Ann. Inst. Statist. Math.
Vol. 44, No. 1, 85–95 (1992)

LARGE SAMPLE PROPERTIES OF ESTIMATES OF A DISCRETE GRADE OF MEMBERSHIP MODEL

H. DENNIS TOLLEY[1] AND KENNETH G. MANTON[2]

[1] *Department of Statistics, Brigham Young University, Provo, UT 84602, U.S.A.*
[2] *Center for Demographic Studies, Duke University, Durham, NC 27706, U.S.A.*

(Received December 28, 1989; revised June 11, 1990)

Abstract. Increasingly, fuzzy partitions are being used in multivariate classification problems as an alternative to the crisp classification procedures commonly used. One such fuzzy partition, the grade of membership model, partitions individuals into fuzzy sets using multivariate categorical data. Although the statistical methods used to estimate fuzzy membership for this model are based on maximum likelihood methods, large sample properties of the estimation procedure are problematic for two reasons. First, the number of incidental parameters increases with the size of the sample. Second, estimated parameters fall on the boundary of the parameter space with non-zero probability. This paper examines the consistency of the likelihood approach when estimating the components of a particular probability model that gives rise to a fuzzy partition. The results of the consistency proof are used to determine the large sample distribution of the estimates. Common methods of classifying individuals based on multivariate observations attempt to place each individual into crisply defined sets. The fuzzy partition allows for individual to individual heterogeneity, beyond simply errors in measurement, by defining a set of pure type characteristics and determining each individual's distance from these pure types. Both the profiles of the pure types and the heterogeneity of the individuals must be estimated from data. These estimates empirically define the fuzzy partition. In the current paper, this data is assumed to be categorical data. Because of the large number of parameters to be estimated and the limitations of categorical data, one may be concerned about whether or not the fuzzy partition can be estimated consistently. This paper shows that if heterogeneity is measured with respect to a fixed number of moments of the grade of membership scores of each individual, the estimated fuzzy partition is consistent.

Key words and phrases: Consistency, fuzzy partition, grade of membership.

1. Introduction

This paper examines the consistency of maximum likelihood estimates of a discrete grade of membership (GOM) model. The GOM model has been proposed by Woodbury and Clive (1974) and Woodbury *et al.* (1978) as a method for modeling heterogeneity in high dimensional discrete multivariate data. Individual heterogeneity is accounted for by estimating a set of grade of membership scores for each individual. These grade of membership scores reflect each individual's heterogeneity by characterizing the individual relative to a set of "pure type" conditions. Since the GOM scores can be any value in a bounded range of values no individual is required to be in a cluster as in discriminant analysis. This flexibility has proven extremely useful as regards modeling the high dimensional discrete multivariate data associated with a wide range of studies as illustrated by Woodbury and Manton (1982), Clive *et al.* (1983), Manton *et al.* (1985, 1987), Vertrees and Manton (1986), Berkman *et al.* (1989) and Blazer *et al.* (1989).

Though proving useful in a number of empirical contexts, there has not previously been a formal demonstration of the statistical properties of the GOM parameter estimates. Although the procedure is based on maximum likelihood methods, determining its large sample statistical properties involves resolving the issues posed by the properties of the model. First, the number of parameters estimated increases with sample size. Explicitly each individual has an individual specific set of GOM scores that must be estimated. Hence, one can not expect such properties as consistency or asymptotic normality of parameter estimates without some kind of restrictions as noted by Neyman and Scott (1948). Second, the GOM scores are bounded within a simplex so that many estimates frequently take values on the boundaries. This can cause difficulties in the usual arguments for maximum likelihood estimates which assume the existence of a derivative in an open neighborhood of the true parameters (see Lehman (1983)). In this paper, we show that by suitably selecting a metric, and generalizing the results of Kiefer and Wolfowitz (1956), asymptotically consistent estimates of the pure type parameters can be obtained. From this result, the large sample distribution of the estimated pure type parameters is derived following the arguments of Moran (1971) and Chant (1974).

