

2010/59



Nonnegative factorization
and the maximum edge biclique problem

Nicolas Gillis and François Glineur



CORE

DISCUSSION PAPER

Center for Operations Research
and Econometrics

Voie du Roman Pays, 34
B-1348 Louvain-la-Neuve
Belgium

<http://www.uclouvain.be/core>

CORE DISCUSSION PAPER
2010/59

**Nonnegative factorization
and the maximum edge biclique problem**

Nicolas GILLIS¹ and François GLINEUR²

October 2010

Abstract

Nonnegative matrix factorization (NMF) is a data analysis technique based on the approximation of a nonnegative matrix with a product of two nonnegative factors, which allows compression and interpretation of nonnegative data.

In this paper, we study the case of rank-one factorization and show that when the matrix to be factored is not required to be nonnegative, the corresponding problem (RINF) becomes NP-hard. This sheds new light on the complexity of NMF since any algorithm for fixed-rank NMF must be able to solve at least implicitly such rank-one subproblems.

Our proof relies on a reduction of the maximum edge biclique problem to RINF. We also link stationary points of RINF to feasible solutions of the biclique problem, which allows us to design a new type of biclique finding algorithm based on the application of a block-coordinate descent scheme to RINF. We show that this algorithm, whose algorithmic complexity per iteration is proportional to the number of edges in the graph, is guaranteed to converge to a biclique and that it performs competitively with existing methods on random graphs and text mining datasets.

Keywords: nonnegative matrix factorization, rank-one factorization, maximum edge biclique problem, algorithmic complexity, biclique finding algorithm.

JEL Classification: 15A23, 68Q25, 90C06, 90C27 90C35, 90C59

¹ Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium. E-mail: Nicolas.gillis@uclouvain.be.

² Université catholique de Louvain, CORE, B-1348 Louvain-la-Neuve, Belgium. E-mail: francois.glineur@uclouvain.be.
This author is also member of ECORE, the association between CORE and ECARES.

We thank Pr. Paul Van Dooren and Pr. Laurence Wolsey for helpful discussions and advice. The first author is a research fellow of the Fonds de la Recherche Scientifique (F.R.S.-FNRS).

This paper 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.

1 Introduction

(Approximate) *Nonnegative matrix factorization* (NMF) is the problem of approximating a given nonnegative matrix by the product of two low-rank nonnegative matrices: given a $m \times n$ nonnegative real matrix $M \in \mathbb{R}_+^{m \times n}$ and a factorization rank r , one has to compute two nonnegative matrices $V \geq 0$ and $W \geq 0$ of dimensions $m \times r$ and $r \times n$ such that

$$M \approx VW . \tag{1.1}$$

Typically the quality of the approximation is measured by the Frobenius norm¹ of the residual error matrix, and one tries to solve:

$$\min_{V \in \mathbb{R}^{m \times r}, W \in \mathbb{R}^{r \times n}} \|M - VW\|_F^2 \quad \text{such that } V \geq 0, W \geq 0. \tag{NMF}$$

This problem was first introduced in 1994 by Paatero and Tapper [26], and more recently received a considerable interest after the publication of two papers by Lee and Seung [21, 22]. It is now well established that NMF is useful in the framework of compression and interpretation of nonnegative data ; it has for example been applied in analysis of image databases, text mining, interpretation of spectra, computational biology and many other applications (see [2, 8, 9] and references therein). Unfortunately (NMF) is a NP-hard optimization problem [29] and therefore we cannot expect to solve it up to global optimality in a reasonable computational time. Therefore most practical algorithms proposed to find approximate solutions of (NMF), based on iterative optimization schemes (see, e.g., [2, 3, 5, 6, 9, 14, 19, 23]), offer no guarantee on the global optimality of the solutions they provide. How can one interpret the outcome of a NMF? Assume each column $M_{:j}$ of matrix M represents an element of a data set: expression (1.1) can be equivalently written as

$$M_{:j} \approx \sum_k V_{:k} W_{kj}, \quad \forall j, \tag{1.2}$$

where each element $M_{:j}$ is decomposed into a nonnegative linear combination (with weights W_{kj}) of nonnegative basis elements ($\{V_{:k}\}$, the columns of V). Nonnegativity of V allows interpretation of the basis elements in the same way as the original nonnegative elements in M , which is crucial in applications where the nonnegativity property is a requirement (e.g., where elements are images described by pixel intensities or texts represented by vectors of word counts). Moreover, nonnegativity of the weight matrix W corresponds to an essentially additive reconstruction which leads to a *part-based representation*: basis elements will represent similar parts of the columns of M . Sparsity is another important consideration: finding sparse factors improves compression and leads to a better part-based representation of the data [18]. In particular, when dealing with sparse matrices, NMF can be interpreted as a biclustering model, see [10, 20] and references therein. In fact, each rank-one factor of the decomposition will correspond to a dense rectangular submatrix of M (a bicluster), enabling NMF to detect interactions between columns and rows of the matrix M (e.g., in text mining applications, NMF extracts closely related sets of texts and words [11]).

In the special case where we seek a rank-one factorization (i.e. when $r = 1$), NMF is known to be polynomially solvable (using the Eckart-Young and Perron-Frobenius Theorems, it reduces to computing the dominant left and right singular vectors). The central problem studied in this paper, called rank-one nonnegative factorization (R1NF), is an extension of rank-one NMF where the matrix to be approximated by the outer product of two nonnegative vectors is now allowed to contain negative elements.

R1NF is introduced in Section 2, where it is shown that allowing negative elements in the matrix transforms the polynomially solvable rank-one NMF problem into a NP-hard problem. The reduction

¹ $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$.

used on the proof is based on the problem of finding a maximum edge biclique in a bipartite graph. Because any algorithm designed to solve NMF must at least implicitly solve R1NF problems, this hardness result sheds new light on the limitations of NMF algorithms and the complexity of NMF when the factorization rank r is fixed.

In Section 3, stationary points of the R1NF problem used in the above-mentioned reduction are shown to coincide with bicliques of the corresponding graph. Building on that fact, Section 4 introduces a new type of biclique finding algorithm that relies on the application of a simple nonlinear optimization scheme (block-coordinate descent) to the equivalent R1NF problem considered earlier, which only requires for each iteration a number of operations proportional to the number of edges of the graph. This method is then compared to a greedy heuristic and an existing algorithm [12] on some synthetic and text mining datasets, and is shown to perform competitively.

2 Rank-one Nonnegative Factorization (R1NF)

2.1 Motivation

Solving (NMF) amounts to finding r nonnegative rank-one factors $V_{:k}W_{k:}$, each having to satisfy the following equality as well as possible

$$V_{:k}W_{k:} \approx M - \sum_{i \neq k} V_{:i}W_{i:} \doteq R_k \not\geq \mathbf{0} \quad \forall k,$$

i.e. each of them should be the best possible nonnegative rank-one approximation of the corresponding residual matrix, denoted R_k . It is important to notice here that, unlike input matrix M , matrices R_k can contain negative elements. Therefore, any NMF algorithm has to solve, at least implicitly, the following subproblems

$$\min_{V_{:k} \in \mathbb{R}^m, W_{k:} \in \mathbb{R}^n} \|M - VW\|_F^2 = \|R_k - V_{:k}W_{k:}\|_F^2 \quad \text{such that } V_{:k} \geq \mathbf{0}, W_{k:} \geq \mathbf{0}, \quad (2.1)$$

for each k . We may wonder whether these subproblems can be solved efficiently, i.e., ask ourselves

Is it possible to compute efficiently the best rank-one nonnegative approximation of a matrix which is not necessarily nonnegative?

An interesting observation is that computing the globally² optimal value of $V_{:k}$ for a given value of $W_{k:}$ can be done in closed-form (and similarly for computing the optimal value of $W_{k:}$ for a fixed $V_{:k}$):

$$V_{:k}^* = \operatorname{argmin}_{V_{:k} \geq \mathbf{0}} \|R_k - V_{:k}W_{k:}\|_F^2 = \max\left(\mathbf{0}, \frac{R_k W_{k:}^T}{\|W_{k:}\|_2^2}\right), \quad (2.2)$$

$$W_{k:}^* = \operatorname{argmin}_{W_{k:} \geq \mathbf{0}} \|R_k - V_{:k}W_{k:}\|_F^2 = \max\left(\mathbf{0}, \frac{V_{:k}^T R_k}{\|V_{:k}\|_2^2}\right). \quad (2.3)$$

One can therefore try to solve (2.1) and, more generally, (NMF) by updating successively the columns of V and rows of W . This scheme, which amounts to a block-coordinate descent method, was proposed by Cichocki et al. [5] and called hierarchical alternating least squares (HALS) (see also [4, 17]). It has been observed to work remarkably well in practice, and in particular it clearly outperforms the standard multiplicative updates (MU) of Lee and Seung [22].

²Indeed, each of these two subproblems is convex.

