#446-77

INSTITUTE FOR RESEARCH ON POVERTY DISCUSSION PAPERS

FITTING STOCHASTIC MODELS TO LONGITUDINAL SURVEY DATA --SOME EXAMPLES IN THE SOCIAL SCIENCES

Burton Singer and Seymour Spilerman

UNIVERSITY OF WISCONSIN - MADISON

Fitting Stochastic Models To Longitudinal Survey Data

--Some Examples in the Social Sciences

Burton Singer

Columbia University
U.S.A.

Seymour Spilerman

Russell Sage Foundation
U.S.A.
On leave from University of Wisconsin

The work reported here was supported by National Science Foundation grants SOC76-17706 at Columbia University and SOC76-07698 at University of Wisconsin-Madison. Support from the Institute for Research on Poverty at the University of Wisconsin is also gratefully acknowledged.

ABSTRACT

An important feature of longitudinal data which has no counterpart in cross-sectional surveys is that one may carry out empirical studies in which individual histories are the basic unit of analysis. This opportunity for research aimed at understanding individual economic and social dynamics has focused attention on the dearth of analytical tools which are available for exploiting this unique feature of longitudinal data. We present examples, arising in the social sciences, of some new procedures for testing a commonly occurring form of longitudinal data --(multi-wave panel data)-- for compatibility with continuous time Markov chain models and mixtures of them. The tests exhibited herein are the simplest prototype of analytical procedures which are in serious need of development, particularly for assessing and characterizing path dependencies in individual histories.

Fitting Stochastic Models to Longitudinal Survey Data

--Some Examples in the Social Sciences

1. INTRODUCTION

The recent availability of large longitudinal data sets has focused attention on the dearth of analytical tools which are available for exploiting the unique features of such data. Particularly prominent among existing longitudinal surveys are the National Longitudinal Survey of Labor Force Experience (Parnes [14]) and the Michigan Panel Study of Income Dynamics (Morgan [13]), each of which attempts to measure various facets of the labor force experience of individuals over a substantial portion of their lives. Also of considerable interest are the National Crime Survey and several victimization surveys as described in Fienberg, [7]. An important feature of these data sets which has no counterpart in cross-sectional samples is that one may carry out empirical studies in which individual histories --or household histories-- are the basic unit of analysis. This focus immediately highlights new kinds of questions which can be answered with longitudinal data and that cannot be addressed otherwise. For example, in the context of labor force participation, accurate individual histories can be utilized to construct distributions of the durations of employment and unemployment for persons in particular occupational groups, age ranges, and geographical regions. Conditional probabilities of persons transferring from one job category to another given their age and earlier employment

history can be computed from work history data, whereas these probabilities are empirically outside the scope of cross-sectional surveys.

If quantities such as durations of unemployment and the above mentioned conditional probabilities of transfer between pairs of jobs are judged to be of central importance when a longitudinal survey is being planned, then continous histories for each individual represent the ideal form of data collection. Thus in the employment, unemployment, out of the labor market trichotomy it would be desirable to know in which of these states each individual is situated for all times after, say, age 17. Unfortunately, few longitudinal surveys have been designed with such questions in mind --for an exception see the retrospective survey of Coleman, et al., [6]. As a result, if questions answerable in terms of detailed individual histories are of interest to a researcher, he is usually confronted with data where the histories contain gaps of various kinds. The methodological issue then is how to utilize such fragmentary data to test theories of individual movement which incorporate both the observed and unobserved events.

The purpose of this paper is to outline and illustrate a general conceptual framework for such tests utilizing a commonly occurring form of fragmentary data from multi-wave panel studies. In Section 2 we describe a form of discrete-state multi-wave panel data in detail and present a formal significance test of the null hypothesis

H_0: Two waves of panel data on a two state process were generated by a continuous time Markov chain.

This test represents the simplest prototype of formal inferential methods to assess whether observed histories with gaps could have been generated by at least one member of a family of stochastic process models.

Unfortunately, as discussed in Section 3, there has been very little development of appropriate inferential methods for multi-wave panel data. As a result, we are usually forced to rely on formal

procedures, which can, nevertheless, be quite informative about dynamic processes which could have generated some sort of fragmentary data.

Sections 4-6 illustrate such informal tests and describe some unsolved problems whose solution would place the currently available methods for the analysis of longitudinal survey data on a firmer foundation.

