Perturbation of matrices and non-negative rank with a view toward statistical models

Cristiano Bocci
Università di Siena
bocci24@unisi.it

Enrico Carlini
Politecnico di Torino
enrico.carlini@polito.it

Fabio Rapallo
Università del Piemonte Orientale
fabio.rapallo@mfn.unipmn.it

October 19, 2010

Abstract

In this paper we study how perturbing a matrix changes its non-negative rank. We show that the non-negative rank is upper-semi-continuos and we describe some special families of perturbations. We apply our results to the study of statistical models.

Key words: Euclidean topology, Jacobian matrix, mixture models, independence of random variables.

MSC: 15B51, 62H17.

1 Introduction

The rank of a matrix gives the least number of rank one matrices, or dyadic products, needed to write the matrix as a sum of dyads. More precisely a $n \times m$ matrix P such that $\mathrm{rk}(P) = k$ can be written as

$$P = c_1(r_1)^t + \ldots + c_k(r_k)^t \,, \tag{1}$$

where the column vectors c_i and r_i have the proper size. Even if P has non-negative entries, the vectors c_i and r_i, are allowed to have negative entries. If we require the vectors to have non-negative entries the least number of summands is called the non-negative rank of P, namely $\mathrm{rk}_+(P)$. The non-negativity constraints make the situation more complex, and the non-negative rank of a matrix is harder to study than the ordinary rank, see e.g. [CR88]. In general $\mathrm{rk}_+(P) \geq \mathrm{rk}(P)$. Therefore, it could not be possible to decompose a rank k matrix into the sum of exactly k dyadic products $c_i(r_i)^t$, where c_i and r_i are non-negative vectors. The relations between the ordinary rank and the non-negative rank have received an increasing attention in the last years, both from a theoretical and an applied point of view. Some recent references are [BL09], [DLC08], [LC10], [PPP06] and [CR10].

As far as we know, there is no efficient way to compute the non-negative rank of a matrix in the general situation. However, there are many recently proposed algorithms to deal with the analogous problem of non-negative matrix factorization, e.g. see [LS01] or [HVD08] for an application to stochastic matrices.

In this paper we study how the non-negative rank of a matrix is affected by small perturbations of the matrix. This is of particular interest when the matrix arises in Probability of Statistics.

Here, a perturbation is intended in the following topological sense. Given a matrix P we consider a neighborhood of P in the Euclidean metric topology. We call any matrix in the neighborhood a perturbation of P. Clearly this notion is more meaningful and interesting when a small neighborhood is considered and hence matrices close to P are studied.

We show that the non-negative rank is upper-semicontinuos in the Euclidean topology, see Theorem 3.1, and hence it cannot decrease by small perturbations of the matrix. We also produce examples of perturbations preserving the non-negative rank, see Proposition 3.2. Using a Jacobian analytic approach we show that, under suitable conditions, perturbing a matrix leaving the ordinary rank fixed also leaves the non-negative rank unchanged, see Proposition 4.2.

The notion of non-negative rank has also relevant applications in Probability and Statistics. In fact, a probability matrix with dyadic expansion as in Equation (1) belongs to the mixture of k independence models for categorical data. Mixture models play a central role in applied probability, as they are the key tool in modelling partially observed phenomena, see [Agr02] for more details. More recently, mixture models have been considered also in the framework of Algebraic Statistics, a branch of Statistics which makes use of notions and techniques from Computational Algebra and Algebraic Geometry, see [PS05, DSS09, GRRW10].

The paper is structured as follows. In Section 2 we recall some basic notions. In Section 3 and Section 4 we use a topological and analytic approach to study perturbations. In Section 5 we use our results to work out some significant examples. Finally, in Section 6 we show how our results relate to Statistics.

2 Basic facts

In this section, we recall some known fact about the non-negative rank. The definitions and the results presented below will be use throughout the paper.
Non-negative matrices. A non-negative $n \times m$ matrix is a point in $\mathbb{R}^{nm}_{\geq 0}$ where

$$\mathbb{R}^{nm}_{\geq 0} = \{(p_{i,j}) : p_{i,j} \in \mathbb{R}, p_{i,j} \geq 0\}.$$