2.2 Definition of R1NF and Implications for NMF

In order to shed some light on the above question, we define the problem of rank-one nonnegative factorization³ (R1NF) to be the variant of rank-one NMF where the matrix to be factorized can be *any real matrix*, i.e., is not necessarily nonnegative. Formally, given an $m \times n$ real matrix $R \in \mathbb{R}_+^{m \times n}$, one has to find a nonnegative column vector $v \in \mathbb{R}^m$ and a nonnegative row vector $w \in \mathbb{R}^n$ such that the nonnegative rank-one product⁴ vw is the best possible approximation (in the Frobenius norm) of matrix R :

$$\min_{v \in \mathbb{R}^m, w \in \mathbb{R}^n} \|R - vw\|_F^2 \quad \text{such that } v \geq 0, w \geq 0. \quad (\text{R1NF})$$

The next subsection shows that, in contrast with standard rank-one NMF, this problem is NP-hard, which provides the following new insights about the NMF problem:

- We cannot expect to be able to solve subproblems (2.1) efficiently up to global optimality, and the HALS algorithm most probably cannot be improved with a better scheme for successively computing rank-one factors $V_{:k}W_k$: arising in (2.1). More generally, any algorithm for NMF cannot expect to solve at each iteration a subproblem where a given column of $V_{:k}$ and its corresponding row W_k : are to be optimized simultaneously which shows that, in that sense, the partition of variables for block-coordinate schemes such as alternative nonnegative least squares (ANLS, optimizing V and W alternatively) [20] and (implicitly) HALS is best possible.
- Recall that the NP-hardness result characterizing NMF requires both the dimensions of matrix M and the factorization rank r of M increase, and that the complexity of NMF for a fixed rank r is currently not known (except⁵ in the polynomially solvable rank-one case). Our hardness result on (R1NF) therefore suggests that NMF is also a difficult problem for any fixed rank $r \geq 2$. Indeed, even if one was given the optimal solution of a NMF problem except for a single rank-one factor, it is not guaranteed that one would be able to find this last factor in polynomial-time, since the corresponding residual matrix is not necessarily nonnegative.

2.3 Complexity of R1NF and the Maximum Edge Biclique Problem in Bipartite Graphs

In this section, we show how the optimization version of the maximum edge biclique problem (MB) can be formulated as a specific rank-one nonnegative factorization problem (R1NF-MB). Since the decision version of (MB) is NP-complete [27], this implies that rank-one nonnegative factorization (R1NF) is in general NP-hard.

A *bipartite graph* G_b is a graph whose vertices can be divided into two disjoint sets V_1 and V_2 such that there is no edge between two vertices in the same set

$$G_b = (V, E) = \left(V_1 \cup V_2, E \subseteq (V_1 \times V_2) \right).$$

A *biclique* K_b is a complete bipartite graph, i.e., a bipartite graph where all the vertices are connected

$$K_b = (V', E') = \left(V'_1 \cup V'_2, E' = (V'_1 \times V'_2) \right).$$

³This terminology has already been used for the problem of finding a symmetric nonnegative factorization, i.e., one where $V=W$, but we assign it a different meaning in this paper.

⁴In the sequel, it is always assumed that v and w are respectively a column and a row vector, i.e. that the rank-one matrix vw is equal to the outer product of v and w .

⁵In fact, testing whether a nonnegative matrix admits a rank-two nonnegative factorization can also be done in polynomial time [7], but, when the answer is negative, finding the best possible rank-two approximate nonnegative factorization has unknown complexity status.

The so-called maximum edge biclique problem in a bipartite graph $G_b = (V, E)$ is the problem of finding a biclique $K_b = (V', E')$ in G_b (i.e., $V' \subseteq V$ and $E' \subseteq E$) maximizing the number of edges. The decision problem: *Given B , does G_b contain a biclique with at least B edges?* has been shown to be NP-complete [27], and the corresponding optimization problem is at least NP-hard.

Let $M_b \in \{0, 1\}^{m \times n}$ be the biadjacency matrix of the unweighted bipartite graph $G_b = (V_1 \cup V_2, E)$ with $V_1 = \{s_1, \dots, s_m\}$ and $V_2 = \{t_1, \dots, t_n\}$, i.e., $M_b(i, j) = 1$ if and only if $(s_i, t_j) \in E$. We denote by $|E|$ the cardinality of E , i.e., the number of edges in G_b ; note that $|E| = \|M_b\|_F^2$. The set of zero values will be denoted $Z = \{(i, j) \mid M_b(i, j) = 0\}$, and its cardinality $|Z|$, with $|E| + |Z| = mn$. With this notation, the maximum biclique problem in G_b can be formulated as

$$\begin{aligned} \min_{v, w} \quad & \|M_b - vw\|_F^2 \\ & v_i w_j \leq M_b(i, j), \forall i, j, \\ & v \in \{0, 1\}^m, w \in \{0, 1\}^n. \end{aligned} \tag{MB}$$

In fact, one can check easily that this objective is equivalent to $\max_{v, w} \sum_{i, j} v_i w_j$ since M_b , v and w are binary: instead of maximizing the number of edges inside the biclique, one minimizes the number of edges outside.

Feasible solutions of (MB) correspond to bicliques of G_b . We will be particularly interested in *maximal* bicliques, which are bicliques not contained in a larger biclique.

The corresponding rank-one nonnegative factorization problem is defined as

$$\min_{v \in \mathbb{R}^m, w \in \mathbb{R}^n} \|M_d - vw\|_F^2 \quad \text{such that} \quad v \geq \mathbf{0}, w \geq \mathbf{0} \tag{R1NF-MB}$$

where M_d is the matrix M_b where the zero values have been replaced by $-d$, i.e.

$$M_d = (1 + d)M_b - d \mathbf{1}_{m \times n}, \quad d > 0, \tag{2.4}$$

and $\mathbf{1}_{m \times n}$ is the matrix of all ones with dimension $m \times n$. Although (R1NF-MB) is a continuous optimization problem, we are going to show that, for a sufficiently large value of d , any of its optimal solutions has to *coincide* with a binary optimal solution of the corresponding (discrete) biclique problem (MB), which will then imply NP-hardness of (R1NF).

Intuitively, if a $-d$ entry of M_d is approximated by a positive value, say p , the corresponding term in the squared Frobenius norm of the error is $d^2 + 2pd + p^2$. As d increases, it becomes more and more costly to approximate $-d$ by a positive number and we will show that, for d is sufficiently large, negatives values of M_d have to be approximated by zeros. Since the remaining values (not approximated by zeros) are all ones, the optimal rank-one solution will be binary.

From now on, we say that a solution (v, w) *coincides* with another solution (v', w') if and only if $vw = v'w'$ (i.e., if and only if $v' = \lambda v$ and $w' = \lambda^{-1}w$ for some $\lambda > 0$). We also let $M_+ = \max(\mathbf{0}, M)$, $M_- = \max(\mathbf{0}, -M)$, $\min(M) = \min_{i, j}(M_{ij})$ and $\|M\|_2$ be the standard matrix 2-norm of M , i.e. $\|M\|_2 = \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|Mx\|_2 = \sigma_{\max}(M)$ where $\sigma_{\max}(M)$ is the maximum singular value of M .

Lemma 1. *Any optimal rank-one approximation with respect to the Frobenius norm of a matrix M for which $\min(M) \leq -\|M_+\|_F$ contains at least one nonpositive entry.*

Proof. If $M = \mathbf{0}$, the result is trivial. If not, we have $\min(M) < 0$ since $\min(M) \leq -\|M_+\|_F$. Suppose now $(v, w) > 0$ is a best rank-one approximation of M . Therefore, since the negative values of M are approximated by positive ones and since M has at least one negative entry, we have

$$\|M - vw\|_F^2 > \|M_-\|_F^2. \tag{2.5}$$

By the Eckart-Young theorem, the optimal rank-one approximation vw must satisfy

$$\|M - vw\|_F^2 = \|M\|_F^2 - \sigma_{\max}(M)^2 = \|M\|_F^2 - \|M\|_2^2.$$

Clearly,

$$\|M\|_F^2 = \|M_+\|_F^2 + \|M_-\|_F^2 \quad \text{and} \quad \|M\|_2^2 \geq \min(M)^2$$

so that we can write

$$\|M - vw\|_F^2 \leq \|M_+\|_F^2 + \|M_-\|_F^2 - \min(M)^2 \leq \|M_-\|_F^2$$

which is in contradiction with (2.5). \square

We will need to use the following well-known result concerning (unconstrained) low-rank approximations (see, e.g., [17, p. 29]).

Lemma 2. *The local minima of the best rank-one approximation problem with respect to the Frobenius norm are global minima.*

We can now state the main result about the equivalence of (R1NF-MB) and (MB).

