ON THE IDENTIFIABILITY PROBLEM FOR FUNCTIONS[1] OF FINITE MARKOV CHAINS

By David Blackwell and Lambert Koopmans

University of California, Berkeley

1. Summary. Let $M = \| m_{ij} \|$ be a 4×4 irreducible aperiodic Markov matrix such that $h_1 \neq h_2$, $h_3 \neq h_4$, where $h_i = m_{i1} + m_{i2}$. Let $x_1, x_2, \cdots$ be a stationary Markov process with transition matrix M, and let $y_n = 0$ when $x_n = 1$ or 2, $y_n = 1$ when $x_n = 3$ or 4. For any finite sequence $s = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)$ of 0's and 1's, let $p(s) = \Pr\{y_1 = \epsilon_1, \cdots, y_n = \epsilon_n\}$. If

$$p^2(00) \neq p(0)p(000) \quad \text{and} \quad p^2(01) \neq p(1)p(010), \tag{1}$$

the joint distribution of $y_1, y_2, \cdots$ is uniquely determined by the eight probabilities $p(0)$, $p(00)$, $p(000)$, $p(010)$, $p(0000)$, $p(0010)$, $p(0100)$, $p(0110)$, so that two matrices M determine the same joint distribution of $y_1, y_2, \cdots$ whenever the eight probabilities listed agree, provided (1) is satisfied. The method consists in showing that the function p satisfies the recurrence relation

$$p(s, \epsilon, \delta, 0) = p(s, \epsilon, 0)a(\epsilon, \delta) + p(s, \epsilon)b(\epsilon, \delta) \tag{2}$$

for all s and $\epsilon = 0$ or 1, $\delta = 0$ or 1, where $a(\epsilon, \delta)$, $b(\epsilon, \delta)$ are (easily computed) functions of M, and noting that, if (1) is satisfied, $a(\epsilon, \delta)$ and $b(\epsilon, \delta)$ are determined by the eight probabilities listed. The class of doubly stochastic matrices yielding the same joint distribution for $y_1, y_2, \cdots$ is described somewhat more explicitly, and the case of a larger number of states is considered briefly.

2. Introduction. Suppose a certain process is known to be a stationary Markov process with N states, say $1, 2, \cdots, N$, and unknown transition matrix M, supposed irreducible and aperiodic. To what extent can we identify M by successive observations on the process, if by observation we are unable to distinguish between certain states of the process? More precisely, if $\{X_n\}$ is a stationary Markov process with states $1, 2, \cdots, N$ and $N \times N$ irreducible aperiodic transition matrix M, and $y_n = \phi(X_n)$, call two such M's equivalent (for the given ϕ) if they determine the same joint distribution of $y_1, y_2, \cdots$. Call a finite set of functions $f_1, \cdots, f_k$, each defined on the set of all $N \times N$ irreducible aperiodic Markov matrices, a *complete set of invariants* if M_1 and M_2 are equivalent if and only if $f_i(M_1) = f_i(M_2)$ for $i = 1, \cdots, k$. Our problem is that of finding a minimal complete set of invariants, i.e., a complete set of which no proper subset is complete. We do not solve this problem, even in special cases, but almost solve it in the two special cases (a) ϕ has only two values, one of which

Received December 20, 1956.

[1] This paper was supported in part by funds provided under Contract AF-41(657)-29 with the Air Research and Development Command, USAF School of Aviation Medicine, Randolph Field, Texas.

is assumed at only a single state and (b) $N = 4$, ϕ has two values, each assumed on two states. By almost solving the problem, we mean the following. Call a set of functions $f_1, \cdots, f_k$ a complete set of invariants relative to a set X of matrices if (1) X is a union of equivalence classes, (2) $f_i(M_1) = f_i(M_2)$ implies M_1 is equivalent to M_2, (3) if M_1 and M_2 are equivalent and in X, $f_i(M_1) = f_i(M_2)$. Thus a complete set of invariants relative to X fails to be a complete set only because two matrices M_1, M_2 not in X may be equivalent even though $f_i(M_1) \neq f_i(M_2)$. In the two special cases above, we find a complete set of invariants relative to a set X containing most matrices.