The GOM model differs from classical discrete multivariate procedures in that the model can be considered as generating a fuzzy partition. For example cluster analysis and discriminant analysis are based upon traditional crisply defined sets. The GOM model assumes that each individual is a "fuzzy" member of any set; the membership being determined by the grade of membership scores of the individual. The theory of fuzzy sets has proven to be a good foundation upon which to model uncertainty. Examples of such models of uncertainty are given by Klir and Folger (1988). However, unlike many applications of fuzzy sets where uncertainty is an outcome of either stochasticity in the data, incomplete information or linguistic system, or "chaotic" determinism, the GOM model assumes that the population of individuals is intrinsically fuzzy. That is, the individuals are very heterogeneous with this heterogeneity being the result of a continuously weighted convex combination of a few pure type characteristics. Thus, the methods presented here can be used to empirically determine grade of membership scores for certain fuzzy partitions using discrete data.

2. Notation and definitions

The following notation will be used to define the model:

J = number of categorical variables observed on each individual
L_j = number of levels or categories of the j-th categorical variable ($L_j \geq 2$ for all j)
X_{ijl} = binary random variable taking the value 0 if the i-th individual did not have response level l for the j-th categorical variable and 1 otherwise
n = number of individuals in the sample $= \sum_i \sum_l X_{i1l}$
p_{ijl} = probability that the i-th individual will make the l-th level response to the j-th categorical variable
λ_{kjl} = probability that an individual of pure type k will make response l to the j-th categorical variable
K = number of pure types (see below)
r $= K\sum_{j=1}^{J}(L_j - 1).$

When J is large, few, if any, of the individuals are likely to have all of the characteristics represented by a pure type. Most will have characteristics (i.e., response probabilities) which are a combination of the various pure type probabilities. Explicitly, we define the grade of membership model as parameterizing the probabilities p_{ijl} as

$$p_{ijl} = \sum_{k=1}^{K} g_{ik}\lambda_{kjl}, \tag{2.1}$$

where $\lambda_{kjl} \geq 0$, $g_{ik} \geq 0$, $\sum_{l=1}^{L_j} = 1$ and $\sum_{k=1}^{K} g_{ik} = 1$. In this model K is assumed known and represents the number of "pure type" characteristics. It is assumed that (2.1) is unique in g and λ using, for example a singlar value decomposition representation.

The set of unknown parameters g_{ik} are the grade of membership (GOM) scores or mixing coefficients. The g_{ik} may be interpreted as the proportion of pure type k present in individual i. The g_{ik} parameters should not be confused with the mixing proportions of a discrete mixture model. In the mixture model the mixing proportions represent the a priori probabilities that the random variable is sampled from the population with the associated probability mass function.

In the current model, the g_{ik} values represent the degree of membership of the i-th individual to the k-th pure type. As pointed out by Singer (1989), these are not probabilities associated with the sampling scheme or the response categories, even though $\sum_k g_{ik} = 1$ still holds. This will be illustrated below by contrasting the structure of the GOM and discrete mixture likelihoods for multinomial variables.

Conditional upon the values of g_{ik}, we assume that an individual's responses are independent of all other individual's response. Additionally, we assume that, again conditional on g_{ik}, the responses of an individual for all categorical variables are jointly independent, *within* an individual. Thus the X_{ijl} random variables have a product multinomial distribution. The likelihood, conditional on the values g_{ik} may be written

$$L = \prod_{i=1}^{n} h_i(\boldsymbol{X}_i, \lambda, \boldsymbol{g}_i), \tag{2.2}$$

where $h_i(\boldsymbol{X}_i, \lambda, \boldsymbol{g}_i) = \prod_{j=1}^{J} \prod_{l=1}^{L_j} p_{ijl}^{X_{ikl}}$, and where p_{ijl} satisfies (2.1). The vector λ is of length r and contains the parameters λ_{kjl}. The vectors $\boldsymbol{X}_i$ and $\boldsymbol{g}_i$ contain the variables X_{ijl} and g_{ik}, respectively, for all values of j and l and for a fixed value of i. By definition, the probability mass function h_i assigns non-zero mass to only a finite number of values of $\boldsymbol{x}$. Thus h_i is bounded for all values of λ and $\boldsymbol{g}_i$.