Theorem 1. *For $d \geq \sqrt{|E|}$, any optimal solution (v, w) of (R1NF-MB) coincides with an optimal solution of (MB), i.e., vw is binary and $vw \leq M_b$.*

Proof. We focus on the entries of vw which are positive and define their support as

$$K = \left\{ i \in \{1, 2, \dots, m\} \mid v_i > 0 \right\} \quad \text{and} \quad L = \left\{ j \in \{1, 2, \dots, n\} \mid w_j > 0 \right\}. \quad (2.6)$$

We also define $v' = v(K)$, $w' = w(L)$ and $M'_d = M_d(K, L)$ to be the subvector and submatrix with indexes in K , L and $K \times L$. Since (v, w) is optimal for M_d , (v', w') must be optimal for M'_d . Suppose there is a $-d$ entry in M'_d , then

$$\min(M'_d) = -d \leq -\sqrt{|E|} = -\|(M_d)_+\|_F \leq -\|(M'_d)_+\|_F,$$

so that Lemma 1 holds for M'_d . Since (v', w') is positive (i.e., it is located inside the feasible domain) and is an optimal solution of (R1NF-MB) for M'_d , (v', w') is a local minimum of the unconstrained problem, i.e., the problem of best rank-one approximation. By Lemma 2, this must be a global minimum. This is a contradiction with Lemma 1: (v', w') should contain at least one nonpositive entry. Therefore M'_d does not contain any $-d$ entry, and we have $M'_d = \mathbf{1}_{|K| \times |L|}$ which implies that $v'w' = M'_d$ by optimality (it is the unique rank-one solution $v'w'$ with objective value equals to zero) and finally allows to conclude that vw is binary and $vw \leq M_b$. \square

We have just proven the following theorem:

Theorem 2. *Rank-one nonnegative factorization (R1NF) is NP-hard.*

3 Stationary Points of (R1NF-MB)

We have shown that optimal solutions of (R1NF-MB) coincide with optimal solutions of (MB) for $d \geq \sqrt{|E|}$, whose computation is NP-hard. In this section, we focus on stationary points of (R1NF-MB) instead: we show how they are related to the feasible solutions of (MB). This result will be used in Section 4 to design a new type of biclique finding algorithm.

3.1 Definitions and Notations

The pair (v, w) is a stationary point for problem (R1NF-MB) if and only if it satisfies its first-order optimality conditions, i.e. if and only if

$$v \geq \mathbf{0}, \quad \mu = (vw - M_d)w^T \geq \mathbf{0} \quad \text{and} \quad v \circ \mu = \mathbf{0}, \quad (3.1)$$

$$w \geq \mathbf{0}, \quad \lambda = v^T(vw - M_d) \geq \mathbf{0} \quad \text{and} \quad w \circ \lambda = \mathbf{0}, \quad (3.2)$$

where \circ denotes the component-wise multiplication. Of course, we are only interested in nontrivial solutions and, assuming that $v \neq \mathbf{0}$ and $w \neq \mathbf{0}$, one can check that conditions (3.1)-(3.2) are equivalent to

$$v = \max\left(\mathbf{0}, \frac{M_d w^T}{\|w\|_2^2}\right) \quad \text{and} \quad w = \max\left(\mathbf{0}, \frac{v^T M_d}{\|v\|_2^2}\right). \quad (3.3)$$

We define three sets of rank-one matrices:

1. Given a positive real number d , S_d is the set of nontrivial stationary points of (R1NF-MB), i.e.⁶

$$S_d = \{vw \in \mathbb{R}_0^{m \times n} \mid (v, w) \text{ satisfies (3.3)}\};$$

2. F is the set of feasible solutions of (MB), i.e.

$$F = \{vw \in \mathbb{R}^{m \times n} \mid (v, w) \text{ is a feasible for (MB)}\},$$

3. B is the set of maximal bicliques of (MB), i.e., $vw \in B$ if and only if $vw \in F$ and vw coincides with a maximal biclique.

3.2 Stationarity of Maximal Bicliques

The next theorem states that, for d sufficiently large, the only nontrivial feasible solutions of (MB) that are stationary points of (R1NF-MB) are the maximal bicliques.

Theorem 3. For $d > \max(m, n) - 1$, $F \cap S_d = B$.

Proof. Let show that $vw \in B$ if and only if $vw \in F$ and $vw \in S_d$. By definition, vw belongs to B if and only if vw belongs to F and is maximal, i.e.,

$$(*) \quad \nexists i \text{ such that } v_i = 0 \text{ and } M_d(i, j) = 1, \forall j \text{ s.t. } w_j \neq 0,$$

$$(**) \quad \nexists j \text{ such that } w_j = 0 \text{ and } M_d(i, j) = 1, \forall i \text{ s.t. } v_i \neq 0.$$

Since vw is binary and $v \neq \mathbf{0}$, the nonzero entries of w must be equal to each other. Noting L the support of w (see Equation (2.6)), we then have

$$w_j = \frac{\|w\|_1}{|L|} = C, \quad \forall j \in L,$$

for some $C \in \mathbb{R}_+$, where $\|x\|_1 = \sum_{i=1}^n |x_i|$ for $x \in \mathbb{R}^n$. Moreover, $d > \max(m, n) - 1$ so that $(*)$ is equivalent to

$$\begin{aligned} & \nexists i \quad \text{such that} \quad v_i = 0 \quad \text{and} \quad M_d(i, :)w^T > 0 \\ & \iff \\ & v_i = 0 \Rightarrow M_d(i, :)w^T \leq 0 \quad \text{and} \quad v_i \neq 0 \Rightarrow v_i = \frac{1}{C} = \frac{\|M_d(i, :)\|_1}{\|w\|_1} = \frac{M_d(i, :)w^T}{\|w\|_2^2}. \end{aligned}$$

⁶ $\mathbb{R}_0 = \mathbb{R} \setminus \{0\}$

These are exactly the stationarity conditions for $v \neq \mathbf{0}$, cf. (3.3). By symmetry, (***) is equivalent to the stationarity conditions for w , so that we can conclude that $vw \in B$ if and only if $vw \in F$ and $vw \in S_d$. \square

Theorem 3 implies that, for d sufficiently large, $B \subset S_d$. It would be interesting to have the converse affirmation, i.e. to show that for d sufficiently large, any stationary point of (R1NF-MB) corresponds to a maximal biclique of (MB). As we will see later, this property unfortunately does not hold. However, the following slightly weaker result can be proved: as d goes to infinity, the points in S_d get closer and closer to feasible solutions of (MB), i.e. to bicliques of the graph G_b . As a consequence, rounding stationary points of (R1NF-MB) for d sufficiently large will generate bicliques of G_b .

3.3 Limit Points of S_d

Lemma 3. *The set S_d is bounded, i.e., $\forall d > 0, \forall vw \in S_d$:*

$$\|vw\|_2 = \|v\|_2 \|w\|_2 \leq \sqrt{|E|}.$$

Proof. For any $vw \in S_d$, we have by (3.3)

$$\|v\|_2 = \left\| \max \left(\mathbf{0}, \frac{M_d w^T}{\|w\|_2^2} \right) \right\|_2 \leq \frac{\|\max(\mathbf{0}, M_d) w^T\|_2}{\|w\|_2^2} \leq \frac{\|\max(\mathbf{0}, M_d)\|_F}{\|w\|_2} = \frac{\sqrt{|E|}}{\|w\|_2}.$$

\square

Lemma 4. *For $vw \in S_d$, if $M_d(i, j) = -d$ and if $(vw)_{ij} > 0$, then*

$$0 < v_i < \frac{\|v\|_1}{d+1} \quad \text{and} \quad 0 < w_j < \frac{\|w\|_1}{d+1}.$$

Proof. By (3.3), we have

$$0 < w_j \|v\|_2^2 = v^T M_d(:, j) \leq \|v\|_1 - (d+1)v_i \Rightarrow 0 < v_i < \frac{\|v\|_1}{d+1}.$$

The corresponding result for w is obtained similarly. \square

Theorem 4. *As d goes to infinity, stationary points of (R1NF-MB) get arbitrarily close to feasible solutions of (MB), i.e., $\forall \epsilon > 0, \exists D$ s.t. $\forall d > D$:*

$$\max_{vw \in S_d} \min_{v_b w_b \in F} \|vw - v_b w_b\|_F < \epsilon. \quad (3.4)$$

Proof. Let $vw \in S_d$. We can assume w.l.o.g. that $vw > 0$; otherwise, we consider the subproblem with the vectors $v(K)$ and $v(L)$ where K (resp. L) is the support of v (resp. w) and the matrix $M(K, L)$, see Equation (2.6). In fact, it is clear that if $(v(K), w(L))$ is close to a feasible solution of (MB) for $M_b(K, L)$, then (v, w) is for M_b . We also assume w.l.o.g. that $\|w\|_2 = 1$; in fact, if $vw \in S_d$, $(\lambda v \frac{1}{\lambda} w) \in S_d, \forall \lambda > 0$. Note that Lemma 3 implies $\|v\|_2 \leq \sqrt{|E|}$. By (3.3),

$$v = M_d w^T \quad \text{and} \quad w = \frac{v^T M_d}{\|v\|_2^2}. \quad (3.5)$$

Therefore, $(v/\|v\|_2, w) > 0$ is a pair of singular vectors of M_d associated with the singular value $\|v\|_2 > 0$. If $M_d = \mathbf{1}_{m \times n}$, the only pair of positive singular vectors of M_d is $(\frac{1}{\sqrt{m}}\mathbf{1}_m, \frac{1}{\sqrt{n}}\mathbf{1}_n)$ so that $vw = M_b$ coincides with a feasible solution of (MB).

