A RELATIONSHIP BETWEEN ARBITRARY POSITIVE MATRICES AND DOUBLY STOCHASTIC MATRICES

By Richard Sinkhorn

University of Houston

1. Introduction. Suppose one observes n transitions of a Markov chain with N states and stochastic matrix $P = (p_{ij})$. The usual estimate of p_{ij} is $t_{ij} = a_{ij}/\lambda_i$ where a_{ij} is the number of transitions from i to j which are observed, and $\lambda_i = \sum_j a_{ij}$. (Cf. [1].) This amounts to a normalization of the rows of $A = (a_{ij})$, and can be expressed as a matrix equation $T = D_1A$ where $T = (t_{ij})$ and $D_1 = \text{diag}[\lambda_1^{-1}, \cdots, \lambda_N^{-1}]$.

If it is known that the stochastic matrix P is in fact doubly stochastic, (i.e., $\sum_i p_{ij} = 1$), what then is a good estimate of T? The maximum likelihood equations are difficult to solve. One estimate which has been used (for example, by Welch [4]) is to alternately normalize the rows and columns of A, in the belief that this iterative process converges to a doubly stochastic matrix, T, which might be, in some sense, a good estimate.

It is not the intent of this paper to obtain properties of this estimate but only to examine the mechanics of the iteration itself. In the next section we shall study this in detail and show that it is always convergent if the matrix A is strictly positive (i.e., $a_{ij} > 0$ for all i, j), and in fact that there exist diagonal matrices D_1 and D_2 (unique up to a scalar factor) with positive diagonals such that $T = D_1AD_2$. T is the only doubly stochastic matrix expressable in this form for a given strictly positive A.

For completeness we shall include a corollary to this result due to Marcus and Newman [3] which states that if A is symmetric and has positive entries, then there exists a diagonal matrix D with positive main diagonal entries such that DAD is doubly stochastic.

Finally in the last section we shall show by example that convergence need not occur at all if some $a_{ij} = 0$, or even if it does there need exist no associated diagonal matrices D_1 and D_2 as in the strictly positive case. Even the apparently natural artifice of replacing the zero entries by "small" functions $a_{ij}(\epsilon)$ of a parameter ϵ, getting $T(\epsilon)$ and letting $\epsilon \to 0$ leads to difficulties.

2. The alternating iteration for positive matrices.

Theorem 1. *To a given strictly positive $N \times N$ matrix A there corresponds exactly one doubly stochastic matrix T_A which can be expressed in the form $T_A = D_1AD_2$ where D_1 and D_2 are diagonal matrices with positive diagonals. The matrices D_1 and D_2 are themselves unique up to a scalar factor.*

Proof. We shall establish only the uniqueness part here. The existence will be demonstrated constructively in the proof of Theorem 2.

If there exist two different pairs of diagonal matrices D_1, D_2 and C_1, C_2 such

Received 17 December 1962; revised 25 November 1963.

that both D_1AD_2 and C_1AC_2 are doubly stochastic, then this means that there exists a positive doubly stochastic matrix P and matrices $B_1 = \mathrm{diag}[b_{11}, b_{12}, \cdots, b_{1N}]$, $B_2 = \mathrm{diag}[b_{21}, b_{22}, \cdots, b_{2N}]$ which are not multiples of the identity matrix, for which B_1PB_2 is also doubly stochastic. But this is impossible since by convexity, one obtains

$$\min_j b_{2j} \leqq 1/b_{1i} \leqq \max_j b_{2j}\,; \qquad \min_i b_{1i} \leqq 1/b_{2j} \leqq \max_i b_{1i}$$

which leads to a contradiction if $b_{1i}b_{2j} \neq 1$ for some i and j. It follows that $C_1 = pD_1$, $C_2 = p^{-1}D_2$ for some positive number p.

