CHANGE OVER TIME: OBSERVATIONAL STATE, MISSING DATA, AND REPEATED MEASURES IN THE GRADE OF MEMBERSHIP MODELL

Max A. Woodbury, Larry S. Corder, Kenneth G. Manton, Duke University
Larry S. Corder, Center for Demographic Studies, Duke University, 2117 Campus Drive, Box 90408,
Durham, North Carolina 27708-0408

KEY WORDS: Longitudinal analysis, missing data, Grade of Membership

ABSTRACT

In longitudinal studies a series of examinations are made of panels of subjects. Each "exam" represents the sampling of information at fixed times from a continuous time stochastic process. The temporal "density" of measurement determines how well the discrete time model, for which parameters are estimated, represents the underlying process. Since measurement is at fixed times there is missing data on the continuous process. Whether that missing data causes bias, and affects the "embeddability" of the discrete-time-measured-process in the continuous process, depends on the time between measurements, and whether it exceeds the Nyquist interval. That is., how much spectral energy (mean-square frequency amplitude) lies out side of the bandwidth implied by the Nyquist interval. This in turn depends on the time spent in measurement, and the rate at which the continuous time process generates events. We consider the nature of information lost during a data collection period vs. that lost outside of data collection periods. Each panel member is available for measurement only in specific field survey periods which do not overlap in time. Thus, data in a panel is nonsystematically missing before and after data collection periods. It is systematically missing if the respondent dies during the field survey period before assessment. The proposed model identifies parameters describing a.) individuals, b.) variables, and c.) the time trajectories of both. An example using the 1982, 1984, and 1989 NLTCS is provided.

INTRODUCTION

Fuzzy sets models can be used as a general non-parametric estimation strategy (no specific distributional assumptions are made) with the consistent estimation of individual scores (incidental parameters) is made possible by the geometric properties of the probability space defining the solution. With this general form, it is possible to specify models of temporal processes where individuals have their own identifiable trajectories. The identification of individual trajectories depends upon the density of measurements over time, the number of distinct variables measured, and the period of time during which observations are gathered. Censoring events occurring during the data collection period precludes measurements of the covariates and hence are more problematic than those occurring between measurement intervals. This is because the updating (and thus "completeness") of state information is not directly affected when "censoring" occurs outside of data collection periods.

We present a modification of a fuzzy set model to provide estimates of individual trajectories and to identify the effect of censoring on groups of persons within a data collection period.

THE MODEL

Fuzzy set models can be viewed as generalizations of statistical procedures used to analyze multiple discrete variables. For example, log-linear models are used to describe multi-way contingency tables (Bishop et al., 1975). In those models the cell probabilities, λ_{jl}, for J variables, each with L_j>2 categories, are estimated. If there are multiple independent populations then this is manifested by there being K independent sets of parameters λ_{kjl}. Alternately, the K independent populations may be represented by defining a population indicator variable, g_{ik}, for each individual. The g_{ik} is 1.0 if a person is in the kth population and 0.0 otherwise. The g_{ik}, in log-linear models, is assumed observed and measured without error.

One generalization of contingency table procedures is to assume that the K independent populations are unobserved (or that there is sufficient noise in the g_{ik}s such that groups cannot be directly identified). In this case whether a person is in group k, or not, is assessed by ML estimation procedures applied to the data (i.e., the latent class model; LCM; Lazarsfeld and Henry, 1968). Generally, instead of estimating g_{ik}s directly, the probability of being in a group (i.e., $\hat{p}_{ik}$= PROB(g_{ik} = 1.0)) is estimated. Thus, LCM resolves heterogeneity not represented by contingency tables by identifying K latent tables with cell probabilitiees λ_{kjl} and the probability of being in population k, $\hat{p}_{ik}$ (Everitt, 1984).

The fuzzy set model generalizes the LCM by suggesting that g_{ik} itself, (not the probability of g_{ik} = 1.0) varies between 0 and 1.0. This allows measurements on persons to be represented as convex mixtures of K latent J-way tables. Thus, the g_{ik} and λ_{kjl} must be estimated simultaneously. If the multinomial data can be coded as L_j binary variables, say, y_{ijl}, then the model is written,

$$\text{PROB}(x_{ijl}=1.0)=\sum_k g_{ik}\cdot\lambda_k . \qquad (1)$$

This model has well defined characteristics because once J variables are selected, the "possible" response space M is defined with Sj Lj basis vectors. The basis vectors and the constraints on the g_{ik}s and λ_{kjl}s used to estimate the probabilities in (1) define a linear space L_B. The intersection of LB and M (LB∩M) defines another lower dimensional convex polytope.B, which we can call the "actualized" possibility space (defined by the totality of traits of individuals which can be characterized, either by its vertices (the coordinates λ_{kjl}), or its faces (the Σ_j L_j half spaces defined by $\lambda_{kjl} \geq 0$; (Woodbury et al., 1993; Weyl, 1949).