Otherwise, when $M_d \neq \mathbf{1}_{m \times n}$, we define

$$A = \left\{ i \mid M_d(i, j) = 1, \forall j \right\} \quad \text{and} \quad B = \left\{ j \mid M_d(i, j) = 1, \forall i \right\}, \quad (3.6)$$

and their complements $\bar{A} = \{1, 2, \dots, m\} \setminus A$, $\bar{B} = \{1, 2, \dots, n\} \setminus B$; hence,

$$M_d(A, \cdot) = \mathbf{1}_{|A| \times n} \quad \text{and} \quad M_d(\cdot, B) = \mathbf{1}_{m \times |B|}.$$

These two sets clearly define the biclique $A \times B$ in graph G_b , or, equivalently, a (binary) feasible solution (\bar{v}_A, \bar{w}_B) for problem (MB), where \bar{v}_A is equal to one for indices in A and to zero otherwise (similarly for \bar{w}_B and B). We are now going to show that, for d sufficiently large, vw is arbitrarily close to $\bar{v}_A \bar{w}_B$, which will prove our claim.

Using Lemma 4 and the fact that $\|x\|_1 \leq \sqrt{n}\|x\|_2, \forall x \in \mathbb{R}^n$, we get

$$\mathbf{0} < v(\bar{A}) < \frac{\sqrt{m|E|}}{d+1} \mathbf{1}_{|\bar{A}|} \quad \text{and} \quad \mathbf{0} < w(\bar{B}) < \frac{\sqrt{n}}{d+1} \mathbf{1}_{|\bar{B}|}. \quad (3.7)$$

Therefore, since $\|w\|_2 = 1$ and $\|v\|_2 \leq \sqrt{|E|}$, we obtain

$$\|v(\bar{A})w - \mathbf{0}\|_F = \|v(\bar{A})\|_2 \|w\|_2 < \frac{1}{d+1} \left(m\sqrt{|E|} \right), \quad \text{and} \quad (3.8)$$

$$\|vw(\bar{B}) - \mathbf{0}\|_F = \|v\|_2 \|w(\bar{B})\|_2 < \frac{1}{d+1} \left(n\sqrt{|E|} \right). \quad (3.9)$$

It remains to show that $v(A)w(B)$ coincides with a biclique of the (complete) graph generated by $M_b(A, B) = \mathbf{1}_{|A| \times |B|}$ since $v(\bar{A})w$ and $vw(\bar{B})$ tend to zero as d goes to infinity.

Noting $k_w = \frac{\|v\|_1}{\|v\|_2^2}$ and using (3.5), we get $w(B) = k_w \mathbf{1}_{|B|}$. Combining this with (3.7) gives

$$1 - |\bar{B}| \frac{\sqrt{n}}{d+1} < \|w\|_2^2 - \|w(\bar{B})\|_2^2 = \|w(B)\|_2^2 = |B|k_w^2 \leq \|w\|_2^2 = 1. \quad (3.10)$$

Moreover, (3.5) implies $v(A) = \mathbf{1}_{|A| \times m} w^T = \|w\|_1 \mathbf{1}_{|A|}$ so that

$$|B|k_w \leq v(A) = (\|w(B)\|_1 + \|w(\bar{B})\|_1) \mathbf{1}_{|A|} < |B|k_w + |\bar{B}| \frac{\sqrt{n}}{d+1}. \quad (3.11)$$

Finally, multiplying (3.11) by k_w , combining it with (3.10) and noting that, since $\|w\|_2 = 1$, we have $k_w \leq 1$, we obtain

$$\left(1 - \frac{|\bar{B}|\sqrt{n}}{d+1} \right) \mathbf{1}_{|A| \times |B|} < v(A)w(B) < \left(1 + \frac{|\bar{B}|\sqrt{n}}{d+1} \right) \mathbf{1}_{|A| \times |B|}. \quad (3.12)$$

We can conclude that, for d sufficiently large, vw is arbitrarily close to a feasible solution $\bar{v}_A \bar{w}_B$ of (MB) which corresponds to the biclique (A, B) . \square

3.4 Example

Let

$$M_b = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad M_d = \begin{pmatrix} -d & 1 \\ 1 & 1 \end{pmatrix}.$$

Clearly, $\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$ belongs to the set B , i.e., it corresponds to a maximal biclique of the graph generated by M_b . By Theorem 3, for $d > 1$, it belongs to S_d , i.e., $[(1 \ 1)^T, (0 \ 1)]$ is a stationary point of (R1NF-MB).

For $d > 1$, one can also check that the singular values of M_d are disjoint and that the second pair of singular vectors is positive. Since it is a positive stationary point of the unconstrained problem, it is also a stationary point of (R1NF-MB). As d goes to infinity, it must get closer to a biclique of (MB) (Theorem 4). Moreover M_d is symmetric so that the right and left singular vectors are equal to each other. Figure 1 shows the evolution⁷ with respect to d of this positive singular vector (v_1, v_2) , which is such that $(v_1 \ v_2)^T (v_1 \ v_2) \in S_d$. It converges to $(0 \ 1)$, which means that the outer product of the left and right singular vectors converges to $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$, which is a biclique, i.e. a member of F . We also note that this biclique is not maximal, which shows that the converse to Theorem 3 is false, even asymptotically as d goes to infinity.

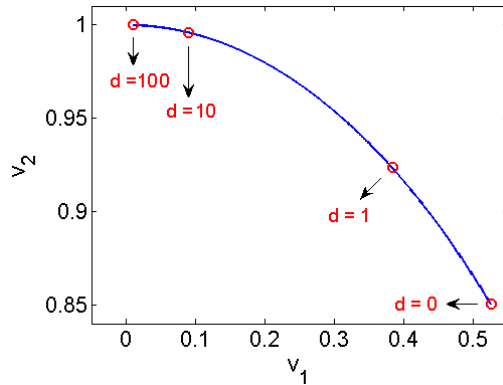


Figure 1: Evolution of (v_1, v_2) .

Corollary 1. For

$$d \geq 2\max(m, n)\sqrt{|E|}, \quad (3.13)$$

any stationary point $vw \in S_d$ of (R1NF-MB) can be rounded⁸ to generate a biclique of the graph G_b generated by M_b .

Proof. The condition

$$\max_{vw \in S_d} \min_{v_b w_b \in F} \max_{ij} (vw - v_b w_b)_{ij} < \frac{1}{2},$$

is clearly sufficient to guarantee that rounding any stationary point of (R1NF-MB) will generate a biclique of G_b . Looking back at Theorem 4, one can check that this is satisfied (cf. Equations (3.8), (3.9) and (3.12)) for d given by (3.13) (note that w.l.o.g. $|E| \geq \max(m, n)$, i.e., that each row and each column of M_b has at least one nonzero entry, otherwise they can be removed). \square

⁷By Wedin's theorem (cf. matrix perturbation theory [28]), singular subspaces of M_d associated with a positive singular value are continuously deformed with respect to d .

⁸Values smaller than 0.5 are set to 0, and set to 1 otherwise.

4 Biclique Finding Algorithm

Many real world applications rely on the discovery of maximal biclique subgraphs, e.g., web community discovery, biological data analysis, text mining, ... [24]. Some algorithms aim at detecting all the maximal bicliques, which is computationally challenging. In fact, there might be an exponential number of such bicliques and the problem is at least NP-hard since it would solve (MB), cf. [1] and references therein. For large datasets, it is in general hopeless to extract all the maximal bicliques in a reasonable computational time. Therefore, one can be interested in finding only large maximal bicliques, which is what we focus on in this section.

For example, a recent data analysis technique called binary matrix factorization (BMF) aims at expressing a binary matrix M as the product of two binary matrices [31, 25, 30]. Each rank-one factor of the decomposition corresponds to a bicluster in the bipartite graph G_b generated by M . Finding bicliques in G allows to solve BMF recursively, since bicliques of G correspond to binary rank-one underapproximations of M [15] (see also the formulation of the optimization problem (MB)).

In this section, we present a heuristic scheme designed to find large bicliques in a given graph, whose main iteration requires a number of operations proportional to the number of edges $|E|$ in the graph. It is based on the reduction of the maximum edge biclique problem to (R1NF-MB) (Theorems 1, 3 and 4). We compare its performance on random graphs and text mining datasets with two other algorithms requiring $O(|E|)$ operations per iteration.