Stochastic matrices. A stochastic matrix is a non-negative matrix having column sum equal to one. To each non-negative matrix we can associate a stochastic matrix. Denote by $P = [c_1, c_2, \ldots, c_m]$ the set of columns of a nonnegative matrix. Define the scaling factor $\sigma(P)$ by

$$\sigma(P) := \operatorname{diag}\{||c_1||_1, \ldots, ||c_m||_1\}$$

where $||\cdot||_1$ is the 1-norm in $\mathbb{R}^n$, and the pullback map $\theta(A)$ by

$$\theta(A) = A\sigma(A)^{-1}.$$

Remark 2.1. *In Probability, stochastic matrices are defined as the non-negative matrices having row sums equal to one. Here we adopt the convention of normalizing the columns. As the rank and the non-negative rank are clearly invariant under transposition this convention do not affect our results.*

Simplex. The n-simplex in $\mathbb{R}^n$ is

$$\Delta^n = \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n \ : \ x_i \geq 0, \sum_{i=1}^n x_i \leq 1\right\}.$$

Note that a $n \times m$ stochastic matrix P can be seen as a collection of m points in Δ^n. More precisely we consider the map π_n assigning to a matrix the set of its columns, that is $\pi_n(P) = \{c_1, \ldots, c_m\} \subset \Delta^n$. All the points c_i lie on the same face of the n-simplex, which is a $(n-1)$-simplex. Hence, by dropping the last component of each c_i, we have a map π_{n-1} sending P into a collection of m points in Δ^{n-1}.

Non-negative rank. Given a $n \times m$ non-negative matrix P, the non-negative rank of P is the smallest integer k such that

$$P = c_1(r_1)^t + \ldots + c_k(r_k)^t$$

where the vectors $c_i \in \mathbb{R}^n$ and the vectors $r_i \in \mathbb{R}^m$ have non-negative entries. The non-negative rank of the matrix P is denoted with $\mathrm{rk}_+(P)$.

Remark 2.2. *Notice that, for a non-negative matrix P, the following holds:* $\mathrm{rk}_+(P) = \mathrm{rk}_+(\theta(P))$. *Hence, the study of the non-negative rank of stochastic matrices and the study of the non-negative rank of non-negative matrices coincide (see [LC10]).*

Geometry and non-negative rank. Let Z be a set of points in $\Delta^{n-1} \subset \mathbb{R}^{n-1}$. The set Z is a (k, r)-set if k is the minimum integer i such that $Z \subset \Delta^{k-1}$ and r is the minimum integer j such that $Z \subset \Delta^n \cap H_{j-1}$ where H_t is an affine space of dimension t. To apply this to non-negative rank we proceed as follows.

Lemma 2.3. *Let P be a $n \times m$ non-negative matrix and let $Z = \pi_{n-1}\theta(P)$. Then, Z is a (k, r)-set if and only if $rk_+(P) = k$ and $rk(P) = r$.*

Another geometric interpretation is given by the nested polygons problem [GG10]. Given a set of points Z in a r-side convex polygon $\mathcal{P}_r$ does there exist a s-side convex polygon $\mathcal{P}_s$ with $s < r$ such that

$$Z \subset \mathcal{P}_s \subset \mathcal{P}_r?$$

If we consider a $n \times m$ non-negative matrix P we let $Z = \pi_{n-1}(\theta(P))$. Then $\mathcal{P}_r$ is $\Delta^{n-1} \cap H$ where H is the linear span of Z. Then we have the following Lemma.

Lemma 2.4. *$rk_+(P) = s$ if and only if s is the minimal integer such that there exists a s-side convex polygon $\mathcal{P}_s$, with $Z \subset \mathcal{P}_s \subset \mathcal{P}_r$.*

3 Upper-semicontinuity of non-negative rank

In this section we will use the ideas recalled in Section 2 to show that the non-negative rank is upper-semicontiuos in the Euclidean topology.

Given a non-negative matrix $P \in \mathbb{R}^{nm}_{\geq 0}$ and $\epsilon > 0$ define the ball of center P and radius ϵ

$$B(P,\epsilon) = \left\{ N = (n_{i,j}) \in \mathbb{R}^{nm}_{\geq 0} : \sqrt{\sum (p_{i,j} - n_{i,j})^2} < \epsilon \right\}.$$

Theorem 3.1. *Let P be an $n \times m$ matrix of non-negative rank k, then there exists a ball $B(P,\epsilon)$ such that, for all $N \in B(P,\epsilon)$, $rk_+(N) \geq k$.*

Proof. We give a proof by contradiction. Suppose that for all natural numbers r there exists $N(r) \in B(P, \frac{1}{r})$ such that $\mathrm{rk}_+(N(r)) < k$. Clearly, the limit of the sequence $N(r)$ is P. By hypothesis we know that there exist convex polygons $\mathcal{P}(r) \subset \Delta^{n-1}$, each having less than k sides, such that

$$\pi_{n-1}(\theta(N(r))) \subset \mathcal{P}(r).$$

Let the vertices of $\mathcal{P}(r)$ be

$$q_1(r), \ldots, q_h(r) \in \Delta^{n-1}$$

and notice that each sequence $q_i(r)$ has a converging subsequence having limit point $\bar{q}_i \in \Delta^{n-1}$. Thus there exists a h-side limit polygon $\bar{\mathcal{P}} \subset \Delta^{n-1}$. As $h < k$ it is enough to show that $\pi_{n-1}(\theta(P)) \subset \bar{\mathcal{P}}$ to get a contradiction using Lemma 2.4.