The GoM model can be further generalized by removing the restrictions that B and its dual space B* be equal in dimension. Thus, in place of (1) we write

$$\text{PROB}(x_{ijl}=1.0)=\sum_k\left(\sum_r \gamma_{ir}\varphi_{rk}\right)\lambda_{kjl} \qquad (2)$$

In (2) the number of vertices of B remains fixed at K, but we "factor" the individual case space g_{ik} into R dimensions, or "groups," of cases. That is, there may be multiple groups of cases characterized by the same number, K, of extreme probability vectors. Alternatively, if R ≠ K then the question is, if (1) is identified because LB«M = B is uniquely defined, is (2) identified? This is satisfied for R < K, i.e., the giks can be viewed as "new" variables where the corresponding possible response space M* is K dimensional so that the "internal" factorization is unique.

The matrix φ_{rk}, when R≠K, is the identity matrix I,. When R ≠ K, then φ_{rk} is the pseudo-inverse of a non-negative matrix ϕ_{kr}. It appears then that φ_{rk} may have negative elements, i.e., φ is the probability matrix relating R and K. If, however, we are analyzing episodes distributed longitudinally, then there are additional dimensions of data (replications over time) that may be exploited to identify parameters (assuming constancy of process parameters). If we designate episodes by t, then we can rewrite (2) as

$$p_{jl}^{i\,t}=\sum_k\sum_m g_{im}\,C_k^{mt}\,\lambda_{jl} . \qquad (3)$$

If we substitute, in (3), for g_{im}, the K component fuzzy set decomposition,

$$g_{im}=\sum_r \gamma_r^i\ \varphi_l^l \qquad (4)$$

we get, with the solution constrained over t,

$$p_{jl}^{i\,t}=\sum_k\sum_m\left(\sum_r \gamma_r^i\ \varphi_m^r\right)C_k^{mt}\,\lambda_{jl}^k$$

$$=\sum_k\sum_r \gamma_r^i\left(\sum_m \varphi_m^r\,C_k^{mt}\right)\lambda_j^k$$

$$=\sum_k\sum_r \gamma_r^i\ V_k^{r\,t}\,\lambda_{jl}^k \qquad (5)$$

In (5), $V_k^{r\,t}$, is the matrix relating the R groups of cases to the K probability profiles defining the unit simplex. We should not expect $V_k^{r\,t}$ to be an identity matrix if it is constant over time t. In this case, it would be the probability matrix ϕ, independent of time.

ANALYSIS

Data

The model was applied to data on functional disability (9 measures of ADL, 7 measures of IADL, and 10 measures of physical impairment) and health (29 medical conditions) from the 1982, 1984, and 1989 National Long Term Care Survey. This survey is based on a list sample of persons aged 65 and over who were Medicare eligible. The surveys are designed to be both longitudinally (i.e., persons surviving to the next survey date are reassessed) and cross-sectionally (by adding in a new sample of 5,000 persons who pass age 65 between survey dates) representative of the U.S. elderly population at the survey time. In the analysis below we examine 27 functional measures with 16,585 observations in the three surveys (9,468 persons; some measured multiple times). In addition to the three surveys, records are linked to Medicare files on service use and mortality from 1982 to 1991. Approximately 11,000 deaths are recorded for both detailed responders and for persons identified as nondisabled on the screen (i.e., the total sample size is 45,000 over all three surveys including nondisabled persons and age-in samples).