4.1 Description

For d sufficiently large, stationary points of (R1NF-MB) are close to bicliques of (MB) (Corollary 1). Since (R1NF-MB) is a continuous optimization problem, any standard nonlinear optimization technique can in principle be used to compute such a stationary point. One can therefore think of applying an algorithm that finds a stationary point of (R1NF-MB) in order to localize a large biclique of the graph generated by M_b . Moreover, since the two problems have the same objective function, stationary points with larger objective functions will correspond to larger bicliques.

Of course, solving (R1NF-MB) up to global optimality, i.e. finding the best stationary point, is as hard as solving (MB). However, one can hope that the nonlinear optimization scheme used will converge to a relatively large biclique of G_b (i.e. with an objective function close to the global optimum) ; this hope will be confirmed empirically later in this section.

We choose to use the coordinate descent method presented earlier, i.e. solve alternatively the problem in the variable v for w fixed, then in the variable w for v fixed, since the optimal solutions for each of these steps can be written in closed form, cf. Equation (3.3). We also propose, instead of fixing the value of parameter d to the value recommended by Corollary 1, to start with a lower initial value d_0 and gradually increase it (with a multiplicative factor $\gamma > 1$) until it reaches the upper bound D equal to the recommended value. Convergence of the resulting scheme, Algorithm BF-NF, is proved in the next Theorem.

Theorem 5. *The rounding of every limit point of Algorithm BF-NF generates a biclique of G_b , the bipartite graph generated by M_b .*

Proof. When an exact two-block coordinate descent is applied to an optimization problem with a continuously differentiable objective function and a feasible domain equal to the Cartesian product of two closed convex sets (i.e. the two blocks correspond to \mathbb{R}_+^m and \mathbb{R}_+^n in this case), every limit point of the iterates is a stationary point [16].

After a finite number of steps of Algorithm BF-NF, parameter d attains the upper bound $D = 2\max(m, n)|E|$ and no longer changes, so that we can invoke this result and, using Corollary 1, guarantee that the resulting limit points can be rounded to generate a feasible solution of (MB), i.e. a biclique of G_b . \square

Algorithm BF-NF Biclique Finding Algorithm based on Nonnegative Factorization

Require: Bipartite graph $G_b = (V, E)$ described by biadjacency matrix $M_b \in \{0, 1\}^{m \times n}$, initial values $w_0 \in \mathbb{R}_{++}^n$ and $d_0 > 0$, parameter $\gamma > 1$.

- 1: **Set** parameter $D = 2\max(m, n)|E|$ and **initialize** variables $d \leftarrow d_0$, $w \leftarrow w_0$
- 2: **for** $k = 1, 2, \dots$ **do**
- 3:

$$v \leftarrow (1 + d)M_b w^T - d\|w\|_1; \quad (4.1)$$

$$v \leftarrow v / \max(v);$$

$$w \leftarrow \frac{(1 + d)v^T M_b - d\|v\|_1}{\|v\|_2^2}; \quad (4.2)$$

$$d \leftarrow \min(\gamma d, D);$$

4: **end for**

Note that the normalization of v ($v \leftarrow v / \max(v)$) performed by Algorithm BF-NF only changes the scaling of the solution vw and allows (v, w) to converge to binary vectors. Finally, one can easily check that Algorithm BF-NF requires only $O(|E|)$ operations per iteration, the main cost being the computation of the matrix-vector products $M_b w^T$ and $v^T M_b$ (the rest of an iteration requiring only $O(\max(m, n))$ operations).

Parameters

It is not clear a priori how the initial value d_0 should be selected. We observed that it should not be chosen too large: otherwise, the algorithm often converges to the trivial solution: the empty biclique. In fact, in that case, the negative terms ($d\|w\|_1$ and $d\|v\|_1$) in (4.1) and (4.2) will dominate, even during the initial steps of the algorithm, and the solution will be set to zero⁹.

On the other hand, the algorithm with $d = 0$ is equivalent to the power method applied to M_b , and then converges (under some mild assumptions) to the best rank-one approximation of M_b [14]. Hence we observed that when d_0 is chosen too small, the iterates will in general converge to the same solution.

In order to balance positive and negative entries in M_d , we found appropriate to choose an initial value of d such that $\|(M_d)_+\|_F \approx \|(M_d)_-\|_F$, i.e.,

$$d_0 \approx \frac{\|M_b\|_F}{\sqrt{|Z|}} = \sqrt{\frac{|E|}{|Z|}}, \quad (4.3)$$

(recall $|Z|$ is the number of zero entries in M_b). For our tests we chose $d_0 = 2\sqrt{\frac{|E|}{|Z|}}$, which appears to work well in practice.

Finally, the algorithm does not seem to be very sensitive to multiplicative factor γ and selecting values around 1.1 gives good results; this value will be used for the computations below.

4.2 Other Algorithms in $O(|E|)$ Operations

We briefly present here two other algorithms designed to find large bicliques using $O(|E|)$ operations per iteration.

⁹In practice, we used a safety procedure which reduces the value of d whenever v (resp. w) is set to zero and reinitializes v (resp. w) to its previous value.

Greedy Heuristic

The simplest heuristic one can imagine is to add, at each step, the vertex which is connected to the most vertices in the other side of the bipartite graph. Once a vertex is selected, the vertices which are not connected to the chosen vertex are deleted. The procedure is repeated on the remaining graph until one obtains a biclique, which is then necessarily maximal.

Motzkin-Strauss Formalism

In [12], Ding and co-authors extend the generalized Motzkin-Strauss formalism, defined for cliques, to bicliques by defining the optimization problem

$$\max_{\mathbf{x} \in F_x^\alpha, \mathbf{y} \in F_y^\beta} \mathbf{x}^T M_b \mathbf{y}$$

where $F_x^\alpha = \{x \in \mathbb{R}_+^n \mid \sum_{i=1}^n x_i^\alpha = 1\}$, $F_y^\beta = \{y \in \mathbb{R}_+^n \mid \sum_{i=1}^n y_i^\beta = 1\}$ and $1 < \alpha, \beta \ll 2$. Multiplicative updates for this problem are then provided:

$$\mathbf{x} \leftarrow \left(\mathbf{x} \circ \frac{M_b \mathbf{y}}{\mathbf{x}^T M_b \mathbf{y}} \right)^{\frac{1}{\alpha}}, \quad \mathbf{y} \leftarrow \left(\mathbf{y} \circ \frac{M_b^T \mathbf{x}}{\mathbf{x}^T M_b \mathbf{y}} \right)^{\frac{1}{\beta}}. \quad (\text{MS})$$

This algorithm does not necessarily converge to a biclique: if α and β are not sufficiently small, it may only converge to a dense bipartite subgraph (a bicluster). In particular, for $\alpha = \beta = 2$, it converges to an optimal rank-one solution of the unconstrained problem, as Algorithm BF-NF does for $d = 0$. For our tests, we chose $\alpha = \beta = 1.05$ as recommended in [12].

In order to evaluate the quality of the solutions provided by this algorithm when it did not converge to a biclique, we used the following two post-processing procedures to convert a bicluster into a biclique:

1. Greedy (MS): extract from the generated bicluster a biclique using the greedy heuristic presented above.
2. Recursive (MS): use the algorithm recursively on the extracted bicluster, i.e. rerun it on the positive submatrix while decreasing the values of parameters α and β with $\alpha \leftarrow 1 + \frac{\alpha-1}{2}$ and $\beta \leftarrow 1 + \frac{\beta-1}{2}$.

4.3 Results

Synthetic Data

We first present numerical experiments with random graphs: for each density (0.1, 0.3, 0.5, 0.7 and 0.9), 100 bipartite graphs with 200 vertices (100 on each side, i.e., $m = n = 100$) were randomly generated (the probability that an edge belongs to the graph is equal to the density). We then performed, for each graph, 100 runs with the same random initializations and each algorithm was allotted 100 iterations, except for the greedy heuristic which was always run until completion and only once for each graph (since it does not require a random initialization). Actual amounts of CPU time spent by Algorithms MS and BF-NF were comparable, as expected from their similar iteration complexity, while the greedy heuristic was faster.