In the fully parameterized form (2.2) may be written as

$$L = \prod_i \prod_j \prod_l \Big(\sum_k g_{ik}\lambda_{kjl}\Big)^{X_{ijl}}. \tag{2.3}$$

In contrast, the multinomial form of the discrete mixture model, with classification probability P_k is

$$L = \prod_i \Big(\sum_k P_k \prod_j \prod_l \lambda_{kjl}^{X_{ijl}}\Big). \tag{2.4}$$

We see that the structure of the two models, for the same data, are mathematically distinct with the response assumed independent only within the K discrete mixtures.

We assume that for individual i the k-th grade of membership score, the g_{ik} value, is the k-th component of the realization of the random vector $\xi_i = (\xi_{i1}, \ldots, \xi_{iK})$, $i = 1, 2, \ldots$, where ξ_i are jointly independent and identically distributed. In this paper consistency is with respect to the distribution of these random variables. In this regard we will need the following notation and definitions.

1. Ω is the parameter space of values λ. Ω is the direct product of the L_j-simplexes for $j = 1, 2, \ldots, J$ which contains its boundaries.

2. Γ is the space of all K-dimensional distribution functions of $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)$ such that for every distribution function $G \in \Gamma$,

i. $\sum_{k=1}^{K} \xi_k = 1$, a.s. dG.

ii. $\xi_i \geq 0$, $i = 1, \ldots, K$, a.s. dG.

iii. G has at most a countable number of points with probability greater than zero.

3. γ denotes the pair $(\boldsymbol{\lambda}, G)$ where $\gamma \in \Omega \times \Gamma$. The parameter $\gamma_0 \in \Omega \times \Gamma$ will be assumed to be the true parameter in the sense that $\boldsymbol{\lambda}_0$ obtains and that G_0 is the true distribution of $\boldsymbol{\xi}$, of which the g_{ik}'s are realizations.

4. For fixed integer R, $\delta_R(\cdot,\cdot)$ is defined on $\Omega \times \Gamma$ as $\delta_R(\gamma_1,\gamma_2) = \sum_{i=1}^{r} |\arctan \lambda_1^{(i)} - \arctan \lambda_2^{(i)}| + \sum |\mu^{(1)}_{a_1 \cdots a_K} - \mu^{(2)}_{a_1 \cdots a_K}|$, where the second sum is taken over all K-tuples of non-negative integers $(a_1, \ldots, a_K)$ such that $\sum_{k=1}^{K} a_K \leq R$, and where $\mu^{(i)}_{a_1 \cdots a_K}$ is the raw moment defined as $\mu^{(i)}_{a_1 \cdots a_K} = E_{G_i} \xi_1^{a_2} \xi_2^{a_2} \cdots \xi_K^{a_K}$, $i = 1$ or 2.

5. The marginal probability mass function is defined as

$$h(\boldsymbol{x} \mid \gamma) = \int h(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{y}) dG(\boldsymbol{y}).$$

By construction the function $h(\boldsymbol{x} \mid \gamma)$ does not depend on i.

6. Define $w(\boldsymbol{x} \mid \gamma, \boldsymbol{\rho}) = \sup h(\boldsymbol{x} \mid \gamma')$, where the supremum is taken over all values $\gamma' \in \Omega \times \Gamma$ for which $\delta_R(\gamma, \gamma') > r$.

7. For fixed ρ and fixed R define the set $A(\rho)$ as $A(\rho) = \{\text{all } \gamma \text{ such that } \gamma \in \Omega \times \Gamma \text{ and } \delta_R(\gamma, \gamma_0) > \rho\}$.

8. $\bar{\Gamma}$ is the completion of Γ under

$$d_2(G_1, G_2) = \sum |\mu^{(1)}_{a_1 \cdots a_K} - \mu^{(2)}_{a_1 \cdots a_K}|.$$

That is $\bar{\Gamma}$ contains the limits of all Cauchy sequences as measured using $d_2(\cdot,\cdot)$.

Definition 1. $\hat{\gamma}_n$ is called the maximum likelihood estimate of γ if $\hat{\boldsymbol{\lambda}}$ and $\hat{\mu}_{a_1 \cdots a_k}$ are chosen such that the likelihood given in equation (2.2) is maximized subject to the conditions of equation (2.1), and that $\sum_{k=1}^{K} a_k \leq R$.

