REDUCTION OF A MATRIX WITH POSITIVE ELEMENTS TO A DOUBLY STOCHASTIC MATRIX

M. V. MENON

1. **Introduction.** The following theorem is proved in [2], the motivation being the problem of estimating the transition probability matrix of a finite dimensional Markov chain.

THEOREM. *To a given square matrix A with strictly positive elements there corresponds exactly one doubly stochastic matrix T which can be expressed in the form $T = D_1 A D_2$, where D_1 and D_2 are diagonal matrices with strictly positive diagonal elements. The matrices D_1 and D_2 are themselves unique up to a scalar factor.*

The existence of T, D_1 and D_2 is established by a "constructive" but "limiting" procedure in [2]. There the author also shows by means of specific examples that the assumption made in the theorem that the elements of A are strictly positive cannot, in general, be relaxed to the one that they are nonnegative. In this paper, we give an "existence proof" of the theorem whose main advantage besides its simplicity is that it not only makes clear *why the theorem holds, but also explains* the reasons why the assumption that the elements be positive cannot, in general, be replaced by the assumption that they be only nonnegative.

The proof, in outline, is as follows. It is shown that A defines in a natural manner a continuous mapping of a certain subset of a finite dimensional Euclidean space into itself, which maps lines through the origin onto lines. By an appeal to a fixed-point theorem, one finds that the mapping leaves a line fixed. This means that the co-ordinates of any point on the line are proportional to those of its image on the line. A simple substitution establishes next that the constant of proportionality is unity. Thus all points on the line are fixed points. The co-ordinates of any one of those points form the diagonal elements of the diagonal matrix D_2 of the theorem, and the other diagonal matrix D_1 is determined at once by the mapping.

2. **Proof.** Solely in order to decrease the complexity of notation, it will be assumed that A is a 3×3 matrix with positive entries,

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix}.$$

Let S be the subset of a 3-dimensional Euclidean space of points with nonnegative co-ordinates which will be denoted either by (x', y', z') or by (x, y, z). Let T be an identical copy of S. We shall denote the co-ordinates of any point in it by (u, v, w). Then

(1)
$$\begin{cases} u(ax' + by' + cz') = 1, \\ v(dx' + ey' + fz') = 1, \\ w(gx' + hy' + kz') = 1, \end{cases}$$

defines a mapping of the set of points in S other than the origin into T, and

(2)
$$\begin{cases} x(au + dv + gw) = 1, \\ y(bu + ev + hw) = 1, \\ z(cu + fv + kw) = 1, \end{cases}$$

defines a mapping of the set of points in T other than $(0, 0, 0)$ into S. Hence, (1) and (2) in succession determine a mapping f, which together with the definition $f(0, 0, 0) = (0, 0, 0)$, constitutes a mapping of S into S, i.e., if $(x', y', z') \in S$, then $f(x', y', z') = (x, y, z)$.

Clearly, f has the following properties:

(a) f is continuous,

(b) $f(x', y', z') = 0$, if and only if $x' = y' = z' = 0$,

(c) any line through the origin lying in S whose equations are $x' = \theta l'$, $y' = \theta m'$, $z' = \theta n'$, where $l' \geqq 0$, $m' \geqq 0$, $n' \geqq 0$, $(l')^2 + (m')^2 + (n')^2 > 0$, and θ is a positive parameter, is mapped into a line through the origin with equations $x = \phi l$, $y = \phi m$, $z = \phi n$, where ϕ is a positive parameter, and $l > 0$, $m > 0$, $n > 0$. Note that the fact that l, m, n are strictly positive is a consequence of the *assumption that the elements of A are strictly positive.*

Suppose now that l', m', n' are the direction cosines of a line through the origin in S. Then $(l')^2 + (m')^2 + (n')^2 = 1$, and this equation defines a closed compact subregion Σ say, of S. Indeed, Σ is that portion of the surface of the unit sphere with centre the origin, which lies in the first orthant. And now by (a), (b) and (c) given above f induces a mapping which we denote by the same symbol f, which maps Σ continuously into a subset $f(\Sigma)$ of itself. We have $f(l', m', n') = (l, m, n) \in \Sigma$. l, m, n are the direction cosines of the line into which f maps the line through the origin determined by l', m', n'. By (b), l, m, $n > 0$ for all $(l, m, n) \in f(\Sigma)$. Hence since Σ is obviously homeomorphic to the unit square, we have, by Brouwer's fixed-point theorem ([1, p. 453]) that there exists a point (l, m, n) with strictly positive co-ordinates in Σ such that $f(l, m, n) = (l, m, n)$.

We have thus shown that there exists a line with strictly positive direction cosines, through the origin in S, which is transformed into itself by f. Hence, if (x', y', z') is any point other than the origin on this line, then $f(x', y', z') = (\theta x', \theta y', \theta z')$ for some $\theta > 0$.

But now, if the sum of the left hand sides of (1) is denoted by σ, we have from (1), $\sigma = 3$. On the other hand, substituting $x = \theta x'$, $y = \theta y'$, $z = \theta z'$ in (2) and adding yields $\theta \sigma = 3$. Hence $\theta = 1$. In other words every point on the line is a fixed point of f. Thus if (p, q, r) are the co-ordinates of a fixed point other than $(0, 0, 0)$, then (1) and (2) are simultaneously satisfied if we put $x' = x = p$, $y' = y = q$, $z' = z = r$. The diagonal matrices D_2 and D_1 of the theorem have as diagonal elements p, q, r and u, v, w, respectively, the latter being, of course, determined by relations (1) with $x' = p$, $y' = q$ and $z' = r$.

This demonstrates the "existence" part of the theorem. The uniqueness part is proved in [2].

3. **Positiveness assumption.** As mentioned earlier, examples are given in [2] to show that the assumption that the entries of A are strictly positive cannot be relaxed. It may not be amiss to point out that the singularity or nonsingularity of A is really irrelevant to the question. For one can readily verify that if

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix},$$

then the theorem will not hold.

The explanation for the importance of the assumption of positive entries for A is furnished by property (c) in the previous section. When A has some zero elements, the line fixed by the transformation f of §2 *may* have one or more of its direction cosines zero. This would imply (see the penultimate paragraph of §2) that D_2 has a zero diagonal entry, a situation which is obviously impossible. For example, the "fixed line" corresponding to the matrix A in the preceding paragraph, has direction cosines of the form $l, 0, n$.

ADDED IN PROOF. The methods used above enable one to establish also the following more general theorem:

Let $r_1, \cdots, r_m; c_1, \cdots, c_n$ be positive numbers. $A_{m \times n}$ is a positive matrix. $B_{m \times n}$ is a nonnegative matrix, no column of which consists entirely of zeroes. Then there exists a unique positive number θ and two diagonal matrices D and E such that DAE has row sums equal to the r_i, and θDBE has column sums equal to the c_j.

The author has also obtained extensions of this result to certain wider classes of pairs of nonnegative matrices A and B.

References

1. N. Dunford and J. T. Schwartz, *Linear operators. Part 1, General theory*, Interscience, New York, 1958.

2. R. Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, Ann. Math. Statist. **35** (1964), 876–879.

IBM Research Laboratory, San Jose, California