Figure 2 displays the performance profile for these experiments [13], where the performance function at $\rho \leq 1$ is defined as the percentage, among all graphs and all runs, of bicliques whose sizes (i.e. number of edges) is larger than ρ times the size the largest biclique found by any algorithm in the corresponding graph, i.e.,

$$\text{performance}(\rho) = \frac{\#\{\text{bicliques} \mid \text{size} \geq \rho \times \text{size of best biclique found}\}}{\#\text{runs}}.$$

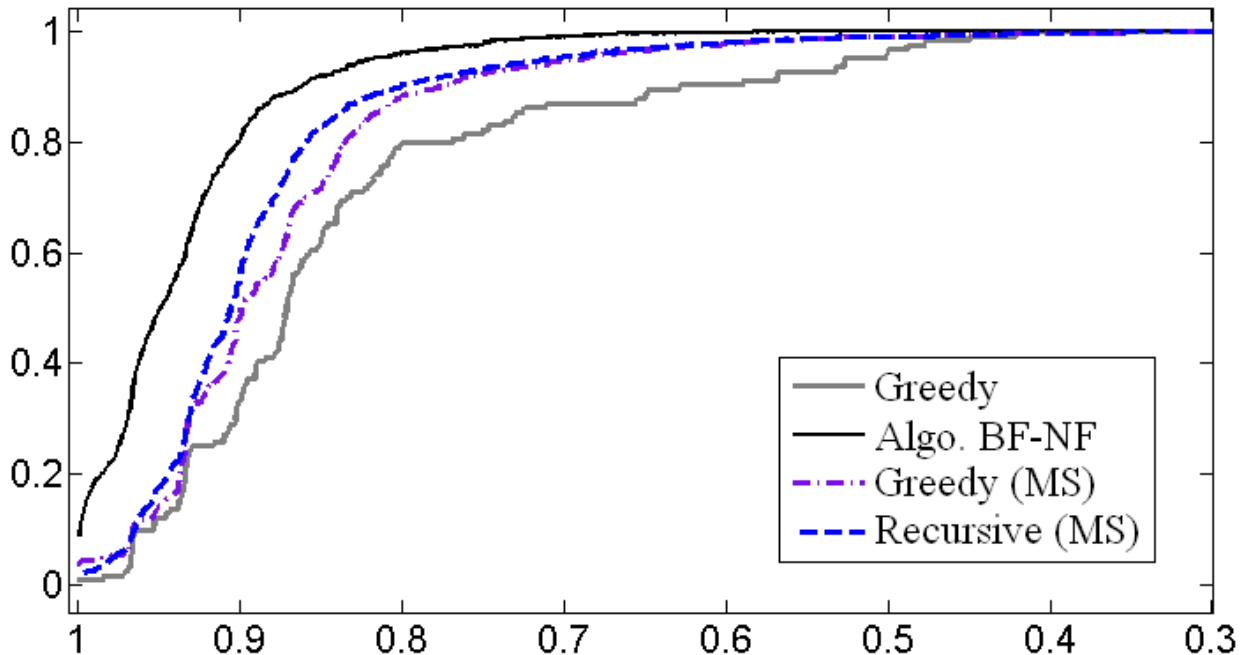


Figure 2: Performance profile for random graphs (densities from 0.1 to 0.9).

On such a performance profile, the higher the curve, the better ; more specifically, the left part of the graph measures efficiency, i.e. how often a given algorithm produces the best biclique among its peers, while the right part estimates robustness, i.e. how far from the best non-optimal solutions are. These two aspects are also reported more quantitatively in Table 1, which displays the value of the performance function at $\rho = 1$ (Efficiency, i.e. how often a given algorithm find a biclique with largest size) and the smallest value of ρ such that the performance function is equal to 100% (Robustness, i.e. the relative size of the worst biclique found).

We observe on the performance profile that both Algorithm BF-NF and (MS) perform better than the greedy heuristic. The variant of (MS) using recursive post-processing performs slightly better than the one based on the use of the greedy heuristic. Nevertheless, Algorithm BF-NF generates in general better solutions: it is more efficient (9% of its solutions are ‘optimal’, twice better than the greedy (MS)) and more robust (all solutions are at most a factor 0.56 away from the best solution, better than 0.42 and 0.30 of other algorithms).

Densities	Greedy	Algo. 1	Greedy M.-S.	Recursive M.-S.
Both (Fig. 2)	1% — 0.42	9% — 0.56	4% — 0.30	2% — 0.30
Sparse (Fig. 3)	0% — 0.33	24% — 0.39	14% — 0.28	14% — 0.28
Dense (Fig. 3)	2% — 0.76	16% — 0.80	6% — 0.68	2% — 0.70

Table 1: Efficiency — Robustness.

It is worth noting that the algorithms behave quite differently on sparse and dense graphs. Using the same setting as before, Figure 3 displays performance profiles for sparse graphs (on the left, with densities 0.05, 0.1, 0.15 and 0.2) and dense graphs (on the right, with densities 0.8, 0.85, 0.9 and 0.95). For sparse graphs, both versions of (MS) seem to coincide and the greedy heuristic performs significantly worse. For dense graphs, the greedy heuristic coincides with the greedy (MS) and performs almost as well as the recursive (MS). However, in all cases, Algorithm BF-NF performs better. It

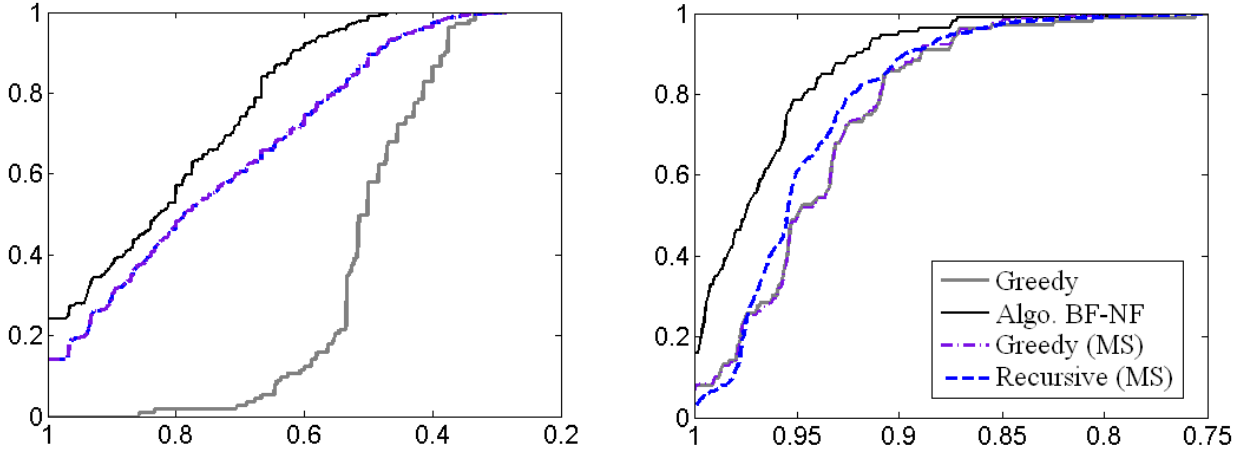


Figure 3: Performance profiles for random graphs: sparse (left, from 0.05 to 0.2) and dense (right, from 0.8 to 0.95).

is more efficient: it finds the best solution in 24% (resp. 16%) of the runs for sparse (resp. dense) graphs while (MS) only achieves 14% (resp. 6%) and the greedy heuristic 0% (resp. 2%). It is also more robust: all solutions are at most a factor 0.39 (resp. 0.80) away from the best solution for sparse (resp. dense) graphs, bigger than the best factor 0.33 (resp. 0.76) of the other algorithms.

Text Datasets

If parameter D in Algorithm BF-NF is chosen smaller than the value recommended by Corollary 1, the algorithm is no longer guaranteed to converge to a biclique. However, the negative entries in M_d will force the corresponding entries of the solutions of (R1NF-MB) to be small (cf. Theorem 4). Therefore, instead of a biclique, one gets a dense submatrix of M_b , i.e., a *bicluster*. Algorithm BF-NF can then be used as a *biclustering algorithm* and the density of the corresponding submatrix will depend on the choice of parameter D between 0 and $2 \max(m, n)|E|$. We test this approach on the six text mining datasets (with sparse matrices) described in Table 2.

Data	m	n	$ E $	sparsity
classic	7094	41681	223839	99.92
sports	8580	14870	1091723	99.14
reviews	4069	18483	758635	98.99
hitech	2301	10080	331373	98.57
ohscal	11162	11465	674365	99.47
la1	3204	31472	484024	99.52

Table 2: Text mining datasets [32] (sparsity is given in %: $100 * |Z|/(mn)$).