Let

$$\pi_{n-1}(\theta(N(r))) = \{c_1(r), \ldots, c_m(r)\}$$

and

$$\pi_{n-1}(\theta(P)) = \{c_1, \ldots, c_m\}$$

and notice that the limit of $c_i(r)$ is c_i. Also notice that for each i we have

$$c_i(r) = \alpha_{i,1}(r)q_1(r) + \ldots + \alpha_{i,h}(r)q_h(r)$$

where the coefficients $\alpha_{i,j}(r)$ vary in the compact set $[0,1]$. Taking the limit we get for each i

$$c_i = \bar{\alpha}_{i,1}\bar{q}_1 + \ldots + \bar{\alpha}_{i,h}\bar{q}_h$$

and hence $c_i \in \bar{\mathcal{P}}$. This completes the proof.

$\square$

Thus, for a matrix M of a given non-negative rank, we have that, in a suitable neighborhood of M the non-negative rank can only increase, i.e. the non-negative rank is upper-semicontinous.

Clearly each neighborhood of a matrix P contains a matrix having the same non-negative rank of P, the matrix P itself. But even more is true.

Proposition 3.2 (Barycentric perturbation)**.** *Let P be a non-negative $n \times m$ matrix. For any $\epsilon > 0$ there exists $N \in B(P, \epsilon)$ such that $N \neq P$ and* $\mathrm{rk}_+(N) = \mathrm{rk}_+(P)$.

Proof. Let P have columns c_i and consider the vector $b = \frac{1}{m} \sum_i c_i$. Roughly speaking b correspond to the barycenter of the points $\pi_{n-1}(\theta(P))$. Then we consider the $n \times m$ matrix N_ϵ having the i-th column defined as

$$c_i + \epsilon(b - c_i).$$

As ϵ increases the points $\pi_{n-1}(\theta(N_\epsilon))$ approach the barycenter of $\pi_{n-1}(\theta(P))$. It is enough to prove the statement for $0 < \epsilon < 1$. Thus, by Lemma 2.4, $\mathrm{rk}_+(N_\epsilon) \leq \mathrm{rk}_+(P)$. Then, the conclusion follows applying Theorem 3.1. $\square$

4 Jacobian approach

Throughout this section we assume $k \leq \min\{n, m\}$ and we let $X_{n\times m,k} \subset \mathbb{R}^{mn}$ be the variety of $n \times m$ matrices of rank at most k. It is well-known that $\dim(X_{n\times m,k}) = k(n + m - k)$.

Consider the map $f : \mathbb{R}^{k(n+m)} \to \mathbb{R}^{mn}$ which sends the point

$$p = (x_{1,1}, \dots, x_{1,n}, y_{1,1}, \dots, y_{1,m}, \dots, x_{k,1}, \dots, x_{k,n}, y_{k,1}, \dots, y_{k,m})$$

to the matrix

$$f(p) = \sum_{i=1}^{k} \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,n} \end{pmatrix} \begin{pmatrix} y_{i,1} & \dots & y_{i,m} \end{pmatrix}$$

Let f_+ be the restriction of f to the non-negative octant $\mathbb{R}^{k(n+m)}_{\geq 0}$. The image of f_+ is the variety $X^+_{n\times m,k}$ of $n \times m$ matrices of non-negative rank at most k. It is clear that $X^+_{n\times m,k} \subset X_{n\times m,k}$.

The local rank of f_+^*, the Jacobian matrix of f_+, gives the local dimension of the image.

Thus, if a matrix $P \in f_+^{-1}(p)$, where p is not on the boundary of the octant $\mathbb{R}_+^{k(n+m)}$ and $f_+^*(P)$ has maximal rank, then there exists a neighborhood of P of matrices of rank at most k and non-negative rank at most k.

