JOURNAL OF MATHEMATICAL PSYCHOLOGY 11, 391–403 (1974)

Equivalence Classes of Functions of Finite Markov Chains[1]

BRIAN A. WANDELL[2]

School of Social Sciences, University of California, Irvine, California 92664

AND

JAMES G. GREENO AND DENNIS E. EGAN

University of Michigan, Human Performance Center, Ann Arbor, Michigan

A matrical representation of a Markov chain consists of the initial vector and transition matrix of the chain, along with matrices that specify which observable response occurs for each state. The likelihood function based on a Markov model can be stated in a general way using the components of the model's matrical representation. It follows directly from that statement that two models are equivalent in likelihood if they are related through matrix operations that constitute a change of basis of the matrical representation. Two necessary properties of a change matrix associating two Markov models that are members of the same equivalence class with respect to likelihood are derived. Examples are provided, involving use of the results in analyzing identifiability of Markov models, including a useful application of diagonalization that provides a connection between the problem of identifiability and the eigenvalue problem.

The main development of this paper will be to characterize classes of Markov theories that possess identical likelihood functions. Earlier analyses (Greeno & Steiner, 1964; Greeno, 1967, 1968; Steiner & Greeno, 1969) have attempted to characterize these classes of theories by specifying a set of identifiable parameters of a theory that are given as functions of the theoretical parameters. Then equivalent versions of the model are cases where the theoretical parameters lead to the same values of the identifiable parameters. This approach has not led to very general characterizations. The approach taken here is to consider a model as a set of matrices and vectors—a vector of initial probabilities, a transition matrix, a set of matrices that relate the states of the model to observable responses, and a summation vector. The characterization of models that are equivalent in likelihood is a relationship among the matrices defining

[1] This research was supported by National Science Foundation Grant GB-31045. The authors are grateful to Mr. G. Avrunin for many helpful discussions.

[2] Now at the University of California, Irvine.

the two models (see (2)). Many readers will immediately recognize this relationship as simply reexpressing the matrical representation of the model with respect to a new basis.

One implication of this fact is that we now have a means of generating alternative representations of a Markov theory that may contain fewer parameters than the original. The existence of such a representation implies that the parameters in the original model are not independently estimable. A new, reduced, parameter space is provided by the constructed formulation. A model that is stated in a form without independently estimable parameters is said to lack identifiability. In general, this occurs when probabilities of data are determined by a smaller set of parameters than that specified in the theory. While estimates of the smaller parameter set can be found, no amount of data can provide estimates of all the theoretical parameters, unless new kinds of experiments are devised that provide new kinds of data. (For a fuller discussion of the general issue of identifiability see Restle & Greeno, 1970, Chapter 10. The problem is a standard topic in econometrics, and a basic series of papers is in Koopmans, 1950.)

THEORY

The notation to be used is patterned after Erickson (1970). The proposition to be proved is demonstrated almost trivially once it is properly formulated.

Let $X = \{x_t, I\}$ be a standard Markov chain with stationary transition probabilities and state space I, assumed to be finite and minimal (cf. Chung, 1960). Let v denote the start vector; elements of v are $p(x_1 = i)$ for $i \in I$. Let A be the matrix of transition probabilities, $p(x_{t+1} = j \mid x_t = i)$, for i and j in I.

Let D be the set of observable outcomes, usually a set of responses. Let f denote a function that takes I onto D. Note that this function associates with each state a unique response from D and thus naturally defines the process that concerns us, $y_t = f(x_t)$. The ordered pair $Y = (f, x)$ is called a function of the finite Markov chain.