2. INDIVIDUAL HISTORIES AND PANEL DATA--THE SIMPLEST PROTOTYPE

Let S be a finite set containing r elements, each of which is identified with a possible state of a stochastic process. Then define

$$\Omega = \{\omega:\omega(t),\ t\geq 0 \text{ is a step function taking values in } S\} .$$

One of the most frequently enountered forms of longitudinal data is a panel survey of observations from Ω of the form

$$\{\omega_j(k\Delta),\ 1 \leq j \leq N, \text{ where } N = \text{number of persons surveyed},\ 0 \leq k \leq n, \text{ where } n+1 = \text{number of waves in the panel study},\ 0 < \Delta = \text{spacing between observations}\} . \qquad (1)$$

Thus the states $\{\omega_j(t),\ t\neq k\Delta\}$ occupied by individual j are not observed, and (1) represents fragmentary information about the movement of the N individuals. Complete individual histories over a time interval [0,T] would be $\{\omega_j(t),\ 0\leq t\leq T,\ 1\leq j\leq N\}$; such data is rarely available in economic and sociological surveys --(but see Coleman, et. al. [6] for an exception).

Now let

$$\underset{\sim}{\overline{p}} = \{\overline{p}_{\underline{i}} = \text{Prob}\left[\omega:\omega(0)=i_0,\ \omega(\Delta)=i_1,\ \ldots,\ \omega(n\Delta)=i_n\right]$$
$$\underline{i} = (i_0, \ldots, i_n) \in S^{n+1}\}$$

be a probability measure on S^{n+1} = (n+1)-fold Cartesian product of S with itself, and observe that the maximum likelihood estimate of $\{\bar{p}_{\underline{i}}, \underline{i} \in S^{n+1}\}$ using the data (1) is

$$\left\{ \hat{p}_{\underline{i}} = \frac{1}{N} \sum_{j=1}^{N} \psi_{\underline{i}}(j) , \quad \underline{i} \in S^{n+1} \right\} \tag{2}$$

where

$$\psi_{\underline{i}}(j) = \begin{cases} 1 & \text{if } (\omega_j(0), \ldots, \omega_j(n\Delta)) = \underline{i} \\ 0 & \text{otherwise} \end{cases} .$$

One of the primary objectives in the analysis of longitudinal survey data is to assess whether or not (1) could have arisen from observations on a stochastic process whose joint distributions at the sampling times 0, Δ, 2Δ, ..., $n\Delta$ belong to some parametric family

$$\left\{ \bar{p}(\underline{a}) \right\}_{\underline{a} \in A} = \left\{ \bar{p}_{\underline{i}}(\underline{a}) = \text{Prob}_{\underline{a}}(\omega : \omega(0) = i_0, \ldots, \omega(n\Delta) = i_n) \right\}_{\underline{a} \in A} \tag{3}$$
$$\underline{i} = (i_0, \ldots, i_n) \in S^{n+1}$$

where A is a subset of a finite dimensional Euclidean space.

This kind of assessment can be incorporated within the framework of classical hypothesis testing, by using algebraic characterizations of specific parametric families (3) to describe a null hypothesis. We illustrate this idea with a simple example, which is a useful prototype of the general problem of assessing whether multi-wave panel data could have been generated by a restricted class of stochastic process models.

Example 1:

Consider a two-state process with S = {1, 2}, and let n = 1 in the observation plan (1) --(i.e. a two-wave panel study). If our null hypothesis is that (1) was generated by some continuous time

Markov chain then we are asking whether $\{\overline{p}_{\underline{i}},\ \underline{i} \in S^2\}$ can be represented as

$$\overline{p}_{i_0,i_1} = \pi_{i_0}\, p_{i_0,i_1}$$

where

$$\pi_{i_0} = \sum_{i_1} \overline{p}_{i_0,i_1}$$

$$p_{i_0,i_1} = \frac{\overline{p}_{i_0,i_1}}{\pi_{i_0}} = \text{Prob}(\omega:\omega(\Delta)=i_1 \,|\, \omega(0)=i_0)$$

and the stochastic matrix

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

satisfies

$$\text{trace } P = p_{11} + p_{22} > 1. \tag{4}$$

Remarks:

(i) Condition (4) is an algebraic characterization of the class of all 2×2 stochastic matrices whose entries can be transition probabilities for a two-state continuous time Markov chain. This was first established for chains with stationary transition probabilities by D. G. Kendall (see Kingman [10], pg. 15) and then extended to include general two-state non-stationary chains by Goodman [8]. Goodman expresses (4) in the equivalent form

$$\det P = \text{trace } P - 1 > 0\ . \tag{4'}$$

(ii) A formal statistical test which rejects the null hypothesis

$$H_0: P \in \underset{\sim}{P} = \{P : \text{trace } P > 1\}$$

rejects all non-stationary, continuous time Markov chains as possible processes to generate (1) with n=1. However, if H_0 is accepted then a continuum of non-stationary continuous time chains can generate P while there is only one time-homogeneous chain which has transition probabilities