Given a point $p \in \mathbb{R}^{k(n+m)}$ with coordinates

$$p = (x_{1,1}, \dots, x_{1,n}, y_{1,1}, \dots, y_{1,m}, \dots, x_{k,1}, \dots, x_{k,n}, y_{k,1}, \dots, y_{k,m})\,;$$

we say that p satisfies property (+) if

1) $(x_{i,1}, \dots, x_{i,n})$, $i = 1, \dots, k$ are linearly independent vectors of $\mathbb{R}^n$;

2) $(y_{i,1}, \dots, y_{i,m})$, $i = 1, \dots, k$ are linearly independent vectors of $\mathbb{R}^m$;

3) the linear span $V =< (y_{i,1}, \dots, y_{i,m}), i = 1, \dots, k >$ does not contain $m - k$ coordinate vectors e_i of $\mathbb{R}^m$.

Theorem 4.1. *If p satisfies (+) then* $\mathrm{rk}(f_+^*(p)) = k(n+m-k)$.

Proof. Since the Jacobian is given by all possible derivatives with respect to $x_{i,j}$ and $y_{a,b}$, it is enough to show that exactly $k(m+n-k)$ of them are linearly independent. First of all we notice that the derivative with respect to $x_{i,j}$ is a matrix of the form

$$f_{x_{i,j}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{pmatrix} y_{i,1} & \dots & y_{i,m} \end{pmatrix}$$

while the derivative with respect to $y_{a,b}$ is a matrix of the form

$$f_{y_{a,b}} = \begin{pmatrix} x_{a,1} \\ \vdots \\ x_{a,n} \end{pmatrix} \begin{pmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{pmatrix}$$

That is, the derivative with respect to $x_{i,j}$, $j = 1, \dots n$ is a matrix with all zeros but the j-th row consisting of the vector $(y_{i,1}, \dots, y_{i,m})$. Similarly the derivative with respect to $y_{a,b}$, $b = 1, \ldots m$ is a matrix with all zeros but the b-th column consisting of the vector $(x_{a,1}, \ldots, x_{a,n})$.

We now build a set $\mathcal{M}$ consisting of $k(m+n-k)$ independent matrices and hence we prove the statement.

The set of n derivatives $f_{x_{1,j}}$ $j = 1, \ldots n$ are clearly independent. We can consider the other set of m derivatives $f_{x_{2,j}}$ $j = 1, \ldots n$, requiring that $(y_{2,1}, \ldots, y_{2,m})$ is not proportional to $(y_{1,1}, \ldots, y_{1,m})$. This is satisfied by property (+). Iterating the process we add to $\mathcal{M}$ the kn independent matrices of the form $f_{x_{i,j}}$, $i = 1, \ldots, k$, $j = 1, \ldots, n$.

To add the other $k(m-k)$ independent matrices we proceed as follows.

We can assume that condition 3) in (+) is satisfied by the first $m-k$ coordinates vectors.

Consider first the derivative $f_{y_{1,1}}$. This matrix has only the first column different from zero and can be expressed as a linear combination of the previous matrices $f_{x_{i,j}}$ if and only if the vector $(1, 0, \ldots, 0)$ lies in the span $V =< (y_{i,1}, \ldots, y_{i,m}), i = 1, \ldots, k >$. Thus, requiring that $(1, 0, \ldots, 0) \notin V$ we have that $\mathcal{M} \cup f_{y_{1,1}}$ consists of independent matrices. A similar argument applies to all the matrices $f_{y_{a,1}}$, $a = 1, \ldots k$ once we require that the vectors $(x_{a,1}, \ldots, x_{a,n})$ are linearly independent, which is exactly condition 2) in (+). In fact these matrices differ only for the elements in the non-zero column (the first one) where we found, varying a, the vectors $(x_{a,1}, \ldots, x_{a,n})$. If these matrices $f_{x_{i,j}}, f_{y_{a,1}}$ were dependent then we would have a linear combination

$$\alpha_1 f_{y_{1,1}} + \cdots + \alpha_k f_{y_{k,1}} + \sum \beta_{i,j} f_{x_{i,j}} = 0.$$

If we write this combination as

$$\alpha_1 f_{y_{1,1}} + \ldots + \alpha_k f_{y_{k,1}} = -\sum \beta_{i,j} f_{x_{i,j}}$$

we notice that the lefthandside is a matrix with non zero elements only in the first column. By the hypothesis that $(1, 0, \ldots, 0) \notin V$ the previous equality can hold if and only if both sides are equal to zero, i.e.

$$\alpha_1 f_{y_{1,1}} + \ldots + \alpha_k f_{y_{k,1}} = 0$$

and

$$\sum \beta_{i,j} f_{x_{i,j}} = 0$$

which is clearly a contradiction.