For case (a) the solution, following the methods of Feller [1], is straightforward. Say ϕ assumes the value 0 on state 1, the value 1 on all other states. The joint distribution of $y_1, y_2, \cdots$ determines and is determined by the distribution of return times to state 1, i.e., by the sequence of numbers

$$\alpha_n = \Pr\{x_{n+1} = 1, \quad x_j \neq 1 \text{ for } 2 \leqq j \leqq n \mid x_1 = 1\},$$

which determines and is determined by its generating function $A_1(t) = \sum_1^\infty \alpha_n t^n$. Define

$$A_i(t) = \sum_1^\infty \Pr\{x_{n+1} = 1, \quad x_j \neq 1 \text{ for } 2 \leqq j \leqq n \mid x_1 = i\} t^n.$$

Then the functions A_i, $i = 1, \cdots, N$ satisfy the system

$$A_i(t) = t\left[m_{i1} + \sum_{j=2}^N m_{ij} A_j(t)\right], \quad i = 1, \cdots, N.$$

Cramer's rule yields

$$A_1(t) = 1 - (\det(I - tM) / \det(J - tM_1)),$$

where I, J are the $N \times N$ and $(N-1) \times (N-1)$ identity matrices and M_1 is obtained from M by deleting the first row and column. Thus two matrices are equivalent whenever they determine the same polynomials $P(t) = \det(I - tM)$ and $Q(t) = \det(J - tM_1)$ and, if for a given M these polynomials have no common roots, a second M is equivalent if and only if it has the same P and Q. Thus, on the class X of matrices for which P and Q have no common roots, the coefficients of P and Q are a complete set of invariants. That two matrices not in X may be equivalent even though the polynomials P, Q differ is shown by the example

$$M = \begin{vmatrix} \frac{1}{2} & x & \frac{1}{2} - x \\ \frac{1}{2} & y & \frac{1}{2} - y \\ \frac{1}{2} & z & \frac{1}{2} - z \end{vmatrix}.$$

All choices of x, y, z, $0 < x, y, z < 1/2$ lead to equivalent M's, while P, Q do depend on x, y, z.

3. The case (2, 2). Suppose $N = 4$ and that ϕ assumes two values, each on two states. Say $\phi(1) = \phi(2) = 0$; $\phi(3) = \phi(4) = 1$. Let $h_i = m_{i1} + m_{i2}$,

$i = 1, \cdots, 4$; we assume $h_1 \neq h_2$, $h_3 \neq h_4$. For any finite sequence $s = (\epsilon_1, \cdots, \epsilon_n)$ of 0's and 1's, let $p(s) = \Pr\{y_1, \cdots, y_n) = s\}$. We shall prove

THEOREM 1. *The function p satisfies*

$$p(s, \epsilon, \delta, 0) = p(s, \epsilon, 0)a(\epsilon, \delta) + p(s, \epsilon)b(\epsilon, \delta)$$

for all s and $\epsilon = 0$ or 1, $\delta = 0$ or 1, where

$$a(0, \delta) = (\Pr\{(y_2, y_3) = (\delta, 0) \mid x_1 = 1\} - \Pr\{(y_2, y_3) = (\delta, 0) \mid x_1 = 2\}) / (h_1 - h_2),$$

$$a(1, \delta) = (\Pr\{(y_2, y_3) = (\delta, 0) \mid x_1 = 3\} - \Pr\{(y_2, y_3) = (\delta, 0) \mid x_1 = 4\}) / (h_3 - h_4),$$

$$b(0, \delta) = (h_1 \Pr\{(y_2, y_3) = (\delta, 0) \mid x_1 = 2\} - h_2 \Pr\{(y_2, y_3) = (\delta, 0) \mid x_1 = 1\}) / (h_1 - h_2),$$

$$b(1, \delta) = (h_3 \Pr\{(y_2, y_3) = (\delta, 0) \mid x_1 = 4\} - h_4 \Pr\{(y_2, y_3) = (\delta, 0) \mid x_1 = 3\}) / (h_3 - h_4).$$

PROOF. For any s and any $i = 1, \cdots, 4$, let $q(s, i) = \Pr\{(y_1, \cdots, y_n) = s, x_{n+1} = i\}$. Then

$$p(s, \epsilon, \delta, 0) = \sum_{\substack{\phi(i)=\epsilon \\ \phi(j)=\delta}} q(s, i)m_{ij}h_j. \tag{3}$$

Fix i and denote by i^* the state different from i for which $\phi(i) = \phi(i^*)$. Then

$$p(s, \epsilon, 0) - q(s, i)h_i = h_{i^*}(p(s, \epsilon) - q(s, i)), \tag{4}$$

since each side is $\Pr\{(y_1, \cdots, y_n) = s, x_{n+1} = i^*, y_{n+2} = 0\}$. Solving (4) for $q(s, i)$ and substituting in (3) expresses $p(s, \epsilon, \delta, 0)$ as a linear combination of $p(s, \epsilon, 0)$, $p(s, \epsilon)$ whose coefficients are functions of M, ϵ, δ. These coefficients are the quantities denoted by $a(\epsilon, \delta)$, $b(\epsilon, \delta)$ in (2).