$$||p_{ij}(0,\Delta)|| = P .$$

In particular, the unique chain with stationary transition probabilities statisfying

$$||p_{ij}(0,\Delta) = \text{Prob}(\omega:\omega(\Delta)=j \,|\omega(0)=i)|| = P \in \underset{\sim}{P}$$

has transition probabilities for general times $0 \le s < t$

$$p_{ij}(s,t) = \text{Prob}(\omega:\omega(t)=j \,|\omega(s)=i) = \left(e^{(t-s)Q}\right)_{ij}$$

where

$$Q = \frac{1}{\Delta} \log P = \frac{1}{\Delta} \frac{\log(\text{trace } P-1)}{\text{trace } P-2} \begin{pmatrix} p_{11} - 1 & 1 - p_{11} \\ 1 - p_{22} & p_{22} - 1 \end{pmatrix} .$$

There is a natural parametrization of the 2×2 intensity matrices Q; namely,

$$\left\{Q:Q = \begin{pmatrix} -a_1 & a_1 \\ a_2 & -a_2 \end{pmatrix} \text{ where } a_i \ge 0, \quad i = 1,2 \right\} .$$

Thus we can express any $P \in \underset{\sim}{P}$ in the parametric form

$$p_{ij} = \left(e^{\Delta \begin{pmatrix} -a_1 & a_1 \\ a_2 & -a_2 \end{pmatrix}} \right)_{ij}$$

and thereby identify the joint distribution $\{\overline{p}_{\underline{i}}, \underline{i} \in S^2\}$ with a member of the parametric family $\{\overline{p}(\underline{a})\}_{\underline{a} \in A}$ where

$$\overline{p}_{ij} \equiv \overline{p}_{ij}(\underline{a}) = \pi_i \left(e^{\Delta \begin{pmatrix} -a_1 & a_1 \\ a_2 & -a_2 \end{pmatrix}} \right)_{ij}$$

and

$$\pi_i = \sum_j \overline{p}_{ij} \quad .$$

For a formal test of the null hypothesis $H_0: P \in \underset{\sim}{P} = \{P : \text{trace } P > 1\}$ introduce the alternative hypothesis

$$H_1: P \notin \underset{\sim}{P}$$

and the decision rule:

if

$$\begin{aligned} &\text{trace } \hat{P} > 1+\delta_1, && \text{accept } H_0; \\ 1-\delta_2 \leq &\text{trace } \hat{P} \leq 1+\delta_1, && \text{accept that the observations} \\ & && \text{are inadequate to discriminate} \\ & && \text{between } H_0 \text{ and } H_1; \\ &\text{trace } \hat{P} < 1-\delta_2, && \text{accept } H_1. \end{aligned}$$

Here

$$\hat{P} = \left|\left| \frac{n_{ij}}{n_{i+}} \right|\right| \equiv ||\hat{p}_{ij}||$$

n_{ij} = number of individuals in state j at time Δ who were in state i at time 0

$$n_{i+} = \sum_{j=1}^{2} n_{ij}$$

and the constants δ_1 and δ_2 are determined from an a priori specification of Type I and Type II error. Especially, we set

$$\alpha_1 = \text{Prob}_{H_0}(\text{reject } H_0) = \text{Prob}_{H_0}(\text{Type I error}) \tag{5}$$

$$= \sup_{\underline{p}: p_{11}+p_{22}>1} \text{Prob}_{\underline{p}}(\text{trace } \hat{P} < 1-\delta_2)$$

and

$$\alpha_2 = \text{Prob}_{H_1}(\text{accept } H_0) = \text{Prob}_{H_1}(\text{Type II error}) \tag{6}$$

$$= \sup_{\underline{p}: p_{11}+p_{22}\leq 1} \text{Prob}_{\underline{p}}(\text{trace } \hat{P} > 1+\delta_1)$$

where

$$\underline{p} = (p_{11}, p_{22}), \qquad 0 \leq p_{ii} \leq 1, \qquad i = 1,2$$

and $\{\alpha_i, i=1,2\}$ are specified by the researcher.

Simple approximate solutions to (5) and (6) are

$$\delta_1 \approx \frac{1}{2}\sqrt{\frac{1}{n_{1+}} + \frac{1}{n_{2+}}}\ \Phi^{-1}(1-\alpha_2)$$

and

$$\delta_2 \approx -\frac{1}{2}\sqrt{\frac{1}{n_{1+}} + \frac{1}{n_{2+}}}\ \Phi^{-1}(\alpha_1)$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{e^{-u^2/2}}{\sqrt{2\pi}}\, du \ .$$

Some preliminary numerical evidence indicates that this approximation is quite good when $\min(n_{1+}, n_{2+}) > 35$. Full analytical and numerical details about this and other tests of the hypothesis H_0 will be published elsewhere.

