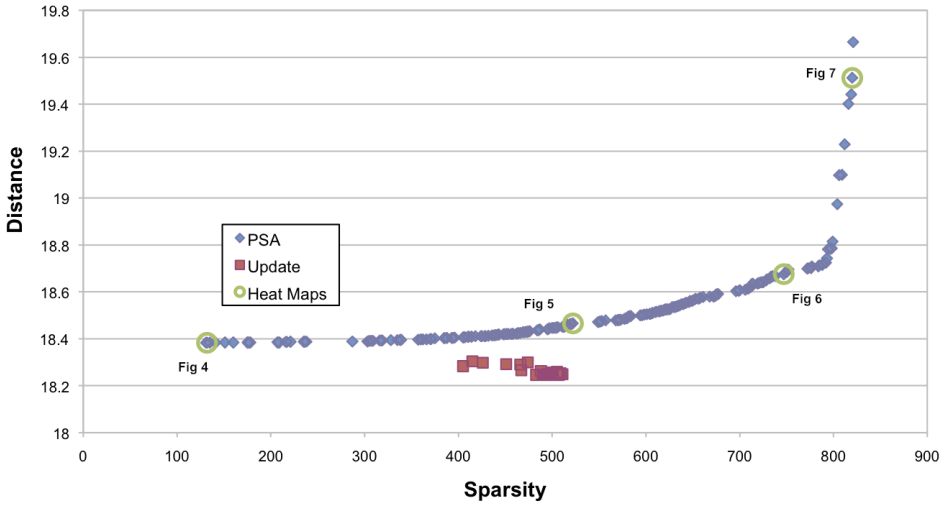

Figure 3: Graph of Pareto-optimal solutions versus standard NMF on the smaller 348 document set, based on the factor $\mathbf{H}$.

matrix cluster together. A good solution would see all documents (*i.e.* rows) pertaining to a particular subject clustering together in the same column, with each document represented by a dark line. Each document cluster should appear in a different column with the rest of the matrix colored gray to white. The documents in the dataset are sorted by topic, so clustering is instantly recognizable.

Moving along the curve in Figure 3 from low sparseness and distance to high sparseness and distance, the effect of increasing sparseness in the $\mathbf{H}$ matrix is observed. Those solutions that have the lowest distance also have the low sparseness values, hence those matrix entries for documents that do not relate to a particular topic are gray in color. The benefit of enhancing sparseness beyond what is achievable using standard NMF optimization methods are best seen in Figure 6 as the emphasis on sparseness forces documents to belong to one cluster only. Figure 7 is a $\mathbf{H}$ matrix sampled at the end of the Pareto curve. Here the clustering can no longer be observed, as the solution consists largely of zero values. These results demonstrate the ability of the algorithm to explore the entire search space.

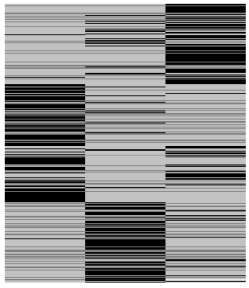

Figure 4: Low Distance/Sparsity

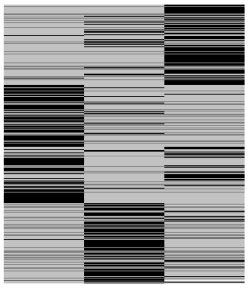

Figure 5: Mid Distance/Sparsity

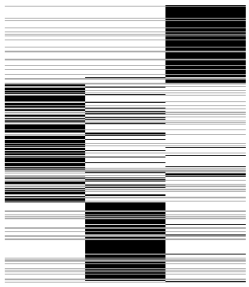

Figure 6: High Distance/Sparsity

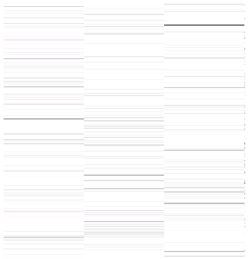

Figure 7: Highest Distance/Sparsity

5 Conclusions

This paper addresses a problem in NMF, which arises from the fact that the standard update-based NMF formulation solely emphasizes reconstruction error. However, matrix factorization tasks may have other objectives that are desirable to optimize in order to produce useful solutions. A notable examples is the sparseness of the resulting factors, which can aid in the interpretability of the algorithm output. Those algorithms that do tackle the problem from a two criteria perspective did so using some form of weighted sum technique.

We recast NMF in a Pareto simulated annealing framework that returns a Pareto set of solutions, which optimizes both distance and sparseness. Because the two criteria are competing with one another, the range of solutions is vast, and the optimal solutions lie along a Pareto front. Standard NMF optimization strategies fail to explore this Pareto frontier. In contrast, the NMF-PSA algorithm proposed here was designed to find as many optimal solutions along the frontier as possible, and allows for the inspection of these solutions to determine if meaningful clustering could be observed at various points along the curve. The differing patterns illustrated by the $\mathbf{H}$ matrix heat-maps demonstrated clustering behavior at various points along the curve, and highlighted the merit of pursuing all solutions that lie on the Pareto frontier, rather than merely focusing on those solutions that optimize by a single criterion.

In future work we propose to explore the prospect of modifying the LNMF algorithm described in section 2 to achieve the multi-criteria objectives addressed in this work. LNMF uses a weight parameter to control the trade-off between distance and sparseness. This parameter could be adjusted over multiple runs to produce a range of solutions. It would be interesting to implement this idea and compare performance against PSA-NMF, both in terms of optimality and computational effort.

Acknowledgments

This work is partly supported by Science Foundation Ireland Grant No. 08/SRC/I140 (Clique: Graph & Network Analysis Cluster).

References

[1] P. Czyzak and A. Jaszkiewicz. Pareto simulated annealing - a metaheuristic technique for multiple-objective combinatorial optimization. *Journal of Multi-Criteria Decision Analysis*, 7(34-37), 1998.

[2] K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley, Kanpur, India, 2001.

[3] K. Devarajan. Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Comput Biol*, 4(7), Jul 2008.

[4] K. Foley. Multi-criteria optimisation of non-negative matrix factorisation problems using pareto simulated annealing techniques. Master's thesis, UCD, 2010.

[5] D. Greene and P. Cunningham. Producing accurate interpretable clusters from high-dimensional data. *Proc. 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05)*, pages 486–494, 2005.

[6] D. Greene and P. Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proc. 23rd International Conference on Machine learning (ICML'06)*, pages 377–384, 2006.

[7] S. Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, 34(5/6), 1984.

[8] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91, October 1999.

[9] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562, 2000.

[10] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances In Neural Information Processing Systems*, 13:556–562, 2001.

[11] S. Z. Li, X. W. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. In *Proc. Computer Vision and Pattern Recognition (CVPR'01)*, volume 1, page 207, 2001.

[12] W. Liu, N. Zheng, and X. Lu. Non-negative matrix factorization for visual coding. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*, volume 3, pages 293–296, April 2003.

[13] V. Pareto. *Cours d'Economie Politique*. Rouge, Lausanne, Switzerland, 1896.

[14] C. J. Petrie, T. A. Webster, and M. R. Cutkosky. Using pareto optimality to coordinated distrubuted agents. *AIEDAM special issue on conflict management*, 9:269–281, 1995.

[15] V. Snášel, J. Platoš, and P. Krömer. On genetic algorithms for boolean matrix factorization. In *Eighth International Conference on Intelligent Systems Design and Applications*, pages 170–175, 2008.

[16] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273, 2003.