The outcome of an experiment is a sequence of observable responses $d_1 d_2 \cdots d_n$, for a finite experiment with n trials. Define a matrix $C(d_k)$ whose (i, j)th entry is one if $i = j$ and $j \in f^{-1}(d_k)$, and zero otherwise. That is $C(d_k)$ has a one in each diagonal entry corresponding to a state that is associated with the response d_k and zero everywhere else. Also define a column vector Σ whose entries are all ones. If the Markov chain X has N states, then $C(d_k)$ is $N \times N$ and Σ is an N-vector. Now denote a sequence of observed responses $d = d_1 \cdots d_n$. The likelihood function of Y, is given by

$$L(d; Y) = vC(d_1)\, AC(d_2)\, A \cdots AC(d_n)\Sigma. \tag{1}$$

To see that (1) is correct, consider the first trial, with response d_1. Clearly,

$$L(d_1; Y) = vC(d_1)\Sigma,$$

the sum of probabilities of the states in I that give response d_1. Now consider a two-trial experiment with outcomes d_1d_2. The likelihood is

$$L(d_1d_2;\, Y) = vC(d_1)\, AC(d_2)\Sigma.$$

The term $vC(d_1)$ is a vector with nonzero probabilities for the states associated with d_1 and the rest zero. $vC(d_1)A$ is the probability vector for the states in I on trial 2, when it is certain that d_1 occurred on trial 1. $vC(d_1)\, AC(d_2)$ modifies $vC(d_1)A$ by placing zeros for all the states that do not give response d_2. Then the likelihood is the sum of these probabilities. In general, if ν_{n-1} is the vector of probabilities of the states in I for trial $n-1$, when it is certain the first $n-1$ outcomes were $d_1 \cdots d_{n-1}$, then $\nu_n = \nu_{n-1}AC(d_n)$, and this sketches an inductive proof of equation (1).

An easy extension of this reasoning gives the likelihood function for an infinite sequence in D that contains a finite number of all but one element d_n. Let $d = d_1 \cdots d_nd_n \cdots$. The likelihood is

$$L(d;\, Y) = \lim_{m\to\infty} vC(d_1)A \cdots AC(d_{n-1})(AC(d_n))^m\, \Sigma.$$

A sequence of this kind would arise with positive probability only if $f^{-1}(d_n)$ were an absorbing class. In the usual case, $f^{-1}(d_n)$ contains an absorbing state and then the limit of $(AC(d_n))^m$, as m grows, is a matrix with one at the diagonal entry for the absorbing state, and zeros elsewhere.

In general, given a function of a Markov chain, Y, a matrical representation of Y is a 4-tuple,

$$M_Y = (v, \{C(d_i)\colon d_i \in D\}, A, \Sigma)$$

and the entries of M_Y are sufficient to determine the likelihood function, (1). We now consider the class of all 4-tuples of the form $(w, \{G(d_i)\colon d_i \in D\}, B, S)$ where w is an N-element row vector, $G(d_i)$ and B are $N \times N$ matrices, and S is an N-element column vector, and where substitution into (1) gives an acceptable likelihood function (that is, a number between zero and one for all permissible substitutions of response sequences in the sequence $d_1d_2 \cdots d_n \cdots$, and the sum over all such sequences is one). Within the class of all such 4-tuples, there are many that give the same likelihood function as M_Y while not conforming to the matrix restrictions discussed previously. To obtain a new matrical representation $M_{Y'}$ that gives the same likelihood function as M_Y, find a nonsingular $N \times N$ matrix, P, and compute the conjugate operators and new vectors

$$\begin{aligned} v' &= vP, \qquad C'(d_i) = P^{-1}C(d_i)P, \\ A' &= P^{-1}AP, \qquad \Sigma' = P^{-1}\Sigma. \end{aligned} \tag{2}$$

Then the matrical representation $M_{Y'} = (v', \{C'(d_i): d_i \in D\}, A', \Sigma')$ generates the same likelihood function as M_Y, for it is obvious that

$$\begin{aligned} L(d_1 d_2 \cdots d_n; Y) &= vC(d_1)\, AC(d_2)\, A \cdots AC(d_n)\Sigma \\ &= v'C'(d_1)\, A'C'(d_2)\, A' \cdots A'C'(d_n)\Sigma'. \end{aligned}$$

To summarize this result we state the following.

PROPOSITION 1. *Let M and M' be matrical representations of a function of a finite Markov chain. If there exists a nonsingular change matrix, P, such that* (2) *holds, then the likelihood functions of M and M' are identical.*

In other words, M and M' are equivalent in likelihood if they are related through (2) by a nonsingular P. Equation (2) states that the square matrices are similar. The similarity relation is obviously an equivalence and any two similar matrices represent the same underlying linear transformation simply written with respect to a different coordinate system or basis. Postmultiplying v and premultiplying Σ by P and P^{-1} respectively has the effect of rewriting these vectors with respect to that same basis.