Similarly, as $(0,1,0,\ldots,0) \notin V$ we can add the set of derivatives $f_{y_{i,2}}$, $i = 1,\ldots,k$ to the set $\mathcal{M}$. The process can go further for $m-k$ times. For each of the k coordinate vectors we add k derivatives to $\mathcal{M}$. In conclusion, in $\mathcal{M}$ we find km derivatives of the form $f_{x_{i,j}}$ ($i = 1,\ldots,k$, $j = 1,\ldots,n$) and $k(m-k)$ derivatives of the form $f_{y_{a,b}}$ ($a = 1,\ldots,k$ and $b = 1,\ldots,m-k$) which are linearly independent. □

We can use this Jacobian approach to investigate properties of the non-negative rank under perturbations preserving the rank.

Proposition 4.2 (Isorank perturbation)**.** *Let $P = f_+(p)$ such that p satisfies (+) and p has positive coordinates. Then there exists a ball $B(P,\epsilon)$ such that for each $N \in B(P,\epsilon) \cap X_{n\times m,k}$ we have*

$$\mathrm{rk}_+(N) = \mathrm{rk}_+(P).$$

Proof. If $P = f_+(p)$ and p satisfies (+) then by Proposition 4.1 we know that f is locally invertible. Hence there exist balls $B(P,\epsilon)$ and $B(p,\delta)$ such that each $N \in B(P,\epsilon) \cap X_{n\times m,k}$ has a unique preimage in $B(p,\delta)$ using the map f. Moreover, if p has positive coordinates we can find a, possibly smaller, ϵ and $q \in f^{-1}(N)$ such that q has positive coordinates. Hence $\mathrm{rk}_+(N) \leq \mathrm{rk}_+(P)$. The conclusion follows by Theorem 3.1. □

Remark 4.3. *The conditions 1)-2)-3) of property (+) are open conditions in the Zariski topology. This means that the matrices in $\mathbb{R}^{nm}_{\geq 0}$ not having property (+) satisfy a set of polynomial equations.*

5 Examples

In this section we will present some interesting examples. Some of these example were inspired to us by the matrix

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

which is the most well-known example of a matrix with rank and non-negative rank which are different (see [CR88]).

Small cases. Let P be an $n \times m$ matrix and assume $n \leq m$. We want to describe how the non-negative rank of P changes under perturbations for small values of n. If $n \leq 3$, then it is easy to show that $\mathrm{rk}(P) = \mathrm{rk}_+(P)$, see [CR88]. Thus the first interesting cases are for $n = 4$ and $\mathrm{rk}(P) = 3$. If $\mathrm{rk}_+(P) = 4$ then any small perturbation will not change the non-negative rank by Theorem 3.1. Thus, let us assume that $\mathrm{rk}_+(P) = 3$. Using Proposition 3.2, we know that there are small perturbations preserving the non-negative rank. Of course, there are small perturbations not preserving it: it is enough to increase the ordinary rank. Hence, we ask: are there small perturbations of P, say P_ϵ, such that $\mathrm{rk}(P_\epsilon) = \mathrm{rk}(P)$ and $\mathrm{rk}_+(P_\epsilon) = 4$? Not surprisingly, the answer depends on the choice of P. It is easy to construct a matrix P with the required ranks and satisfying the hypothesis of Proposition 4.2. Thus, in this case, the answer to our question is no. But, for a different choice of P, the answer can be yes. Consider, for example, P_ϵ defined as follows:

$$P_\epsilon = \begin{pmatrix} 2 & 0 & 2 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 1+\epsilon \\ 2 & 2 & 0 & 1-\epsilon \end{pmatrix},$$

and let $P = P_0$. It easy to see that $\mathrm{rk}(P_\epsilon) = 3$ for all ϵ while $\mathrm{rk}_+(P_0) = 3$ and $\mathrm{rk}_+(P_\epsilon) = 4$ for small positive values of ϵ, see Figure 1. We denote with $c_1, \ldots, c_4$ the points corresponding to the columns of P while $c_4(\epsilon)$ corresponds to the fourth column of P_ϵ. In Figure 1, and in the following figures, we use the graphic representation described in [LC10] and related to the map π_3. More precisely, a 4×4 matrix will be presented as a set of four points in a tetrahedron. This presentation allows for an easy visualization of rank related properties. We notice, for example, that a rank two matrix will correspond to four coplanar points.