The maximum likelihood estimate $\hat{\gamma}_n$ approximates the underlying distrubution function G only up to R-th order raw moments. As one would expect, this means that consistency of the maximum likelihood estimate is defined only in the sense of $\delta_R(\cdot,\cdot)$. The index n in $\hat{\gamma}_n$ refers to the number of individuals.

Definition 2. Condition 1 holds for $C \subset \Omega \times \Gamma$ if $\gamma \in C$ implies that

$$\mathrm{Prob}_{\gamma_0} \left\{ \frac{h(x \mid \gamma)}{h(x \mid \gamma_0)} = 1 \right\} < 1.$$

3. Preliminary results

The proof of consistency of $\hat{\gamma}_n$ is based on the argument given by Kiefer and Wolfowitz (1956). In order to apply this argument, however, the following results will be needed.

Clearly $\delta_R(\cdot,\cdot)$ is a semimetric on $\Omega \times \Gamma$. This follows since every G in Γ is a distribution function with support equal to the direct product space defined in Definition 1, above, all moments of $G \in \Gamma$ exist. Any two elements of Γ are equal with respect to $d_2(\cdot,\cdot)$ if all raw moments of degree R or less are equal. Hence two elements in Γ may be different in their moments of degree greater than R and still be zero distance apart using the metric $d_2(\cdot,\cdot)$.

Let C be the set of all Cauchy sequences in Γ with respect to $d_2(\cdot,\cdot)$. If $G \in C$ then there exists at least one sequance G_m, $m = 1, 2, \ldots$ such that $G_m \in \Gamma$ and $\lim_{m\to\infty} d_2(G_m, G) = 0$.

The relation $\delta_0(G, G') = \lim_{m\to\infty} d_2(G_m, G'_m)$ defines an equivalence class in C, where G, G' are the limits of the sequences $G_m \in \Gamma$ and $G'_m \in \Gamma$, $m = 1, 2, \ldots$. A typical set in this equivalence class is $E_G \subset C$ where there exists a sequence $G_m \in \Gamma$ such that $\delta_0(G_m, G) = 0$. By considering constant sequences, i.e., $G_m = G_i$ for all m and fixed i, then we see that the equivalence classes in C contain those distributions in G which have the same moments up to order R. Let G be the set of equivalence classes of C defined by $\delta_0(\cdot,\cdot)$. Then from a well known theorem regarding metric spaces (e.g., see Maddox (1970), p. 28) we have

LEMMA 3.1. *$\Omega \times \bar{\Gamma}$ is a complete metric space.*

Note that the metric for $\Omega \times \bar{\Gamma}$ is $\delta_R(\cdot,\cdot)$ defined above where the distribution used in the definition is any element from the relevant equivalence class. Notationally we will say that $\gamma \in \Omega \times \bar{\Gamma}$ if $\lambda \in \Omega$ and $G \in E \in \bar{\Gamma}$.

With a complete metric space, we wish to examine convergence of the marginal probability mass functions $h(\boldsymbol{x} \mid \gamma)$. The major result needed is the following continuity type result.

LEMMA 3.2. *Every sequence of γ_m, $\{\gamma_m : \gamma_m \in \Omega\times\Gamma\}$ such that $\delta_0(\gamma_m, \gamma^*) \to 0$ as $m \to \infty$ implies that $h(\boldsymbol{x} \mid \gamma_m) \to h(\boldsymbol{x} \mid \gamma^*)$, where $\gamma^* \in \Omega \times \bar{\Gamma}$.*

PROOF. Recalling the definition of $h(\boldsymbol{x} \mid \gamma)$ we have

$$\lim_{m\to\infty} h(\boldsymbol{x} \mid \gamma_m) = \lim_{m\to\infty} \int \cdots \int h(\boldsymbol{x}, \boldsymbol{\lambda}^{(m)}, \boldsymbol{\xi}) dG_m(\xi_1, \ldots, \xi_K).$$

By the boundedness of $h(\cdot,\cdot,\cdot)$ over the range of integration we have

$$= \int \cdots \int \lim_{m\to\infty} h(\boldsymbol{x}, \boldsymbol{\lambda}^{(m)}, \boldsymbol{\xi}) dG_m(\xi_1, \ldots, \xi_K).$$