Although Proposition 1 permits us to generate a very large number of matrical representations with identical likelihood functions, we have a special interest in matrical representations that are possible theories with psychological content. We attempt to specify this class of matrices by the following.

DEFINITION. Let $M = (w, \{G(d_i): d_i \in D\}, B, S)$ be a matrical representation of a function of a finite Markov chain. Then M is a Markov theory (MT) if and only if

(1) All entries of w and B are nonnegative.

(2) The row sum of w and all the row sums of B equal one.

(3) The (i, j)th entry of $G(d_k)$ is one when $i = j \in f^{-1}(d_k)$ and zero otherwise.

(4) The column vector S has all its entries equal to one.

It is important to ask whether it is possible for two matrical representations with identical likelihood functions to both be MT's. Using Proposition 1 we can often generate new MT's from old ones. A number of examples will be given below.

We add here a remark about necessity. A converse proposition of Proposition 1 is that for any two matrical representations M and M' with the same state space, if M and M' are identical in likelihood, then M' can be obtained from M through (2) with some nonsingular change matrix P. This proposition is true under fairly general conditions, having been proved by Erickson (1970). We do not present Erickson's proof, since it

requires considerable development of notation. Briefly, Erickson showed that if two matrical representations M and M' are of the same size and are both related to a set of observable responses through the same state-response matrices, and if the size of M and M' equals the rank[3] of the matrical representation of M, then there is a similarity transformation yielding M' from M. When the condition of rank(M) = size(M) holds, Erickson's work appears to show the converse of Proposition 1.

The next proposition states some necessary conditions that a change matrix must satisfy if it is to generate a new MT from a given MT. Let the set of observable outcomes be given as $\{d_1, ..., d_i, ..., d_K\}$, let rank $C(d_i) = r_i$, and let $q_i = \sum_{h=1}^{i-1} r_h$. We assume without loss of generality that $r_i \leqslant r_j$ $(\forall i \leqslant j)$, and further we let the response function matrix for the ith observable outcome be

$$
\begin{array}{r}
\\
q_i \\
\\
\\
C(d_i) = q_i + r_i \\
\\
\\
N - q_{i+1}
\end{array}
\left[\begin{array}{c|c|c}
0 & & \\
\ddots & 0 & 0 \\
0 & & \\
\hline
 & 1 & \\
0 & \ddots & 0 \\
 & 1 & \\
\hline
 & & 0 \\
0 & 0 & \ddots \\
 & & 0
\end{array}\right]
$$

This can always be achieved by simply relabeling the states.

PROPOSITION 2. *Let M and M' be Markov theories satisfying* (2), *both written in standard form. Then the nonsingular change matrix P must be in block diagonal form where the* ith *block has dimension $r_i \times r_i$. In addition, the sum of entries of each row of P must be one.*

Proof. Let p_{jk} be the (j, k)th entry of P. Since $P\Sigma = \Sigma$, the jth entry of $P\Sigma$, which equals $\sum_{k=1}^{N} p_{jk}$, must be one, and thus the row sum of P is one.

The rank of $C(d_i)$ equals the rank of $C'(d_i)$ because they are similar; therefore, we may conclude that $C(d_i) = C'(d_i)$. Equation (2) implies that $PC(d_i) = C(d_i)P$.

[3] The definition of the rank of a function of a finite Markov chain is given in Erickson's paper, while the definition of size of $Y = (f, x)$ is simply the order of the image of f.