Failing of upper-semicontinuity. The upper-semicontinuity of the non-negative rank is of course a local property as shown by the following example. Consider the matrix

$$M_\epsilon = \frac{1}{2(1+2\epsilon)} \begin{pmatrix} 1+\epsilon & \epsilon & 1+\epsilon & \epsilon \\ 1+\epsilon & \epsilon & \epsilon & 1+\epsilon \\ \epsilon & 1+\epsilon & 1+\epsilon & \epsilon \\ \epsilon & 1+\epsilon & \epsilon & 1+\epsilon \end{pmatrix}$$

and let $c_i(\epsilon)$'s be the four column vectors where we set $c_i = c_i(0)$. When $\epsilon = 0$ the matrix has non-negative rank equal to four. Since each c_i is a

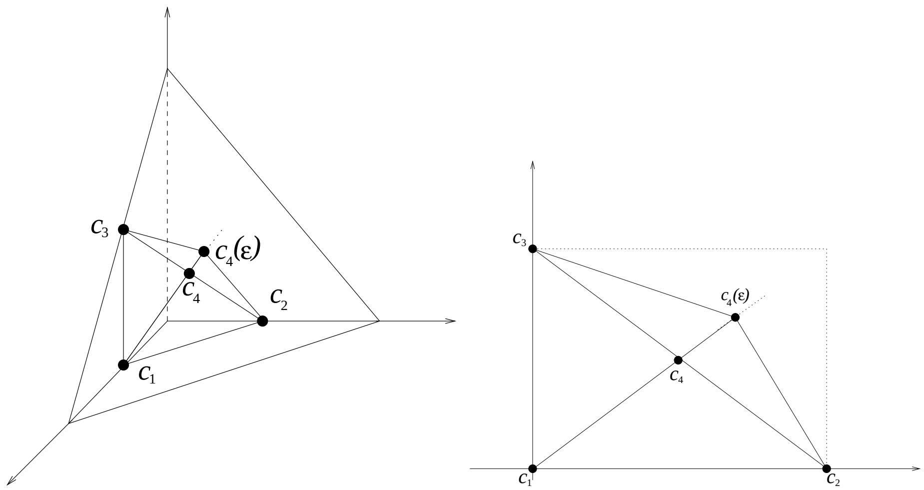

Figure 1: The matrices P_ϵ for $\epsilon = 0$ and a small positive value of ϵ represented in the tetrahedron and in the plane.

vector of sum one, we can use the map π_3 to represent the columns in the simplex Δ^3, which is a tetrahedron in $\mathbb{R}^3$. To simplify the drawings, we have dropped the first coordinate of each column instead of the last one, but of course this does not affect our analysis. The four points for the matrix M_0 are plotted in Figure 2.

The points $c_1(\epsilon), \ldots, c_4(\epsilon)$ are the vertices of a rectangle R_ϵ which we can draw in the plane. As $\epsilon > 0$ increases, the four points move along the main diagonals, as in Figure 3, and R_ϵ will eventually be contained in the triangle ABC where

$$A = (0, \sqrt{2}/2 - 1/2) \qquad B = (\sqrt{2}/4, 1/2) \qquad C = (\sqrt{2}/2, \sqrt{2}/2 - 1/2).$$

It is not hard to show that, for $\epsilon < \sqrt{2}/2$ we have $\mathrm{rk}_+(M_\epsilon) = 4$ while $\mathrm{rk}_+(M_{\sqrt{2}/2}) = 3$. Hence, moving far enough from M_0 the non-negative rank can decrease.

Non-convexity of $X^+_{4\times 4,3}$. In the 4×4 case, the properties of the non-negative rank imply that the unique non-trivial case is the case of rank 3. The matrices in $X_{4\times 4,3}$ can belong to $X^+_{4\times 4,3}$ or to $X^+_{4\times 4,4}$. With the same graphical approach as above, we can show that the set $X^+_{4\times 4,3}$ is not convex even if the ordinary rank is constant. To do this, it is enough to consider the

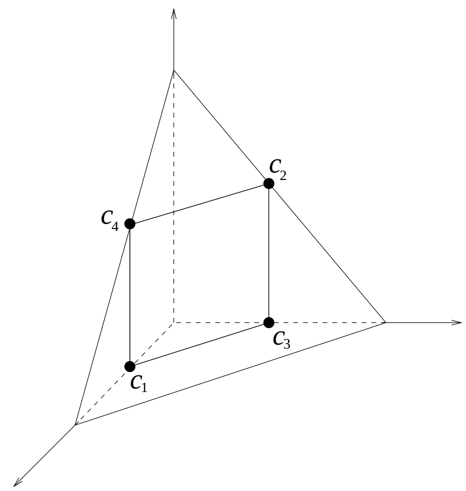

Figure 2: The matrix M_0 in the simplex Δ^3.