3. MORE COMPLICATED HYPOTHESES

For multi-wave panel data such as (1), formal significance tests of $H^{(\ell)}$:{(1) were generated by an ℓ^{th} order Markov chain} where $0 \leq \ell \leq n$ have been given by Anderson and Goodman [1]. Their investigation, however, pays no attention to the distinction between discrete and continuous time processes; hence, tests of the kind exhibited in example 1 have not been previously discussed. The point at which algebraic characterizations such as (4), that distinguish conditional probabilities for a continuous time process from those that can only arise in a discrete-time formulation, enter into test statistics --e.g. the generalized likelihood ratio-- is in specifying the region over which a supremum is to be computed.

Such computations present difficult numerical analysis problems which are far from resolved. In fact, the entire subject of formal inferential test procedures for stochastic process models with observation plans such as (1) is virtually undeveloped.

Because of this paucity of inferential methods we describe some informal test procedures in sections 4-6 which are based on algebraic characterizations of conditional probabilities generated by restricted classes of models. The procedures are analogous to the inferential test illustrated in example 1, except that there is no formal consideration of Type I and Type II errors. This, possibly excessive, reliance on the subjective judgment of a researcher to say when a hypothesis should be rejected is a consequence of the lack of a systematic sampling theory for the algebraic expressions utilized in the proposed tests.

4. UNOBSERVED MULTIPLE TRANSITIONS--AN EXAMPLE

As part of a study of interpersonal relationships among American high school youth in the 1950's, J. Coleman [3] asked students in Northern Illinois high schools in October 1957 and again in May 1958 whether or not:

(1) they perceived themselves to be members of the leading crowd in their school;

(2) they can maintain their principles and be a member of the leading crowd.

Affirmative answers to each question were scored + and negative answers were scored -. Thus, an individual can respond to the above questions in one of four possible ways at each observation time: (Response to (1), Response to (2)) = (+,+), or (+,-), or (-,+), or (-,-). We then identify these responses as possible states of a stochastic process. The observed counts for boys and girls based on the above mentioned two waves of panel data are:

TABLE I

Boys, Observed Counts

Question	(1)	(2)	Response, May 1958 + +	+ -	- +	- -
	+	+	458	140	110	49
Response	+	-	171	182	56	87
October 1957	-	+	184	75	531	281
	-	-	85	97	338	554

Source: Coleman [4], pg. 171

TABLE II

Girls, Observed Counts

Question	(1)	(2)	Response, May 1958 + +	+ -	- +	- -
	+	+	484	93	107	32
Response	+	-	112	110	30	46
October 1957	-	+	129	40	768	321
	-	-	74	75	303	536

Source: Coleman [4], pg. 168

Although the attitudes (1) and (2), held by each student, were assessed at times spaced nine months apart, their attitudes on these questions could have changed multiple times between October 1957 and May 1958. Such changes are, of course, non-observable. In connection with the above data, Coleman [4], pg. 168 utilized a theory about attitude changes in an adolescent population on issues such as (1) and (2). In particular, he suggested that individuals could change their attitude on either issue alone at any one time but could not change their attitude on both issues simultaneously.

Examination of Tables I and II reveals that in both the male and female populations some individuals had changed their attitude on both issues, as observed at the survey times --e.g. 32 girls responded (+,+) in October 1957 and (-,-) in May 1958; 75 boys responded (-,+) in October 1957 and (+,-) in May 1958. Since the times at which an individual changes his/her attitude is unrelated --to the best of our knowledge-- to the survey times, our only recourse in assessing conpatibility of data such as Table I and II with Coleman's theoretical proposition, is to first propose a variety of plausible models of individual attitude change which allow for transitions at arbitrary times. We then assess whether the observed data can --at least to within small errors-- be generated by one or more of the proposed models.

A simple baseline class of models which were suggested by Coleman for comparison with Tables I and II are continuous time Markov chains with stationary transition probabilities governed by the special 4×4 intensity matrices

$$Q \in \underset{\sim}{Q}_1 = \left\{ Q : q_{ii} < 0, \quad q_{ij} \geq 0, \; i \neq j, \quad \sum_{j=1}^{4} q_{ij} = 0 \right\} ;$$

$$q_{14} = q_{23} = q_{32} = q_{41} = 0$$

that is, instantaneous change is possible only on one attitude at a time. Transition probabilities P(0,t) for these models satisfy the matrix differential equations

$$\frac{dP}{dt} = QP, \qquad P(0) = I \tag{7}$$

where $Q \in \underset{\sim}{Q}_1$ (see Coleman [4] for the restricted class $\underset{\sim}{Q}_1$); and P(0,t) can be represented as

$$P(0,t) = e^{tQ} . \tag{8}$$

Note: Transition probabilities between a pair of states conditional on a transition occurring --whether it is observed or not-- are given by $m_{ij} = q_{ij}/(-q_{ii})$, $i \neq j$.