Note that

$$PC(d_i) = \begin{pmatrix} & p_{1,q_i+1} & \cdots & p_{1,q_i+r_i} & \\ & p_{2,q_i+1} & \cdots & p_{2,q_i+r_i} & \\ 0 & & & & 0 \\ & \vdots & & \vdots & \\ & p_{N,q_i+1} & \cdots & p_{N,q_i+r_i} & \end{pmatrix}$$

$$C(d_i)P = \begin{pmatrix} & & 0 & & \\ p_{q_i+1,1} & p_{q_i+1,2} & \cdots & p_{q_i+1,N} \\ \vdots & \vdots & & \vdots \\ p_{q_i+r_i,1} & p_{q_i+r_i,2} & \cdots & p_{q_i+r_i,N} \\ & & 0 & & \end{pmatrix}.$$

Let $R_i = \{j: q_i < j \leqslant q_i + r_i\}$. Clearly, if $j \in R_i$ and $k \notin R_i$, then $p_{jk} = 0$ and $p_{kj} = 0$. But since this holds for all i, the only nonzero p_{jk} have both j and k as members of R_i for some i. This proves that the change matrix is in block diagonal form.

Now we consider the question of reducing the size of the state space. The preceding discussion has dealt with relations between matrical representations that have the same number of states. But it sometimes happens that the likelihood functions generated by two theories are identical when the state spaces are not of the same size. An example is the theory of simple memorizing having a state for short-term retention, which for conditions often used in experiments cannot be distinguished from simple all-or-none learning where only a single state represents all correct responses that occur before learning (see Greeno, 1967; Steiner & Greeno, 1969).

We simply remark here that the notion of collapsibility of a function of a Markov chain, introduced by Burke and Rosenblatt (1958) and discussed further by Kemeny and Snell (1960), can now be viewed in slightly greater generality. A Markov chain $X = (x_t, I)$ is collapsible through the function $g: I \to J$ if the function of X, $Y = (g, x)$ is a Markov chain. Burke and Rosenblatt showed that a sufficient condition for collapsibility is that for any j, $k \in J$, $P(y_{t+1} = k \mid x_t = h_1) = P(y_{t+1} = k \mid x_t = h_2)$ for $h_1, h_2 \in g^{-1}(j)$. Kemeny and Snell showed that the condition is necessary for collapsibility to obtain over all possible initial vectors of the Markov chain. From our earlier discussion, it is apparent that if one matrical representative is collapsible to another, then all the matrical representatives associated with each of the original representatives by (2) are equivalent with respect to likelihood. In effect, then, collapsibility between any two matrical representatives can be viewed as collapsibility between their respective equivalence classes with respect to likelihood as defined by equation (2).

APPLICATIONS

We present three examples to illustrate the concepts developed above. We would like to point out a unifying theme in our approach to all of these examples.

Although no algorithm has been given for reducing the parameter space of an MT, an heuristic that we have used successfully is to attempt to diagonalize the submatrices of the transition matrix that define transitions between states associated with the same response.[4] In the notation given in equation 3, below, the matrices to try to diagonalize are W and Z.

$$A = \begin{array}{c} f^{-1}(d_1)\begin{cases} S_1 \\ \vdots \\ S_{r_1} \end{cases} \\ f^{-1}(d_2)\begin{cases} S_{r_1+1} \\ \vdots \\ S_{r_1+r_2} \end{cases} \end{array} \left(\begin{array}{c|c} W & X \\ \hline Y & Z \end{array}\right). \tag{3}$$

There are two good reasons why this heuristic is useful. Firstly, diagonalizing these submatrices usually reduces the number of nonconstant entries in these submatrices. This may reduce the overall number of parameters unless a larger number of nonconstant parameters are introduced into the off-diagonal submatrices of the representation that results. There is no reason that the overall number must be reduced by diagonalizing W and Z. However, often that is the case.

The second reason is more subtle. Diagonalizing these submatrices is one means of rewriting the matrical representation in a form that can be called "observable." (See, e.g., Restle & Greeno, 1970.) By observable we mean that given the observed sequence of responses, the theoretical state on each trial is definitely determined. Thus, if the observable response on trial k, d_k, along with response history $d_1 \cdots d_{k-1}$, determines that on trial k the underlying Markov process must have been in one particular state, the model is observable. Obviously, an observable model has all of its nonconstant entries exactly estimable from data, and in an observable model, these entries constitute a minimally sufficient parameter set. Since diagonalizing these submatrices gives a model in which transitions are disallowed between states with the same response, this rewriting is in the direction of rendering the model observable,[5] and the parameter space may be reduced.