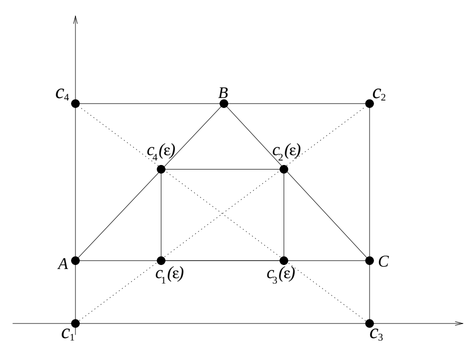

Figure 3: The points $c_1(\epsilon), \ldots, c_4(\epsilon)$ in the critical configuration for $\epsilon = \sqrt{2}/2$.

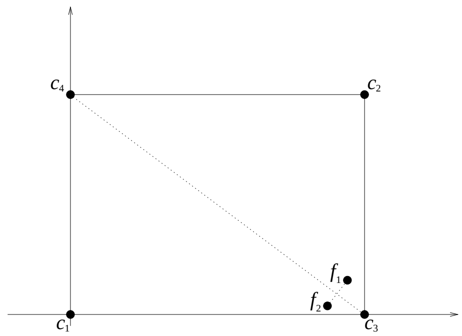

Figure 4: The points $c_1, \ldots, c_4, f_1, f_2$ defining the matrices A_1 and A_2.

two matrices $A_1 = [c_4, c_2, c_3, f_1]$ and $A_2 = [c_4, c_3, c_1, f_2]$ where the columns $c_1, c_2, c_3, c_4, f_1, f_2$ are displayed in Figure 4 in the same plane as in Figure 3. It is immediate to see that both A_1 and A_2 have rank 3 and non-negative rank 3, but the matrix $A = (A_1 + A_2)/2$ has rank 3 (its 4 points are coplanar) but non-negative rank 4. With the same technique, one can also see that the set $X_{4\times 4,3} \setminus X^+_{4\times 4,3}$ is not convex.

$$B_1 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \qquad B_2 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

B_1 and B_2 have rank 3 and non-negative rank 4, as they are obtained from M_0 possibly with permutation of columns, but the matrix $B = (B_1 + B_2)/2$ has non-negative rank 3.

6 Applications to the study of statistical models

The results about the non-negative rank presented above have an useful counterpart in Probability and Mathematical Statistics. In particular the notion of nonnegative rank is useful in the study of mixture of independence models for discrete distributions. We now recall some basic definitions.
Distribution. The distribution (or density) of a random variable X on a set of n possible outcomes $\{1, \ldots, n\}$ is a vector of n non-negative numbers

$(p_1, \ldots, p_n)$ such that

$$p_i \geq 0 \quad \text{for all} \quad i \quad \text{and} \quad \sum_i p_i = 1 \, ,$$

where $p_i = \mathbb{P}(X = i)$ is the probability that X assumes the value i.
Joint distribution. If we consider a pair (X, Y) of random variables on $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$ respectively, the joint distribution of X and Y is a matrix of non-negative numbers $P = (p_{i,j})$ such that

$$p_{i,j} \geq 0 \quad \text{for all} \quad i, j \quad \text{and} \quad \sum_{i,j} p_{i,j} = 1 \, , \tag{2}$$

where $p_{i,j} = \mathbb{P}(X = i, Y = j)$ is the probability that $(X = i)$ and $(Y = j)$.
Probability models. A matrix P satisfying the constraints in Equation (2) is also called a two-way table. The set

$$\Delta = \left\{ P \in \mathbb{R}^{nm} \ : \ p_{i,j} \geq 0 \quad \text{for all} \quad i, j \quad \text{and} \quad \sum_{i,j} p_{i,j} = 1 \right\}$$

is the $n \times m$ (closed) standard simplex and each probability distribution for a pair (X, Y) belongs to Δ. A probability model $\mathcal{M}$ is a subset of Δ. In many cases $\mathcal{M}$ is defined through a set of polynomial equations, and in such case we call $\mathcal{M}$ an algebraic model.
The independence model. For two-way tables, one among the most simple models is the independence model. The construction of the independence model is described for instance in [Agr02]. Under independence of X and Y we have

$$\mathbb{P}(X = i, Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j)$$

for all $i = 1, \ldots, n$ and for all $j = 1, \ldots, m$, and therefore P is a rank one matrix, i.e., there exist vectors r and c such that $P = c(r)^t$. Thus, the independence model for $n \times m$ tables is the set:

$$\mathcal{M}_I = \{P \ : \ \text{rank}(P) = 1\} \cap \Delta \, .$$

Remark 6.1. *It is a well known fact in Linear Algebra that a non-zero matrix P has rank 1 if and only if all 2×2 minors of P vanish. This shows that the* *independence model is an algebraic model. Thus, an equivalent definition of the independence model is as follows. The independence model is the set:*

$$\mathcal{M}_I = \{P \; : p_{i,j}p_{k,h} - p_{i,h}p_{k,j} = 0 \\ \text{for all } \; 1 \le i < k \le n, 1 \le j < h \le m\} \cap \Delta\,.$$