COROLLARY 1. *The distribution of $y_1, y_2, \cdots$ is determined by $p(0)$, $p(00)$ and the functions $a(\epsilon, \delta)$, $b(\epsilon, \delta)$.*

PROOF. We have $p(1) = 1 - p(0)$ and, since the $\{y_n\}$ process is stationary, $p(10) = p(01) = p(0) - p(00)$, so that $p(11) = 1 - 2p(0) + p(00)$. Thus $p(s)$ is determined if the length of s does not exceed 2. (2) determines $p(s, 0)$ in terms of p for shorter sequences and $a(\epsilon, \delta)$, $b(\epsilon, \delta)$, and $p(s, 1) = p(s) - p(s, 0)$, so that, by induction, p is determined for all s.

COROLLARY 2. *On the set X of matrices for which $p^2(00) \neq p(0)p(000)$ and $p^2(01) \neq p(1)p(010)$, the eight functions $p(0)$, $p(00)$, $p(0, \epsilon, 0)$, $p(0, \epsilon, \delta, 0)$, where $\epsilon = 0$ or 1, $\delta = 0$ or 1 are a complete set of invariants.*

PROOF. Letting s be empty and the sequence 0 in (2) yields

$$p(\epsilon, \delta, 0) = p(\epsilon, 0)a(\epsilon, \delta) + p(\epsilon)b(\epsilon, \delta)$$

$$p(0, \epsilon, \delta, 0) = p(0, \epsilon, 0)a(\epsilon, \delta) + p(0, \epsilon)b(\epsilon, \delta).$$

Thus if $p(\epsilon, 0)p(0, \epsilon) \neq p(\epsilon)p(0, \epsilon, 0)$ for $\epsilon = 0$ or 1, the functions $a(\epsilon, \delta)$, $b(\epsilon, \delta)$

are determined by $p(s)$ for s of length not exceeding 4, so that the latter set is a complete set of invariants on X by Corollary 1. Since $\{y_n\}$ is stationary, $p(s)$ for all s of length not exceeding four is determined by the eight probabilities described in the corollary, so that this set is complete on X.

Thus, since there are twelve parameters in a 4×4 Markov matrix and an equivalence class is defined by eight restrictions, there is in general a four-parameter set of matrices equivalent to a given matrix. An explicit parametric representation of the equivalence classes has not been found.

For the case of doubly stochastic matrices, in which there are nine parameters it turns out that an equivalence class is determined by seven restrictions, so that, in general, there is a two-parameter set of doubly stochastic matrices equivalent to a given doubly stochastic matrix. Moreover an explicit representation can be given, as follows: For any 4×4 doubly stochastic matrix for which $h_1 \neq h_2$, $h_3 \neq h_4$, there is a unique set of numbers σ, a, A, b, B, d, D, x, y for which the matrix has the form (U_1, U_2, U_3, U_4), where the column vectors are given by

$$U_1 = \begin{pmatrix} \sigma + a + x + (d/x) \\ \sigma - a - x + (d/x) \\ \tfrac{1}{2} - \sigma - y - (d/x) + b(y/x) \\ \tfrac{1}{2} - \sigma + y - (d/x) - b(y/x) \end{pmatrix}, \quad U_2 = \begin{pmatrix} \sigma - a + x - (d/x) \\ \sigma + a - x - (d/x) \\ \tfrac{1}{2} - \sigma - y + (d/x) - b(y/x) \\ \tfrac{1}{2} - \sigma + y - (d/x) - b(y/x) \end{pmatrix},$$

$$U_3 = \begin{matrix} \tfrac{1}{2} - \sigma - x - (D/y) + B(x/y) \\ \tfrac{1}{2} - \sigma + x - (D/y) - B(x/y) \\ \sigma + A + y + (D/y) \\ \sigma - A - y + (D/y) \end{matrix}, \quad U_4 = \begin{matrix} \tfrac{1}{2} - \sigma - x + (D/y) - B(x/y) \\ \tfrac{1}{2} - \sigma + x + (D/y) + B(x/y) \\ \sigma - A + y - (D/y) \\ \sigma + A - y - (D/y) \end{matrix}$$

It is a tedious but straightforward matter to check that $p(0)$ $(= \tfrac{1}{2})$, $p(00)$ $(= \sigma)$ and the functions $a(\epsilon, \delta)$, $b(\epsilon, \delta)$ determine and are determined by σ, a, A, b, B, d, D, and that the restrictions $p(\epsilon, 0)p(0, \epsilon) \neq p(\epsilon)p(0\epsilon 0)$ assert $d \neq 0$, $D \neq 0$. Thus any choice of x, y for which all elements remain nonnegative produces a doubly stochastic matrix equivalent to the original, and every such matrix may be obtained for some x, y.