We now proceed with three examples. First, consider two theories of all-or-none learning, Model I assuming that learning can occur only following errors (Bower & Trabasso, 1964; Restle, 1962). Each of the theories has three states: L, an absorbing

[4] The usefulness of this technique was pointed out to us by Richard Schensted.

[5] The reader may convince himself that in general for a model to be in observable form each state may have only one nonzero exit probability to each class of the partition defined by the inverse image of f, i.e., the partition whose typical class is of the form $f^{-1}(d_k)$.

state where correct responses occur; S, a transient state where correct responses occur; and E, a transient state where errors occur. Initial vectors and transition matrices for the two cases are as follows:

Model I:

$$v_1 = (p,\ q,\ 1 - p - q)$$
$$A_1 = \begin{pmatrix} 1 & 0 & 0 \\ a & (1-a)g & (1-a)(1-g) \\ a & (1-a)g & (1-a)(1-g) \end{pmatrix}$$

Model II: (4)

$$v_2 = (r,\ s,\ 1 - r - s)$$
$$A_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & j & 1-j \\ k & (1-k)j & (1-k)(1-j) \end{pmatrix}.$$

Since errors occur only in the third state, the state-response matrices for both theories are as follows:

$$C_{\text{correct}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad C_{\text{error}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

Models I and II are equivalent in likelihood, as Greeno and Steiner (1964, 1968) showed, when corresponding values of their parameters are used. (It follows that the theories cannot be distinguished by binary data.) Both Model I and Model II are equivalent in likelihood to the following theory in the same states:

Model III:

$$v_3 = (t,\ u,\ 1 - t - u)$$
$$A_3 = \begin{pmatrix} 1 & 0 & 0 \\ c & (1-c)f & (1-c)(1-f) \\ d & (1-d)f & (1-d)(1-f) \end{pmatrix}. \tag{6}$$

The state-response matrices for Model III are also those given by Eq. (5). To show that these three models are equivalent we apply Eq. (2) using the following change matrix.

$$P_3 = \begin{pmatrix} 1 & 0 & 0 \\ \alpha & 1-\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{7}$$

$$\alpha = \frac{c}{1 - (1-c)f}, \text{ and obtain}$$

Model III′:

$$v_{3'} = v_3 P_3 = (t + u\alpha,\ u(1-\alpha),\ 1 - t - u)$$

$$A_{3'} = P_3^{-1} A_3 P_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & (1-c)f & 1-(1-c)f \\ d + (1-d)f\alpha & (1-d)f(1-\alpha) & (1-d)(1-f) \end{pmatrix}. \quad (8)$$

Clearly, Model III′ is equivalent to Model II except for notation.

Thus Model III is equivalent to Model II which in turn is equivalent to Model I. The substitution relations of parameters of Model II and Model III are

$$j = (1-c)f,$$
$$k = \frac{(1-d)fc}{1-(1-c)f} + d.$$

The relationships between parameters of Model II and Model I are

$$j = (1-a)g,$$
$$k = \frac{(1-a)ga}{1-(1-a)g} + a.$$

Using these substitutions it can be shown that choosing probabilities as parameters in one of the three models leads to parameters in the other two equivalent models that are also probabilities. As noted above, the parameters of an equivalent matrical representation are not always probabilities. However in this case they are and all models do satisfy our definition of a Markov theory. Since Model III is a theory with five parameters, and the theory is equivalent in likelihood to other models having only four parameters, we see that the parameters of Model III are not minimal; that is, (6) is not identifiable.

A second example shows a reduction in the state space of a theory. A model assuming all-or-none learning, but with a state corresponding to short-term retention, has been studied by Atkinson and Crothers (1964), Bernbach (1965), Greeno (1967), Greeno and Steiner (1969), and Kintsch (1966). The states are L, H, and S, giving correct responses, and E, giving errors. Initial and transition probabilities are as follows:

Model IV:

$$v_4 = (t,\ (1-t)s,\ (1-t)(1-s)g,\ (1-t)(1-s)(1-g)),$$

$$A_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ a & (1-a)h & (1-a)(1-h)g & (1-a)(1-h)(1-g) \\ b & (1-b)h & (1-b)(1-h)g & (1-b)(1-h)(1-g) \\ b & (1-b)h & (1-b)(1-h)g & (1-b)(1-h)(1-g) \end{pmatrix}. \quad (9)$$