In order to assess whether the transition matrices induced by Tables I and II according to

$$\hat{P}(0,\Delta) = \left|\left| \frac{n_{ij}}{n_{i+}} \right|\right| \tag{9}$$

(here, n_{ij} = number of individuals in state i in October 1957 who are also in state j in May 1958,

$$n_{i+} = \sum_{j=1}^{4} n_{ij} ,$$

and Δ = 9 months)

can be approximately represented in the form (8), we introduce the matrix norm

$$||A|| = \sqrt{\sum_i \sum_j |a_{ij}|^2}$$

and determine Q_{boys} and Q_{girls} for which

$$\min_{Q \in \underset{\sim}{Q}_1} ||\log \hat{P}_{boys} - Q||$$

$$\min_{Q \in \underset{\sim}{Q}_1} ||\log \hat{P}_{girls} - Q||$$

are attained.

The primary quantities of interest are the probabilities

$$M = \begin{pmatrix} -\frac{1}{q_{11}} & & \underline{0} \\ & \ddots & \\ \underline{0} & & -\frac{1}{q_{44}} \end{pmatrix} Q + I$$

interpreted as probabilities of movement between pairs of states conditional on a change occurring. These probabilities are given in terms of the least squares intensity matrices, Q_{boys} and Q_{girls}, by

$$M_{boys} = \begin{pmatrix} 0 & .6148 & .3852 & 0 \\ .6546 & 0 & 0 & .3454 \\ .3561 & 0 & 0 & .6439 \\ 0 & .2133 & .7867 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{1}{(q_{11})_{boys}} & & \underline{0} \\ & \ddots & \\ \underline{0} & & -\frac{1}{(q_{44})_{boys}} \end{pmatrix} Q_{boys} + I$$

and similarly,

$$M_{girls} = \begin{pmatrix} 0 & .5361 & .4639 & 0 \\ .6897 & 0 & 0 & .3103 \\ .2367 & 0 & 0 & .7633 \\ 0 & .2344 & .7656 & 0 \end{pmatrix}$$

Computing tables of expected values under the model proposed by Coleman we obtain

$$\begin{pmatrix} n_{1+} & & \underline{0} \\ & \ddots & \\ \underline{0} & & n_{4+} \end{pmatrix}_{boys} e^{Q_{boys}} = \begin{pmatrix} 454.3 & 134.9 & 108.8 & 65.6 \\ 174.6 & 182.6 & 47.1 & 91.7 \\ 187.2 & 56.7 & 539.2 & 286.7 \\ 93.2 & 92.7 & 334.2 & 554.0 \end{pmatrix} \quad (10)$$

$$\begin{pmatrix} n_{1+} & & \underline{0} \\ & \ddots & \\ \underline{0} & & n_{4+} \end{pmatrix}_{girls} e^{Q_{girls}} = \begin{pmatrix} 479.1 & 90.6 & 104.7 & 41.6 \\ 111.8 & 112.4 & 22.4 & 51.5 \\ 124.8 & 38.0 & 770.6 & 324.5 \\ 57.6 & 77.5 & 305.6 & 544.4 \end{pmatrix} \quad (11)$$

Comparing (10) and (11) with Tables I and II reveals that constrained time-homogeneous Markov models with $Q \in \underset{\sim}{Q}_1$ provide very good approximations to this data. The key methodological lesson of these calculations is that observations on a process, where multiple transitions occur between the observation times, can still be effectively tested for compatibility with theoretical models which incorporate these non-observable events. Furthermore, the preliminary conclusions about the adolescent society listed below are much more transparent in M_{boys} and M_{girls} then in Tables I and II or in the transition matrices $\hat{P}(0,\Delta)$ induced by them. These conclusions are:

(i) The most probable transitions for both boys and girls are $(+,-) \to (+,+)$; $(-,+) \to (-,-)$; and $(-,-) \to (-,+)$.

(ii) Although both boys and girls who perceive themselves outside of the leading crowd and who don't feel you must give up on principles to be in it will tend to change their mind on the issue of principles, girls have a somewhat higher probability than boys of feeling this way. In particular, $(m_{34})_{girls} = .7633 > (m_{34})_{boys} = .6439$.