Figure 4 compares Algorithms BF-NF and MS for varying values of their parameters: for the Motzkin-Strauss formalism, we tested $\alpha = \beta \in [1.3, 1.9]$ with step size 0.025 and, for Algorithm BF-NF, $D \in d_0 10^{[3,9]}$ with step size 0.25 (d_0 given by Equation (4.3)). For each value, we performed 10 runs (same initializations for both algorithms and 500 iterations) and plotted all the non-dominated solutions (i.e., for which no other solution has both larger size and higher density) for each dataset. We observe that our approach consistently generates better results since its curves dominate the ones

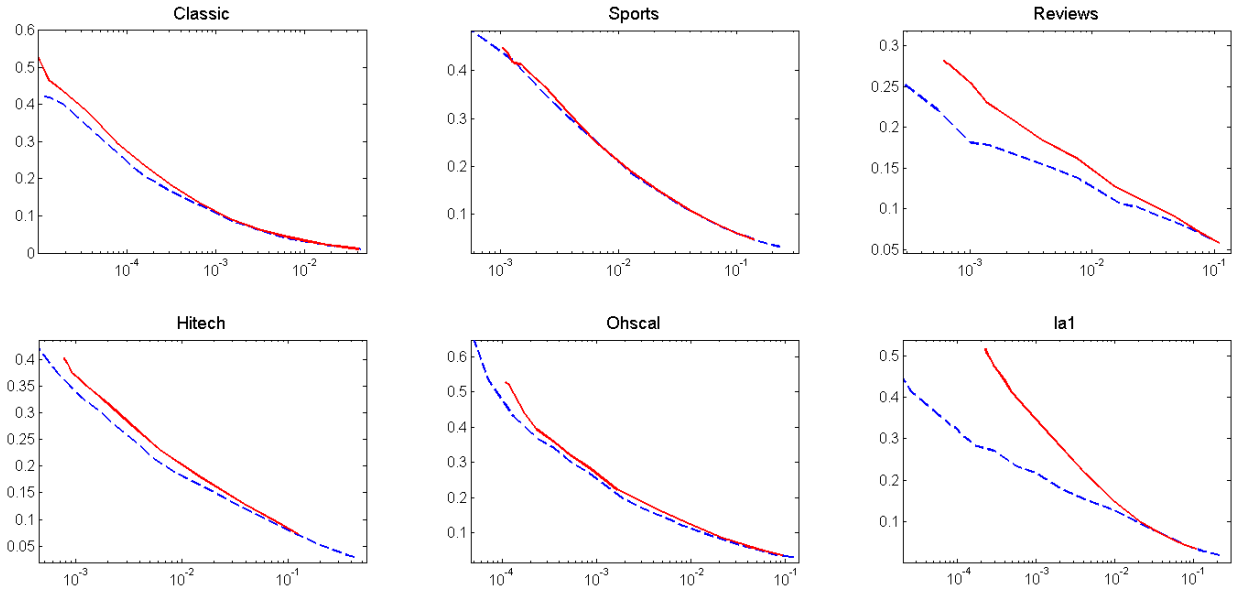


Figure 4: Normalized size vs. density for the Motzkin-Strauss formalism (dashed line) and Algorithm BF-NF based on (R1NF-MB) (solid line). The x-axis indicates the normalized sizes of the extracted clusters (i.e., number of entries in the extracted submatrix divided by the number of entries in the original matrix) while the y-axis indicates the density of these clusters (number of nonzero entries divided by the total number of entries) for the text datasets of Table 2.

of the Motzkin-Strauss formalism, i.e., the biclusters it finds are denser for the same size or larger for the same density.

Finally, we mention that Algorithm BF-NF can be further enhanced in the following ways:

- It is applicable to non-binary matrices, i.e., weighted graphs. Theorem 1 can easily be adapted using $d \geq \|M_+\|_F$ (Lemma 1), and one can show that the resulting algorithm will converge to the optimal rank-one approximation of a positive submatrix of M .
- It is possible to give more weight to a given side of the biclique by adding regularization terms to the cost functions. For example, one can consider the following objective function

$$\min_{v, w \geq 0} \|M_d - vw\|_F^2 + \alpha \|v\|_2^2 + \beta \|w\|_2^2$$

which our algorithm can handle after some straightforward modifications (namely, the optimal solution for v when w is fixed can still be written in closed-form, and vice versa).

- If $M_b \in \{0, 1\}^{n \times n}$ is the adjacency matrix of a (non bipartite) graph $G = (V, E)$ with $V = \{v_1, \dots, v_n\}$, i.e., $M_b(i, j) = 1 \Leftrightarrow (v_i, v_j) \in E$, one can check that formulation (MB) corresponds to the maximum edge biclique problem in any graph. This only requires that the diagonal entries of M_b are set to zero (no self loop in the graph) since a vertex cannot simultaneously belong to both sides of a biclique. Therefore, all the results of this paper are actually valid for not necessarily bipartite graphs.

5 Conclusion

We have introduced Nonnegative Factorization (NF), a generalization of Nonnegative Matrix Factorization (NMF), and proved its NP-hardness in the rank-one case by reduction of the maximum edge

biclique problem. Since finding each rank-one factor in any NMF decomposition implicitly amounts to solving a rank-one NF problem, this suggests that (NMF) is a NP-hard problem for any fixed factorization rank and that no polynomial time algorithm based on the successive optimization of the rank-one factors can be designed, giving more credence to algorithms based on alternating optimization (e.g., the HALS algorithm or the standard alternating nonnegative least squares).

We also presented a heuristic algorithm for detecting large bicliques whose iterations require $O(|E|)$ operations. It is based on results linking stationary points of a specific rank-one nonnegative factorization problem (R1NF-MB) and the maximum edge biclique problem. We experimentally demonstrated its efficiency and robustness on random graphs and text mining datasets.

References

- [1] G. ALEXE, S. ALEXE, Y. CRAMA, S. FOLDES, P.L. HAMMER, AND B. SIMEONE, *Consensus algorithms for the generation of all maximal bicliques*, Discrete Applied Mathematics, 145(1) (2004), pp. 11–21.
- [2] M. BERRY, M. BROWNE, A. LANGVILLE, P. PAUCA, AND R.J. PLEMMONS, *Algorithms and Applications for Approximate Nonnegative Matrix Factorization*, Computational Statistics and Data Analysis, 52 (2007), pp. 155–173.
- [3] M. BIGGS, A. GHODSI, AND S. VAVASIS, *Nonnegative Matrix Factorization via Rank-One Downtdate*, in 25th international conference on machine learning (ICML), 2008.
- [4] C. CICHOCKI AND A-H. PHAN, *Fast local algorithms for large scale Nonnegative Matrix and Tensor Factorizations*, IEICE Transactions on Fundamentals of Electronics, Vol. E92-A No.3 (2009), pp. 708–721.
- [5] C. CICHOCKI, R. ZDUNEK, AND S. AMARI, *Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization*, in ICA07, London, UK, September 9-12, Lecture Notes in Computer Science, Vol. 4666, Springer, pp. 169-176, 2007.
- [6] ———, *Nonnegative Matrix and Tensor Factorization*, IEEE Signal Processing Magazine, (2008), pp. 142–145.
- [7] J.E. COHEN AND U.G. ROTHBLUM, *Nonnegative ranks, Decompositions and Factorization of Nonnegative Matrices*, Linear Algebra and its Applications, 190 (1993), pp. 149–168.
- [8] K. DEVARAJAN, *Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology*, PLoS Computational Biology, 4(7), e1000029 (2008).
- [9] I.S. DHILLON AND S. SRA, *Nonnegative Matrix Approximations: Algorithms and Applications*, tech. report, University of Texas (Austin), 2006. Dept. of Computer Sciences.
- [10] C. DING, X. HE, AND H.D. SIMON, *On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering*, in SIAM Int’l Conf. Data Mining (SDM’05), 2005, pp. 606–610.
- [11] C. DING, T. TAO L, AND M.I. JORDAN, *Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding*, Data Mining, IEEE International Conference on, (2008), pp. 183–192.
- [12] C. DING, Y. ZHANG, T. LI, AND S.R. HOLBROOK, *Biclustering Protein Complex Interactions with a Biclique Finding Algorithm*, in Sixth IEEE International Conference on Data Mining, 2006, pp. 178–187.
- [13] E.D. DOLAN AND J.J. MORÉ, *Benchmarking optimization software with performance profiles*, Math. Program., Ser. A, 91 (2002), pp. 201–213.
- [14] 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.
- [15] N. GILLIS AND F. GLINEUR, *Using underapproximations for sparse nonnegative matrix factorization*, Pattern Recognition, 43(4) (2010), pp. 1676–1687.
- [16] L. GRIPPO AND M. SCIANDRONE, *On the convergence of the block nonlinear Gauss-Seidel method under convex constraints*, Operations Research Letters, 26 (2000), pp. 127–136.

- [17] N.-D. HO, *Nonnegative Matrix Factorization - Algorithms and Applications*, PhD thesis, Université catholique de Louvain, 2008.
- [18] P.O. HOYER, *Nonnegative Matrix Factorization with Sparseness Constraints*, J. Machine Learning Research, 5 (2004), pp. 1457–1469.
- [19] H. KIM AND H. PARK, *Non-negative Matrix Factorization Based on Alternating Non-negativity Constrained Least Squares and Active Set Method*, SIAM J. Matrix Anal. Appl., 30(2) (2008), pp. 713–730.
- [20] J. KIM AND H. PARK, *Sparse Nonnegative Matrix Factorization for Clustering*, Tech. Report GT-CSE-08-01, Georgia Institute of Technology, 2008.
- [21] D.D. LEE AND H.S. SEUNG, *Learning the Parts of Objects by Nonnegative Matrix Factorization*, Nature, 401 (1999), pp. 788–791.
- [22] ———, *Algorithms for Non-negative Matrix Factorization*, In Advances in Neural Information Processing, 13 (2001).
- [23] C.-J. LIN, *Projected Gradient Methods for Nonnegative Matrix Factorization*, Neural Computation, 19 (2007), pp. 2756–2779. MIT press.
- [24] G. LIU, K. SIM, AND J. LI, *Efficient Mining of Large Maximal Bicliques*, Springer Berlin, Lecture Notes in Computer Science, (2006), pp. 437–448.
- [25] P. MIETTINEN, *The boolean column and column-row matrix decompositions*, Data Mining and Knowledge Discovery, 17(1) (2008), pp. 39–56.
- [26] P. PAATERO AND U. TAPPER, *Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics, 5 (1994), pp. 111–126.
- [27] R. PEETERS, *The maximum edge biclique problem is NP-complete*, Discrete Applied Mathematics, 131(3) (2003), pp. 651–654.
- [28] G.W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, San Diego, 1990.
- [29] S.A. VAVASIS, *On the complexity of nonnegative matrix factorization*, SIAM Journal on Optimization, 20 (2009), pp. 1364–1377.
- [30] Z.-Y. ZHANG, T. LI, C. DING, X.W. REN, AND X.-S. ZHANG, *Binary matrix factorization for analyzing gene expression data*, Data Mining and Knowledge Discovery, 20 (1) (2009), pp. 28–52.
- [31] Z.-Y. ZHANG, T. LI, C. DING, AND X.-S. ZHANG, *Binary matrix factorization with applications*, in Proceedings of 2007 IEEE International Conference on Data Mining (ICDM07), 2007.
- [32] S. ZHONG AND J. GHOSH, *Generative model-based document clustering: a comparative study*, Knowledge and Information Systems, 8 (3) (2005), pp. 374–384.