Results

The model was applied to the NLTCS for the years 1982, 1984 and 1989. From (5)

$$p_{jl}^{i\,t}=\sum_k\sum_r \gamma_r^i\ V_k^{r\,t}\lambda_{jl}^k\ . \qquad (6)$$

In (6) all parameters are assumed nonnegative with the identifying constraints.

$$\sum_r \gamma_r^i = 1 \qquad (7a)$$

$$\sum_k V_k^{r\,t} = 1 \qquad (7b)$$

$$\sum_l \lambda_{jl}^k = 1. \qquad (7c)$$

The vector $\underset{\sim}{\gamma}^i$ is 1-> 1 with the temporal trajectory for individual, IDi. We can pair $\underset{\sim}{\gamma}^i$ with

$$\sum_k V_k^{r\,t}\lambda_{jl}^k = \Lambda_{jl}^{r\,t} \qquad (8)$$

which orients all individual trajectories to a convex common variable space. Thus, coefficients combine the profiles, $\underset{\sim}{\lambda}^k$, to present a time specific profile, $\underset{\sim}{\Lambda}_{jl}^t$, with r implicit

In Table 1 we present $V_k^{r\,t}$ where K was set at 6. In this table the six groups defined by health and functioning characteristics represent the columns. The six "groups" represent sets of individuals with common patterns or each of the three survey dates.

The six profiles may be briefly defined as, 1.) Mobility and IADL impaired; 2.) Elderly, Moderately Impaired; 3.) Elderly, Heavily Impaired but without dementia; 4.) Relatively Functional; 5.) Frail with Dementia; and 6.) by examining the λ_{kjl} from (6) Missing (principaly not in the sample or dead).

We do not present those tables because the novel element we are reporting on is present in Table 1 and because of space considerations.

An examination of Table 1 shows that profile 6 (missing) never occurs in group 1, occurs for group 2 in 1982 and 1989, in 1989 for group 3 and 4, etc. The frail with dementia loads only on the last group in 1989. Group 1 and 4 are functional in 1982 and 1984. Thus, the matrix tells us how subgroups change in prevalence over time.

DISCUSSION

The analysis of V indicated patterns of expression of specific groups over time. This is similar to a two-way classification of cases and variables except that the solution space has well defined properties due to the convexity constraints. More detailed analysis would include dependent mortality processes using the information on the 11,000 deaths. Also, examinations of the identity of groups, and the measures of the extreme profiles, would be normally done in a substantive analysis.

REFERENCES

Bishop, Y.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass. and London, England.

Everitt, B.S. 1984. *An Introduction to Latent Variable Models.* Chapman and Hall, London and New York.

Lazarsfeld, P.F., and N.W. Henry. 1968. *Latent Structure Analysis.* Houghton Mifflin, Boston.

Manton, K.G., M.A. Woodbury, and H.D. Tolley. 1992. *Statistical Procedures for the Application of Fuzzy Set Models to High Dimensional Discrete Response Data.* John Wiley, New York.

Weyl, H. 1949. The elementary theory of convex polyhedra. *Annals of Mathematics Study* 24:3-18.

Woodbury, M.A., K.G. Manton, and H.D. Tolley. 1993. Empirical comparisons of discrete and continuous mixture for classification of health behavior and health service use. In review at *Journal of the American Statistical Association.*

Table 1: Distribution of γ^j vectors

	Type					
	1	2	3	4	5	6
	Mobility	Elderly	Elderly		Frail	
	ADL	Moderately	Heavily	Relatively	With	
	Impaired	Impaired	Impaired	Functional	Dementia	Missing
Group 1						
1982	6.46	0.00	9.43	84.11	0.00	0.00
1984	6.53	0.00	6.59	86.88	0.00	0.00
1989	0.15	95.67	4.18	0.00	0.00	0.00
Group 2						
1982	0.00	0.00	0.00	0.00	0.00	100.00
1984	1.01	0.51	27.63	70.84	0.00	0.00
1989	0.00	0.00	0.00	0.00	0.00	100.00
Group 3						
1982	0.00	0.00	99.16	0.84	0.00	0.00
1984	0.00	0.00	100.00	0.00	0.00	0.00
1989	0.00	0.00	0.00	0.00	0.00	100.00
Group 4						
1982	0.14	0.00	0.00	99.86	0.00	0.00
1984	0.34	0.00	0.00	99.66	0.00	0.00
1989	0.00	0.00	0.00	0.00	0.00	100.00
Group 5						
1982	0.00	0.00	48.09	51.91	0.00	0.00
1984	0.00	0.00	0.00	0.00	0.00	100.00
1989	0.00	0.00	0.00	0.00	0.00	100.00
Group 6						
1982	0.00	0.00	0.00	0.00	0.00	100.00
1984	0.00	0.00	0.00	0.00	0.00	100.00
1989	0.00	90.72	0.00	8.11	1.17	0.00