By Lemma 3.1

$$= \int \cdots \int \left(\lim_{m\to\infty} h(\boldsymbol{x}, \lambda^{(m)}, \xi) \right) dG^*(\xi_1, \ldots, \xi_K).$$

From the form of $h(\boldsymbol{x}, \boldsymbol{\lambda}^{(m)}, \boldsymbol{\xi})$ we have

$$\lim_{n\to\infty} h(\boldsymbol{x}, \boldsymbol{\lambda}^{(n)}, \boldsymbol{\xi}) = h(\boldsymbol{x}, \boldsymbol{\lambda}^*, \boldsymbol{\xi}),$$

where $G_m \to G^*$ and $\lambda^{(n)} \to \lambda^*$, using the metric $\delta_R(\cdot,\cdot)$. Setting $\gamma^* = (\lambda^*, G^*) \in \Omega \times \bar{\Gamma}$ the result follows.

4. Identifiability and consistency

Before we can show consistency, we must first show whether or not the distribution can be uniquely identified. The need for identifiability is the major reason for generating the metric space $\bar{\Gamma}$ with the metric $d_2(\cdot,\cdot)$. The marginal probability mass function $h(\boldsymbol{x} \mid \gamma)$ is identifiable provided that $d_R(\gamma_1, \gamma_2) \neq 0$, implies that $h(\boldsymbol{x} \mid \gamma_1) \neq h(\boldsymbol{x} \mid \gamma_2)$ for at least one value of $\boldsymbol{x}$. From the definition of $h(\boldsymbol{x} \mid \gamma)$, we may write

$$h(\boldsymbol{x}_i \mid \gamma) = \int \cdots \int \prod_{j=1}^{J} \prod_{l=1}^{L_j} \Big(\sum_{k=1}^{K} \lambda_{kjl} \xi_{ik} \Big)^{X_{ijl}} dG_i(\xi_1, \ldots, \xi_K).$$

This may be rewritten as

$$(4.1) \qquad \sum_{k_1=1}^{K} \cdots \sum_{k_j=1}^{K} \cdots \sum_{k_J=1}^{K} \prod_{j=1}^{J} \prod_{l=1}^{L_j} \lambda_{k_j jl}^{X_{ijl}} \int \cdots \int \prod_{j=1}^{J} \prod_{l=1}^{L_j} \xi_{ik_j}^{X_{ijl}} dG_i(\xi_1, \ldots, \xi_K).$$

Note that in both of these equations the distribution functions G_i, are identical over values of i. We index them to keep track of the fact that we have different combinations of x_{ijl} possible. For fixed i and j, x_{ijl} has only one non-zero value over values of l, and this value is unity. Therefore, from equation (4.1) we see that the largest moments of ξ_k in specifying $h(\boldsymbol{x} \mid g)$ are of degree J. Put in symbols

$$(4.2) \qquad h(\boldsymbol{x} \mid \gamma) = \sum_{k_1=1}^{K} \cdots \sum_{k_j=1}^{K} \cdots \sum_{k_J=1}^{K} a(k_1, \ldots, k_J) b(k_1, \ldots, k_J),$$

where $a(k_1, k_2, \ldots, k_J)$ is a function of the λ_{kjl} values and $b(k_1, k_2, \ldots, k_J)$ is one of the degree J moments of $(\xi_1 \cdots \xi_K)$. Both are functions of the particular values of $\boldsymbol{x}$. From this result, we will set $R = J$ for the remainder of the paper.

By inspection one can see that if $J = 2$, $L_j = 2$, $j = 1,2$ and $K = 2$, there are four λ values and two moments in the set of equations. However, there are only four different values of $h(\boldsymbol{x} \mid \gamma)$ generated by varying the values of x_{ijl}. Hence the function $h(\boldsymbol{x} \mid \gamma)$ for these conditions is not identifiable. If the value of J is increased to three, the number of parameters is increased to nine (six λ values and three moments). The number of different $h(\boldsymbol{x} \mid \gamma)$ values is eight; again no identifiability. For $J = 4$, however, the number of $h(\boldsymbol{x} \mid \gamma)$ values (and consequently the number of equations generated) is 16 and the number of λ parameters and moments total only 12.