(iii) For persons perceiving themselves outside the leading crowd and feeling you must go against your principles to be in it, it is much more likely that they will change their attitude about the issue of principles before they are in the leading crowd than the reverse. (i.e. $m_{43} > m_{42}$ for both boys and girls)

Having demonstrated that a restricted class of time-homogeneous Markov models provides a readily interpretable and remarkably good approximation to the data in Tables I and II, it is necessary to add a note of caution. In particular, a variety of non-Markovian models of both homogeneous and heterogeneous populations, which are indistinguishable from time-homogeneous Markov models on the basis of two waves of panel data, can also represent Tables I and II. Thus we view the conclusions based on the preceding calculations as suggestive but tentative pending a comparison of

$$e^{(k-j)\Delta Q_{boys}} \quad \text{and} \quad e^{(k-j)\Delta Q_{girls}}$$

with observed matrices $\hat{P}_{boys}(j\Delta,k\Delta)$ and $\hat{P}_{girls}(j\Delta,k\Delta)$ arising from additional waves of the panel study.

5. MODEL REJECTION WITH 2-WAVES OF PANEL DATA --NON-STATIONARY CHAINS

A simple inequality which holds for the transition probabilities $||p_{ij}(s,t) = \text{Prob}(\omega:\omega(t) = j|\omega(s) = i)||$ of arbitrary continuous time non-stationary Markov chains can be converted into a rejection critera for a test of the null hypothesis

H_0 : observations of the form (1) with n=1 were generated by some continuous time non-stationary Markov chains.

In particular G. Goodman [8] proved that

$$\prod_{i=1}^{r} p_{ii} \geq \det||p_{ij}|| > 0 \tag{12}$$

for transition probabilities $p_{ij}(s,t)$ which are solutions of the matrix differential equations

$$\begin{aligned} \frac{\partial P(s,t)}{\partial t} &= P(s,t)Q(t), \qquad P(t,t) = I \\ \frac{\partial P(s,t)}{\partial s} &= -Q(s)P(s,t), \qquad P(s,s) = I \end{aligned} \tag{13}$$

and $\forall s \geq 0$, $Q(s) \in \underset{\sim}{Q} = \{Q:q_{ii} \leq 0,\ q_{ij} \geq 0,\ i \neq j,\ \sum_{1}^{r} q_{ij} = 0\}$.

Thus the entries of a stochastic matrix which satisfies

$$\prod_{i=1}^{r} p_{ii} < \det P \tag{14}$$

cannot be interpreted as transition probabilities of any continuous time non-stationary chain.

Example 2: An informal test

Suppose that

$$\hat{P}(0,\Delta) = \begin{pmatrix} .15 & .35 & .50 \\ .37 & .45 & .18 \\ .20 & .60 & .20 \end{pmatrix} \tag{15}$$

is a stochastic matrix whose entries are conditional frequencies from data of the form (1) with n = 1. Computing $\prod_{i=1}^{3} \hat{p}_{ii} = .0135$ and $\det \hat{P} = .05$ we find that (14) holds and that --provided (14) is based on a large sample-- we may reject H_0 on the basis of this data. A formal test of significance which uses

$$\prod_{i=1}^{r} \hat{p}_{ii} - \det \hat{P} \tag{16}$$

as a test statistic can be specified in principle; however, development of the necessary distribution theory for (16) based on multinomial sampling remains to be carried out.

The inequalities (12) are only necessary conditions for a finite stochastic matrix to be embeddable in a continuous 2-parameter family of matrices satisfying (13). Hence, they can only be used to specify criteria for model rejection. Unfortunately, no computationally tractable necessary and sufficient conditions are known for the general non-stationary Markov chain embedding problem (see Johansen [9] and Kingman [11] for an up to date account of this problem). However, if we restrict consideration to the non-stationary birth and death processes --i.e. continuous time Markov chains whose transition probabilities satisfy (5.3) but with

$$Q(s) \in \underset{\sim}{Q} \cap \{Q: q_{ij} = 0 \text{ if } |i-j| > 1\} = \underset{\sim}{Q}_2 \ \forall s \geq 0, \text{ --}$$

then we have the

Theorem: A non-singular stochastic matrix P is embeddable in a continuous two parameter family of stochastic matrices satisfying (13) with $Q(s) \in \underset{\sim}{Q}_2$ $\forall s \geq 0$ if, it is totally positive.

Note: A non-negative r×r matrix A is called totally positive if for

$$\begin{aligned} & i_1 < i_2 < \ldots < i_k \\ & j_1 < j_2 < \ldots < j_k \\ & \det\left|\left|a_{i_\ell j_m}\right|\right| \geq 0 \qquad (17) \\ & 1 \leq \ell \leq k \\ & 1 \leq m \leq k \end{aligned}$$

for $k = 2, 3, \ldots, r$.