The state-response matrices are

$$C_{\text{correct}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad C_{\text{error}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It has been shown that Model IV is equivalent in likelihood to a three-state theory in the form of Model II (Greeno, 1967). For this to be true, there must be a four-state theory, equivalent in likelihood to Model IV, that is collapsible into Model II. One such four-state theory was found by Greeno, Millward, and Merryman (1971), with transition parameters

$$A_{4'} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & x & y & 1-x-y \\ 0 & x & y & 1-x-y \\ z & v & w & 1-s-v-z \end{pmatrix}, \tag{10}$$

$$x = (1-a)h, \qquad y = (1-b)(1-h)g,$$

$$z = \frac{ah + b(1-h)}{1-x-y}, \qquad v = \frac{(1-a)(1-b)\,b(1-h)(1-g)}{1-x-y},$$

$$w = \frac{(1-b)^2\,(1-h)^2\,g(1-g)}{1-x-y}.$$

A change matrix that transforms A_4 into $A_{4'}$ is

$$P_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \alpha & 1-\alpha & 0 & 0 \\ \beta & 0 & 1-\beta & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{11}$$

$$\alpha = \frac{a-(a-b)(1-h)g}{1-(1-a)h-(1-b)(1-h)g}, \qquad \beta = \frac{ah + b(1-h)}{1-(1-a)h-(1-b)(1-h)g}.$$

$A_{4'}$, clearly is collapsible, so Model II is reached with $j = x + y$, and $(1-k)j = v + w$.

For the final example, we consider the transition matrix of a quite general model of two-stage learning.

Model V:

$$A_5 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ c & (1-c)i & 0 & (1-c)(1-i) & 0 \\ ae & a(1-e)h & (1-a)g & a(1-e)(1-h) & (1-a)(1-g) \\ d & (1-d)i & 0 & (1-d)(1-i) & 0 \\ bf & b(1-f)h & (1-b)g & b(1-f)(1-h) & (1-b)(1-g) \end{pmatrix}, \quad (12)$$

$$C_{\text{correct}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad C_{\text{error}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

A change matrix that gives another Markov learning theory is

$$P_5 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \alpha & 1-\alpha & 0 & 0 & 0 \\ \beta & \gamma & 1-\beta-\gamma & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \delta & 1-\delta \end{pmatrix},$$

$$\alpha = \frac{c}{1-(1-c)i}, \qquad \beta = \frac{ae[1-(1-c)i] + a(1-e)\,hc}{[1-(1-c)i][1-(1-a)g]}, \quad (13)$$

$$\gamma = \frac{a(1-e)\,h(1-c)(1-i)}{[1-(1-c)i][(1-c)i-(1-a)g]}, \qquad \delta = \frac{b(1-f)(1-h)}{(1-d)(1-i)-(1-b)(1-g)}.$$

The transition matrix obtained by applying P_5 is

$$A_{5'} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & (1-c)i & 0 & 1-(1-c)i & 0 \\ 0 & 0 & (1-g)g & 1-(1-a)g-x & 0 \\ 1-(1-d)(1-i)-y & y & 0 & (1-d)(1-i) & 0 \\ 1-(1-b)(1-g)-w-z & z & w & 0 & (1-b)(1-g) \end{pmatrix}, \quad (14)$$

$$x = \frac{(1-\delta)(1-a)(1-g)}{1-\beta-\gamma}, \qquad y = \frac{(1-c)(1-i)(1-d)i}{1-(1-c)i},$$

$$w = \frac{(1-\beta-\gamma)(1-b)g}{1-\delta}, \qquad z = \frac{(1-\alpha)[b(1-f)h-\delta(1-d)i] + \gamma(1-b)g}{1-\delta}.$$

Note first that $A_{5'}$ contains the eigenvalues of the submatrices of transition among states giving correct responses, and among states giving errors. The parameterization obtained thus reduces the parameter space by fixing some of the parameters at zero. Further, some of the dependencies among parameters are maintained. In both A_5 and $A_{5'}$, the entries (3, 3), (3, 5), (5, 3), and (5, 5) form a matrix with determinant zero, as do the entries (2, 2), (2, 4), (4, 2), and (4, 4). Taking these two dependencies into account, we see in $A_{5'}$ that the two-stage theory given as (12) has at most six identifiable parameters in its transition matrix.