Recent titles

CORE Discussion Papers

- 2010/20. Jeroen V.K. ROMBOUITS and Lars STENTOFT. Multivariate option pricing with time varying volatility and correlations.
- 2010/21. Yassine LEFOUILI and Catherine ROUX. Leniency programs for multimarket firms: The effect of Amnesty Plus on cartel formation.
- 2010/22. P. Jean-Jacques HERINGS, Ana MAULEON and Vincent VANNETELBOSCH. Coalition formation among farsighted agents.
- 2010/23. Pierre PESTIEAU and Grégory PONTIERE. Long term care insurance puzzle.
- 2010/24. Elena DEL REY and Miguel Angel LOPEZ-GARCIA. On welfare criteria and optimality in an endogenous growth model.
- 2010/25. Sébastien LAURENT, Jeroen V.K. ROMBOUITS and Francesco VIOLANTE. On the forecasting accuracy of multivariate GARCH models.
- 2010/26. Pierre DEHEZ. Cooperative provision of indivisible public goods.
- 2010/27. Olivier DURAND-LASSERVE, Axel PIERRU and Yves SMEERS. Uncertain long-run emissions targets, CO₂ price and global energy transition: a general equilibrium approach.
- 2010/28. Andreas EHRENMANN and Yves SMEERS. Stochastic equilibrium models for generation capacity expansion.
- 2010/29. Olivier DEVOLDER, François GLINEUR and Yu. NESTEROV. Solving infinite-dimensional optimization problems by polynomial approximation.
- 2010/30. Helmuth CREMER and Pierre PESTIEAU. The economics of wealth transfer tax.
- 2010/31. Thierry BRECHET and Sylvette LY. Technological greening, eco-efficiency, and no-regret strategy.
- 2010/32. Axel GAUTIER and Dimitri PAOLINI. Universal service financing in competitive postal markets: one size does not fit all.
- 2010/33. Daria ONORI. Competition and growth: reinterpreting their relationship.
- 2010/34. Olivier DEVOLDER, François GLINEUR and Yu. NESTEROV. Double smoothing technique for infinite-dimensional optimization problems with applications to optimal control.
- 2010/35. Jean-Jacques DETHIER, Pierre PESTIEAU and Rabia ALI. The impact of a minimum pension on old age poverty and its budgetary cost. Evidence from Latin America.
- 2010/36. Stéphane ZUBER. Justifying social discounting: the rank-discounting utilitarian approach.
- 2010/37. Marc FLEURBAEY, Thibault GAJDOS and Stéphane ZUBER. Social rationality, separability, and equity under uncertainty.
- 2010/38. Helmuth CREMER and Pierre PESTIEAU. Myopia, redistribution and pensions.
- 2010/39. Giacomo SBRANA and Andrea SILVESTRINI. Aggregation of exponential smoothing processes with an application to portfolio risk evaluation.
- 2010/40. Jean-François CARPANTIER. Commodities inventory effect.
- 2010/41. Pierre PESTIEAU and Maria RACIONERO. Tagging with leisure needs.
- 2010/42. Knud J. MUNK. The optimal commodity tax system as a compromise between two objectives.
- 2010/43. Marie-Louise LEROUX and Gregory PONTIERE. Utilitarianism and unequal longevities: A remedy?
- 2010/44. Michel DENUIT, Louis EECKHOUDT, Ilia TSETLIN and Robert L. WINKLER. Multivariate concave and convex stochastic dominance.
- 2010/45. Rüdiger STEPHAN. An extension of disjunctive programming and its impact for compact tree formulations.
- 2010/46. Jorge MANZI, Ernesto SAN MARTIN and Sébastien VAN BELLEGEM. School system evaluation by value-added analysis under endogeneity.
- 2010/47. Nicolas GILLIS and François GLINEUR. A multilevel approach for nonnegative matrix factorization.
- 2010/48. Marie-Louise LEROUX and Pierre PESTIEAU. The political economy of derived pension rights.
- 2010/49. Jeroen V.K. ROMBOUITS and Lars STENTOFT. Option pricing with asymmetric heteroskedastic normal mixture models.

Recent titles

CORE Discussion Papers - continued

- 2010/50. Maik SCHWARZ, Sébastien VAN BELLEGEM and Jean-Pierre FLORENS. Nonparametric frontier estimation from noisy data.
- 2010/51. Nicolas GILLIS and François GLINEUR. On the geometric interpretation of the nonnegative rank.
- 2010/52. Yves SMEERS, Giorgia OGGIONI, Elisabetta ALLEVI and Siegfried SCHAIBLE. Generalized Nash Equilibrium and market coupling in the European power system.
- 2010/53. Giorgia OGGIONI and Yves SMEERS. Market coupling and the organization of counter-trading: separating energy and transmission again?
- 2010/54. Helmuth CREMER, Firouz GAHVARI and Pierre PESTIEAU. Fertility, human capital accumulation, and the pension system.
- 2010/55. Jan JOHANNES, Sébastien VAN BELLEGEM and Anne VANHEMS. Iterative regularization in nonparametric instrumental regression.
- 2010/56. Thierry BRECHET, Pierre-André JOUVET and Gilles ROTILLON. Tradable pollution permits in dynamic general equilibrium: can optimality and acceptability be reconciled?
- 2010/57. Thomas BAUDIN. The optimal trade-off between quality and quantity with uncertain child survival.
- 2010/58. Thomas BAUDIN. Family policies: what does the standard endogenous fertility model tell us?
- 2010/59. Nicolas GILLIS and François GLINEUR. Nonnegative factorization and the maximum edge biclique problem.

Books

- J. GABSZEWICZ (ed.) (2006), *La différenciation des produits*. Paris, La découverte.
- L. BAUWENS, W. POHLMEIER and D. VEREDAS (eds.) (2008), *High frequency financial econometrics: recent developments*. Heidelberg, Physica-Verlag.
- P. VAN HENTENRYCKE and L. WOLSEY (eds.) (2007), *Integration of AI and OR techniques in constraint programming for combinatorial optimization problems*. Berlin, Springer.
- P.-P. COMBES, Th. MAYER and J-F. THISSE (eds.) (2008), *Economic geography: the integration of regions and nations*. Princeton, Princeton University Press.
- J. HINDRIKS (ed.) (2008), *Au-delà de Copernic: de la confusion au consensus ?* Brussels, Academic and Scientific Publishers.
- J-M. HURIOT and J-F. THISSE (eds) (2009), *Economics of cities*. Cambridge, Cambridge University Press.
- P. BELLEFLAMME and M. PEITZ (eds) (2010), *Industrial organization: markets and strategies*. Cambridge University Press.
- M. JUNGER, Th. LIEBLING, D. NADDEF, G. NEMHAUSER, W. PULLEYBLANK, G. REINELT, G. RINALDI and L. WOLSEY (eds) (2010), *50 years of integer programming, 1958-2008: from the early years to the state-of-the-art*. Berlin Springer.

CORE Lecture Series

- C. GOURIÉROUX and A. MONFORT (1995), *Simulation Based Econometric Methods*.
- A. RUBINSTEIN (1996), *Lectures on Modeling Bounded Rationality*.
- J. RENEGAR (1999), *A Mathematical View of Interior-Point Methods in Convex Optimization*.
- B.D. BERNHEIM and M.D. WHINSTON (1999), *Anticompetitive Exclusion and Foreclosure Through Vertical Agreements*.
- D. BIENSTOCK (2001), *Potential function methods for approximately solving linear programming problems: theory and practice*.
- R. AMIR (2002), *Supermodularity and complementarity in economics*.
- R. WEISMANTEL (2006), *Lectures on mixed nonlinear programming*.