Mathematical details concerning the above theorem will be published elsewhere. However, the importance of this result for the analysis of multi-wave panel data lies in the fact that tests of the validity of the inequalities (17) can readily be implemented on a computer. A much more difficult task is the development of a distribution theory for the determinants in (17) based on multinomial sampling. This task would be necessary if formal tests of significance are to replace the informal examination of the determinants in (17) as a means of assessing compatibility of two-wave panel data with birth and death process models.

6. RESIDUALS FROM TIME-HOMOGENEOUS MARKOV CHAINS

There is an extensive empirical literature in sociology and economics --(see Singer and Spilerman [15], [16] for a discussion and references)-- in which an observed stochastic matrix $\hat{P}(0,\Delta)$,

based on two waves of panel data, has been identified with a matrix of transition probabilities for a time-homogeneous Markov chain. Tests of the validity of this identification using several waves of panel data have generally led to its rejection and, at the same time, consistently revealed the inequality

$$\hat{f}(k,\Delta) = \text{trace } \hat{P}^k(0,\Delta) - \text{trace } \hat{P}(0,k\Delta) < 0 \qquad (18)$$

with $\hat{f}(k,\Delta)$ decreasing as k increases.

The most frequently utilized explanation for (18) is that a socially heterogeneous population is being treated as though it was homogeneous --(see, however, Coleman [5] and Singer and Spilerman [15] for a discussion of other explanations). This interpretation suggests that mixtures of Markov chains should provide a better description of the data and, thereby, account for (18). Indeed, in many investigations—(e.g. [2], [12])—(18) has provided excellent clues to readily interpretable mixtures of Markov chains which also fit the observed data quite well.

Motivated by this empirical success, it is natural to ask whether general mixtures of Markov chains must automatically satisfy

$$f(k,\Delta) = \text{trace } P^k(0,\Delta) - \text{trace } P(0,k\Delta) < 0 \qquad (19)$$

for $k = 2, 3, \ldots$ and $\Delta > 0$. In [16] we showed that (19) does <u>not</u> hold in general and exhibited wide classes of mixtures of Markov chains for which

$$f(k,\Delta) > 0, \quad k = 2, 3 \quad \text{and } \Delta \geq 1 . \qquad (20)$$

For a simple example of this behavior, consider the family of mixtures with transition probabilities

$$P(0,t) = sI + (1-s)M^t \qquad t = 0, 1, 2, \ldots \tag{21}$$

and

$$M \in \underset{\sim}{M} = \left\{ M : M = \begin{pmatrix} 1-(2\alpha+\beta) & \alpha & \beta & \alpha \\ \alpha & 1-(2\alpha+\beta) & \alpha & \beta \\ \beta & \alpha & 1-(2\alpha+\beta) & \alpha \\ \alpha & \beta & \alpha & 1-(2\alpha+\beta) \end{pmatrix} \right\}$$

$$0 < \alpha, \beta; \quad 2\alpha + \beta < 1$$

Each $M \in \underset{\sim}{M}$ has eigenvalues $\nu_1 = 1$, $\nu_2 = \nu_3 = 1-2(\alpha+\beta)$, and $\nu_4 = 1-2\alpha$. In terms of $\{\nu_i\}_{i=1,2,3,4}$ a somewhat tedious calculation reveals that with $\Delta = 1$ and $k = 3$,

$$f(3,1) = \text{trace } P^3(0,1) = \text{trace } P(0,3) > 0$$

whenever

$$-3 + 3 \sum_{i=2}^{4} \nu_i^2 - 2 \sum_{i=2}^{4} \nu_i^3 > 0$$

and

$$0 < s < s^*$$

where

$$s^* = \frac{3 - 3 \sum_{i=2}^{4} \nu_i^2 + 2 \sum_{i=2}^{4} \nu_i^3}{\sum_{i=2}^{4} (\nu_i - 1)^3}$$

The importance of such examples for the development of techniques for the analysis of multi-wave panel data lies in the questions they raise about possible interpretations of (18). For example, suppose we restrict attention to mixtures of the form

$$\underset{\sim}{P}_1 = \{P(0,t) : P(0,t) = \int_0^\infty e^{t\lambda(M-I)} d\mu(\lambda), \quad t \geq 0$$

where M is a stochastic matrix with distinct eigenvalues, and μ is an arbitrary probability measure on $[0,+\infty)$ }

or

$$\underset{\sim}{P}_2 = \{P(0,t):P(0,t) = \int_{\Lambda} M^{t\lambda} d\mu(\lambda), \quad t = 0, 1, 2, \ldots$$

where M is a stochastic matrix with distinct eigenvalues, and μ is an arbitrary probability measure on $\Lambda = [0, 1, 2, \ldots)\}$.