THEOREM 2. *The iterative process of alternately normalizing the rows and columns of strictly positive $N \times N$ matrix is convergent to a strictly positive doubly stochastic matrix.*

PROOF. The iteration produces a sequence of positive matrices which alternately have row and column sums one. We shall show that the two subsequences which are composed respectively of the matrices with row sums one and the matrices with column sums one each converge to a positive doubly stochastic limit of the form D_1AD_2 where each D_i has a positive diagonal. The uniqueness part of Theorem 1 will complete the proof. Since the terms of either of the subsequences are generated in the same way as are the transposes of the terms of the other, only one convergence proof is required.

Let $\{A_n\} = \{(a_{nij})\}$ be the sequence with column sums one and let a_n be the minimal element of A_n. We shall show that $\{a_n\}$ is bounded away from zero.

Let A_n have row sums $\lambda_{n1}, \cdots, \lambda_{nN}$ and set

$$\delta_{nj} = \sum_i a_{nij}/\lambda_{ni}\,.$$

Since δ_{nj} is a convex combination of the $1/\lambda_{ni}$ and $\lambda_{n+1,i}$ is a convex combination of the $1/\delta_{nj}$, it follows that

$$(1)\quad \lambda_n(m) \leqq 1 \leqq \lambda_n(M) \Rightarrow \lambda_n(m) \leqq \lambda_{n+1}(m) \leqq 1 \leqq \lambda_{n+1}(M) \leqq \lambda_n(M)$$

where the m and M respectively label minimal and maximal quantities relative to a given A_n. Similarly

$$(2)\quad \delta_n(m) \leqq 1 \leqq \delta_n(M) \Rightarrow \delta_n(m) \leqq \delta_{n+1}(m) \leqq 1 \leqq \delta_{n+1}(M) \leqq \delta_n(M).$$

Therefore the maximum and minimum row and column sums are monotone sequences and hence have limits. To complete the proof, it is necessary to show that these limits all equal one.

Let $x_{ni} = [\lambda_{1i}\lambda_{2i}\cdots\lambda_{ni}]^{-1}$ and $y_{nj} = [\delta_{1j}\delta_{2j}\cdots\delta_{nj}]^{-1}$. Then if $A_1 = (a_{ij})$ has minimal element a,

$$y_{nj} = 1/\sum_i a_{ij}x_{ni} \leqq 1/a_{ij}x_{ni} \leqq 1/ax_{ni}$$

for all i and j. Thus in particular $y_{nj} \leqq 1/ax_n(M)$. Since $\sum_j x_{ni}a_{ij}y_{nj} = \lambda_{n+1,i} \geqq \lambda_{n+1}(m) \geqq \lambda_1(m) = \lambda$, it follows that

$$x_{ni} \geqq \lambda/\sum_j a_{ij}y_{nj} \geqq a\lambda x_n(M)/N.$$

Also $y_{nj} \geqq 1/\sum_i a_{ij}x_{ni} \geqq 1/Nx_n(M)$ and we see that $a_{n+1,\, ij} = x_{ni}a_{ij}y_{nj} \geqq a\lambda/N^2 = \mu$; whence $a_n \geqq \mu > 0$ for all n.

It is clear from (1) that $\lambda_n(M) \to 1 + c$ where $c \geqq 0$. For convenience set $\lambda_n(M) = 1 + c_n$. Then if $\mu_j[\lambda_{ni} \leqq 1] = \sum a_{nij}$ where the sum is taken over all i for which $\lambda_{ni} \leqq 1$, and if $\mu_j[\lambda_{ni} > 1]$ has a corresponding meaning,

$$\delta_{nj} \geqq \mu_j[\lambda_{ni} \leqq 1] + \frac{1}{1 + c_n}\,\mu_j[\lambda_{ni} > 1] = \frac{1 + c_n\,\mu_j[\lambda_{ni} \leqq 1]}{1 + c_n} \geqq \frac{1 + c_n\, a_n}{1 + c_n}.$$

Then if $\lambda_{n+1}(M) = \lambda_{n+1,i_0}$,

$$1 + c \leqq \lambda_{n+1,i_0} = \sum_j a_{ni_0j}/\lambda_{ni_0}\delta_{nj} \leqq (1 + c_n)/(1 + c_n a_n);$$

if $c > 0$, $a_n \to 0$, a contradiction. Thus $c = 0$ and $\lambda_n(M) \to 1$. It readily follows that $\lambda_n(m) \to 1$.