Conclusions

The main development in this paper is a concise way of characterizing classes of Markov theories that are equivalent in likelihood. The fact that similarity equivalence classes are contained in equivalence classes determined by the likelihood function demonstrates a connection of the analysis of Markov learning theories with a fundamental aspect of linear algebra.

One implication is that we now have a convenient way of generating new matrical representations equivalent in likelihood to a given model. This is potentially useful in applications, since equivalent versions of a model may exist in a form that is more convenient for calculations. It is also useful to be able to determine whether different versions of a model, having different psychological interpretations of the states, are empirically distinguishable. The existence of a change matrix relating two models as in (2) is sufficient to demonstrate the models are not distinguishable.

A second implication involves use of the equivalence relation in studying the identifiability of a given model. A finding of considerable potential application is that for some models, a set of sufficient parameters includes the eigenvalues of the submatrices of states that have the same response. We still do not have an algorithm for finding minimal sufficient parameters for a model, but the present results give us one further step in the process of reducing a nonidentifiable parameter space.

References

Atkinson, R. C., and Crothers, E. J. A comparison of paired-associate learning models having different acquisition and retention axioms. *Journal of Mathematical Psychology*, 1964, **1**, 285–315.

Bernbach, H. A. A forgetting model for paired-associate learning. *Journal of Mathematical Psychology*, 1965, **2**, 128–144.

Bower, G. H. Application of a model to paired-associate learning. *Psychometrika*, 1961, **26**, 255–280.

Bower, G. H., and Trabasso, T. R. Concept identification. In R. C. Atkinson (Ed.), *Studies in mathematical psychology*. Stanford, CA: Stanford University Press, 1964. Pp. 32–94.

BURKE, C. J., AND ROSENBLATT, M. A Markovian function of a Markov chain. *Annals of Mathematical Statistics*, 1958, **29**, 1112–1122.

CHUNG, K. L. *Markov chains with stationary transition probabilities*. Berlin: Springer-Verlag, 1960.

ERICKSON, R. V. Functions of Markov chains. *Annals of Mathematical Statistics*, 1970, **41**, 843–850.

GREENO, J. G. Paired-associate learning with short term retention: Mathematical analysis and data regarding identification of parameters. *Journal of Mathematical Psychology*, 1967, **4**, 430–472.

GREENO, J. G. Identifiability and statistical properties of two-stage learning with no successes in the initial state. *Psychometrika*, 1968, **33**, 173–215.

GREENO, J. G., MILLWARD, R. B., AND MERRYMAN, C. T. Matrix analysis of identifiability of some finite Markov chains. *Psychometrika*, 1971, **36**, 389–408.

GREENO, J. G., AND STEINER, T. E. Markovian processes with identifiable states: General considerations and application to all-or-none learning. *Psychometrika*, 1964, **29**, 309–333.

KEMENY, J. G., AND SNELL, J. L. *Finite Markov chains*. Princeton, NJ: Van Nostrand, 1960.

KINTSCH, W. Recognition learning as a function of the length of the retention interval and changes in the retention interval. *Journal of Mathematical Psychology*, 1966, **3**, 412–433.

KOOPMANS, T. C. (Ed.), *Statistical inference in dynamic economic models*. Cowels Commission Monograph No. 10. New York: Wiley, 1950.

RESTLE, F. The selection of strategies in cue learning. *Psychological Review*, 1962, **69**, 329–343.

RESTLE, F., AND GREENO, J. G. *Introduction to mathematical psychology*. Reading, MA: Addison-Wesley, 1970.

STEINER, T. E., AND GREENO, J. G. An analysis of some conditions for representing n state Markov models. *Psychometrika*, 1969, **34**, 461–488.

RECEIVED: June 15, 1973