Notice that the model is defined through pure binomials and that the set of all the 2×2 *minors of a matrix are a system of generator of a toric ideal. This is a general fact in the analysis of algebraic statistical models and the models of this form are called toric models. The reader can refer to [DSS09] and [Rap07] for further details.*

Mixture models. The mixture of two independence models is defined through the following procedure:

- Take two distributions $P_1, P_2 \in \mathcal{M}_I$;
- Toss a (biased) coin and choose P_1 with probability α and P_2 with probability $(1-\alpha)$.

In general, the mixture of k independence models is defined as follows. The mixture of k independence models is the set

$$\mathcal{M}_{kI} = \{P \; : \; P = \alpha_1 c_1 (r_1)^t + \ldots + \alpha_k c_k (r_k)^t\} \tag{3}$$

where the vectors r_i, the vectors c_i and $\alpha = (\alpha_1, \ldots, \alpha_k)$ are probability distributions. Some results and examples about this type of statistical models are presented in [FHRZ10].

Notice that in the decomposition in Equation (3), the components must be non-negative, and therefore the model coincides with the set of $n \times m$ matrices with non-negative rank at most k.

We also notice that it is not possible to approximate a matrix with ordinary rank different from the non-negative rank using matrices with ordinary rank and non-negative rank which coincide.

Corollary 6.2. *Let* P *be a non-negative matrix such that* $\mathrm{rk}(P) = k$ *and* $\mathrm{rk}_+(P) > k$*. Then there is no sequence of matrices* P_n *whose limit is* P *such that* $\mathrm{rk}(P_n) = \mathrm{rk}_+(P_n) = k$.

Proof. The statement is a straightforward consequence of Proposition 3.1. □

References

[Agr02] Alan Agresti. *Categorical Data Analysis*. Wiley, New York, 2 edition, 2002.

[BL09] LeRoy B. Beasley and Thomas J. Laffey. Real rank versus nonnegative rank. *Linear Algebra Appl.*, 431:2330–2335, 2009.

[CR88] Joel E. Cohen and Uriel G. Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra Appl.*, 44:847–860, 1988.

[CR10] Enrico Carlini and Fabio Rapallo. Probability matrices, nonnegative rank, and parameterization of mixture models. *Linear Algebra Appl.*, 433(2):424–432, 2010.

[DLC08] Bo Dong, Matthew M. Lin, and Moody T. Chu. Nonnegative rank factorization via rank reduction. Preprint, 2008.

[DSS09] Manthias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on Algebraic Statistics*. Birkhauser, Basel, 2009.

[FHRZ10] Stephen E. Fienberg, Patricia Hersh, Alessandro Rinaldo, and Yi Zhou. Maximum likelihood estimation in latent class models for contingency table data. In Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin, and Henry P. Wynn, editors, *Algebraic and Geometric Methods in Statistics*, pages 27–62. Cambridge University Press, 2010.

[GG10] Nicolas Gillis and François Glineur. On the geometric interpretation of the nonnegative rank. arXiv:1009.0880, 2010.

[GRRW10] Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin, and Henry P. Wynn. *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, Cambridge, 2010.

[HVD08] Ngoc-Diep Ho and Paul Van Dooren. Non-negative matrix factorization with fixed row and column sums. *Linear Algebra Appl.*, 429:1020–1025, 2008.

[LC10] Matthew M. Lin and Moody T. Chu. On the nonnegative rank of euclidean distance matrices. *Linear Algebra Appl.*, 433:681–689, 2010.

[LS01] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Adv. Neural Inform. Process. Systems*, 13:556–562, 2001.

[PPP06] V.P. Pauca, J. Piper, and Robert J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear Algebra Appl.*, 416:29–47, 2006.

[PS05] Lior Pachter and Bernd Sturmfels. *Algebraic statistics for computational biology.* Cambridge University Press, New York, 2005.

[Rap07] Fabio Rapallo. Toric statistical models: Parametric and binomial representations. *Ann. Inst. Statist. Math.*, 59(4):727–740, 2007.