Corollary. (Marcus and Newman [3]) *If A is symmetric and has positive entries there exists a diagonal matrix D with positive main diagonal entries such that DAD is doubly stochastic.*

Proof. Let $S = D_1AD_2$ be doubly stochastic where the D_i are as in Theorem 1. Then $A = D_1^{-1}SD_2^{-1}$ and $A^T = A$ implies that $D_2D_1^{-1}SD_2^{-1}D_1 = S^T$, and since S^T is doubly stochastic, $D_2D_1^{-1}$ is a scalar multiple of the identity by the uniqueness part of Theorem 1. Thus we can take $D_1 = D_2 = D$.

3. Remarks concerning matrices with zero entries. When A contains zero elements Theorems 1 and 2 need no longer hold. If

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

the iteration oscillates and there certainly exists no D_1 and D_2. If

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{the iteration converges to} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

but again there is no D_1 and D_2.

One might try to overcome these difficulties by replacing the zero elements in A by small quantities. But this approach may be questionable. For instance if

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad \text{is approximated by} \quad A(\epsilon) = \begin{pmatrix} \epsilon & \epsilon & 1 \\ \epsilon & \epsilon & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

the limit of $D_1(\epsilon)A(\epsilon)D_2(\epsilon)$ is

$$\begin{pmatrix} .25 & .25 & .5 \\ .25 & .25 & .5 \\ .5 & .5 & 0 \end{pmatrix} \quad \text{as} \quad \epsilon \to 0.$$

If A is approximated by

$$A'(\epsilon) = \begin{pmatrix} \epsilon & \epsilon^2 & 1 \\ \epsilon & \epsilon^2 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

the doubly stochastic limit becomes

$$\begin{pmatrix} .5 & 0 & .5 \\ .5 & 0 & .5 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{as} \quad \epsilon \to 0,$$

something quite different. In fact, it is possible to have $A(\epsilon) \to A$ without having $D_1(\epsilon)A(\epsilon)D_2(\epsilon)$ converge at all as $\epsilon \to 0$. If

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad A(\epsilon) = \begin{pmatrix} \epsilon & \epsilon \sin^2 1/\epsilon \\ 1 & 1 \end{pmatrix}$$

where $\epsilon > 0$, $\epsilon \neq 1/n\pi$,

$$D_1(\epsilon)A(\epsilon)D_2(\epsilon) = \begin{pmatrix} \alpha_\epsilon & 1 - \alpha_\epsilon \\ 1 - \alpha_\epsilon & \alpha_\epsilon \end{pmatrix}$$

where $\alpha_\epsilon^{-1} = 1 + |\sin 1/\epsilon|$. This has no limit as $\epsilon \to 0$.

Whence any attempt to estimate the transition matrix from an observation matrix by a double normalization or by an alternating row-column iteration may well result in failure when the observation contains zero entries. It may also be a poor policy to use a strictly positive approximation for an observation with zeros in hopes of finding an approximate transition matrix, unless there is some very good reason for a particular selection.

Acknowledgment. The author wishes to thank Professor Ronald Pyke for his many constructive suggestions in the presentation.

REFERENCES

[1] BILLINGSLEY, PATRICK (1962). *Statistical Inference for Markov Processes*. Univ. of Chicago Press.

[2] KEMENY, JOHN G. and SNELL, J. LAURIE (1960). *Finite Markov Chains*. Van Nostrand, Princeton.

[3] MARCUS, MARVIN and NEWMAN, MORRIS (1961). The permanent of a symmetric matrix, Abstract 587-85. *Amer. Math. Soc. Notices* **8** 595.

[4] WELCH, LLOYD, Unpublished Report of the Institute of Defense Analysis. Princeton, New Jersey.