Note: $P(0,t)$ in equation (21) is in $\underset{\sim}{P}_2$.

Then the following issues arise:

(i) If $P(0,t) \in \underset{\sim}{P}_1 \cup \underset{\sim}{P}_2$ and $\forall t \geq 0$ has real positive eigenvalues, then (19) holds for $k = 2, 3, \ldots$ and $\forall \Delta > 0$. It is not known whether most of the observed matrices for which mixture models in $\underset{\sim}{P}_1 \cup \underset{\sim}{P}_2$ --or slight generalizations of these-- have been successfully utilized to account for (18) actually have real positive eigenvalues. There is certainly no clear a priori reason why observed matrices in a wide variety of longitudinal surveys should have this property.

(ii) Since (20) can occur for mixtures in $\underset{\sim}{P}_1 \cup \underset{\sim}{P}_2$ only when $P(0,\Delta)$ has at least one pair of complex conjugate eigenvalues or a negative real eigenvalue, it would be important to know how frequently such observed matrices arise in studies where a convincing substantive argument can be made for mixtures in $\underset{\sim}{P}_1 \cup \underset{\sim}{P}_2$ as a plausible class of models. If (18) occurs for many matrices having complex conjugate eigenvalues, then there must be severe restrictions on the classes of matrices M, mixing measures μ, and observation intervals Δ in mixture models which could have generated the data. Especially, it would be important to understand, form an empirical point of view, why the inequalities (20) exhibited by models such as (21) tend not to occur in economic

and sociological panel data, if this is in fact the case.

(iii) Finally, it should be remarked that even for data generated by mixtures in $\underset{\sim}{P}_1 \cup \underset{\sim}{P}_2$ with real positive eigenvalues, (18) can be violated simply as a consequence of sampling variability. A sampling theory for $\hat{f}(k,\Delta)$ which could place informal tests such as (18) on a stronger foundation remains to be developed.

BIBLIOGRAPHY

1. Anderson, T. W. & Goodman, L. (1957). Statistical inference about Markov chains. Ann. Math. Stat. 28:89-109.

2. Blumen, I., Kogan, M., & McCarthy, P. J. (1955). The Industrial Mobility of Labor as a Probability Process. Cornell Studies in Industrial and Labor Relations, vol. 6, Ithaca, New York: Cornell University Press.

3. Coleman, J. S. (1961). The Adolescent Society. New York: The Free Press.

4. Coleman, J. S. (1964). Introduction to Mathematical Sociology. New York: The Free Press.

5. Coleman, J. S. (1964). Models of Change and Response Uncertainty. Englewood Cliffs, New Jersey: Prentice Hall.

6. Coleman, J. S., Blum, Z. D., Sorenson, A, & Rossi, P. (1972). White and black careers during the first decade of labor force experience. Part I: Occupational status. Soc. Sci. Res. I(3): 243-270.

7. Fienberg, S. E. (1977). Victimization and the National Crime Survey: Problems of design and analysis. Technical Report No. 291, School of Statistics, University of Minnesota.

8. Goodman, G. S. (1970). An intrinsic time for non-stationary finite Markov chains. Z. Wahrscheinlichkeitstheorie 16:165-180.

9. Johansen, S. (1973). A central limit theorem for finite semi-groups and its application to the imbedding problem for finite-state Markov chains. Z. Wahrscheinlichkeitstheorie 26: 171-190.

10. Kingman, J. F. C. (1962). The imbedding problem for finite Markov chains. Z. Wahrscheinlichkeitstheorie 1:14-24.

11. Kingman, J. F. C. (1975). Geometrical aspects of the theory of non-homogeneous Markov chains. Math. Proc. Cambridge Philos. Soc. 77:171-183.

12. McCall, J. J. (1973). Income Mobility, Racial Discrimination, and Economic Growth, Lexington, Massachusetts: D. C. Heath.

13. Morgan, J. N. & Smith, J. D. (1969). A Panel Study of Income Dynamics, Institute for Social Research, Ann Arbor, Michigan.

14. Parnes, H. S. (1975). Sources and uses of panels of microdata --The national longitudinal surveys: New vistas for labor market research. Am. Econ. Rev. 65(2):244-249.

15. Singer, B. & Spilerman, S. (1976). Some methodological issues in the analysis of longitudinal surveys. Ann. Econ. Soc. Meas. 5(4):447-474.

16. Singer, B. & Spilerman, S. (1977). Trace inequalities for mixtures of Markov chains. Adv. Appl. Probab. (in press).

(Key words: Longitudinal Survey, Panel Data; Markov Chain, Hypothesis Testing, Embeddability)