4. A large complete set of invariants. For any $N \times N$ irreducible aperiodic M and any ϕ, let R be the range of ϕ and let S be the set of all finite sequences $s = (r_1, \cdots, r_k)$, $k = 0, 1, 2, \cdots$, $r_i \in R$. For each s the function $p_s(M) = \Pr\{(y_1, \cdots, y_k) = s\}$, as a function of M, is invariant, that is, $p_s(M_1) = p_s(M_2)$ if M_1 and M_2 are equivalent.

THEOREM 2. *There exists a positive integer J, depending only on N and ϕ, such that the set of functions p_s for s not exceeding J in length is a complete set of invariants, that is, the joint distribution of $y_1, y_2, \cdots$ is determined by the joint distribution of $y_1, \cdots, y_J$.*

PROOF. For any $s = (r_1, \cdots, r_k)$, $k \geqq 2$, we have

$$p_s(M) = \sum_{\phi(i_1)=r_1, \cdots, \phi(i_k)=r_k} \lambda_{i_1} m_{i_1 i_2} \cdots m_{i_{k-1} i_k}$$

$$= \lambda M(r_1, r_2) M(r_2, r_3) \cdots M(r_{k-1}, r_k)\delta,$$

$$\lambda = (\lambda_1, \cdots, \lambda_N),$$

where $\lambda_i = \Pr\{x_1 = i\}$, $M(r, r')$ is the matrix obtained from M by replacing m_{ij} by 0 unless $\phi(i) = r$, $\phi(j) = r'$, and δ is the $N \times 1$ column vector with each element unity. Write $M(s) = M(r_1, r_2)M(r_2, r_3), \cdots, M(r_{k-1}, r_k)$. Let F be a second $N \times N$ irreducible aperiodic Markov matrix. Then $p_s(M) = p_s(F)$ if and only if $\lambda M(s)\delta = \mu F(s)\delta$, where μ is the stationary distribution for F. We must find a J such that $\lambda M(s)\delta = \mu F(s)\delta$ for all s of length not exceeding J implies equality for all s. Let $A(s)$ be the $2N \times 2N$ matrix with $M(s)$ in the upper left, $F(s)$ in the lower right, and zeros elsewhere, so that $\lambda M(s) = \mu F(s)\delta$ may be written $\alpha A(s)d = 0$, where $\alpha = (\lambda, -\mu)$ and d is the $2N \times 1$ column vector whose elements are unity. If we consider the class of $2N \times 2N$ matrices as a linear space of $4N^2$ dimensions, the set of matrices $A(s)$ spans a subspace L of dimension $J - 1 \leqq 4N^2$. It remains to show only that the set of matrices $A(s)$ for the length of s not exceeding J already spans L, for if so then any $A(s)$ is a linear combination of these, and $\alpha A(s)d = 0$ whenever the length of s is $\leqq J$ implies $\alpha A(s)d = 0$ for all s. Let L_k denote the linear space spanned by the matrices $A(s)$ for which the length of s does not exceed k. If $L_{k+1} = L_k$ then $L_{k+2} = L_{k+1}$, for say $s = (r, s')$ where $r \varepsilon R$ and s' has length $k + 1$. Then $A(s) = A(r, r')A(s')$, where r' is the initial element of s'. Now by hypothesis $A(s') = \sum_{t \varepsilon T} a(t)A(t)$, where T is the set of sequences of length $\leqq k$, so that $A(s) = \sum_{t \varepsilon T} a(t)A(r, r')A(t)$. Unless r' is the initial element of t, $A(r, r')A(t) = 0$, so that $A(s) = \sum a(t)A(r, t)$ where the sum is over those $t \varepsilon T$ whose initial element is r'. Thus $A(s) \varepsilon L_{k+1}$. We have $L_2 \subset L_3 \subset \cdots \subset L_n \subset \cdots$, with equality for sufficiently large n. If equality first occurs at k, that is, $L_k = L_{k+1}$, we have $L_k = L$. The dimension of L is at least $k - 1$, so that $J - 1 \geqq k - 1$ and $L_J = L$, completing the proof.

The J obtained in the theorem, namely $J = 4N^2 + 1$ is extremely crude. It can be improved somewhat by a more careful bound on the dimension of L. For instance, since all $A(s)$ have zeros in the lower right and upper left places, the actual dimension of L is at most $2N^2$, so that $J = 2N^2 + 1$ will suffice. However, if ϕ is the identity function, L may actually have dimension $2N^2$, while $J = 2$ will suffice, so that, if we are to find the smallest J, a different approach is required.

REFERENCE

[1] WILL FELLER, *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons, Inc., New York, 1950.