In general, the number of equations formed by $h(\boldsymbol{x} \mid \gamma)$ by varying the entries of $\boldsymbol{x}$ is $\prod_{j=1}^{J} L_j$. The number of λ parameters is $r = K \sum_{j=1}^{J} (L_i - 1)$. The number of moments of the ξ_{ik} of degree J or less is $(J + K - 1)!/J!(K - 1)! - 1$. Hence, one cannot expect the model to be identifiable unless

$$\prod_{j=1}^{J} L_j > K \sum_{j=1}^{J} (L_j - 1) + \frac{(J + K - 1)!}{J!(K - 1)!} - 1.$$

If this inequality is satisfied we will assume that the model is identifiable. Under this assumption, we will now examine consistency of the maximum likelihood estimates.

LEMMA 4.1. *Let Condition* 1 *hold for* $\Omega \times \Gamma \supset C$. *Then for every* $\gamma \in C$

$$\lim_{\rho \to 0} E \log \frac{w(X \mid \gamma, \rho)}{h(X \mid \gamma_0)} < 0$$

where the limit of ρ *is restricted to* $\gamma \in C$.

PROOF. The proof follows Kiefer and Wolfowitz ((1956), pp. 892–893).

LEMMA 4.2. *If Condition* 1 *holds for* $A(\rho)$, *there exists a number* $b(\rho) = b$, $0 < b < 1$ *and an* $N = N(\varepsilon, \rho)$ *such that for* $n > N$,

$$\text{Prob}_{\gamma_0} \left\{ \sup_{\gamma \in A(\rho)} \frac{\prod_{i=1}^{n} h(\boldsymbol{X}_i \mid \gamma)}{\prod_{i=1}^{n} j(\boldsymbol{X}_i \mid \gamma_0)} > b^n \right\} < \varepsilon.$$

PROOF. The proof follows Wolfowitz (1949).

THEOREM 4.1. *Let* $\hat{\gamma}_n$ *denote the maximum likelihood estimates of* γ *with respect to* $\Omega \times \Gamma$ *for the GOM model. Then* $\hat{\gamma}_n$ *is consistent.*

PROOF. (Proof follows Wald's (1949) Theorem 2) From Lemma 4.2, we know that for all $n > N$

$$\text{Prob}_{\gamma_0} \left\{ \frac{\sup_{\gamma \in A(\rho)} \prod_{i=1}^{n} h(\boldsymbol{X}_i \mid \gamma)}{\prod_{i=1}^{n} h(\boldsymbol{X}_i \mid \gamma_0)} < h^n \right\} > 1 - \varepsilon,$$

where $h = h(p)$, $0 < h(p) < 1$. Let E be the event that a sequence of maximum likelihood estimates $\hat{\gamma}_n$ has a limit point γ^* where $\delta(\gamma_0, \gamma^*) > \varepsilon$. This means

$$\sup_{\gamma' \in A(\rho)} \prod_{i=1}^{n} h(\boldsymbol{X}_i \mid \gamma') \geq \prod_{i=1}^{n} h(\boldsymbol{X}_i \mid \hat{\gamma}_n)$$

for infinitely many n. This corresponds to

$$\frac{\sup_{\gamma \in A(\rho)} \prod_{i=1}^{n} h(\boldsymbol{X}_i \mid \gamma)}{\prod_{i=1}^{n} h(\boldsymbol{X}_i \mid \gamma_0)} \geq \frac{\prod_{i=1}^{n} h(\boldsymbol{X}_i \mid \hat{\gamma}_n)}{\prod_{i=1}^{n} h(\boldsymbol{X}_i \mid \gamma_0)} \geq 1$$

for infinitely many n. This last inequality comes from the definition of the maximum likelihood estimate. Thus from Lemma 4.2, this event E has probability zero. Hence the result.

5. Applications

Theorem 4.1 shows that the estimates of λ_{kjl} are consistent as n increases. In addition, following Kiefer and Wolfowitz (1956), a corollary to this theorem can be proved which shows that the estimate of the distribution of ξ is also "consistent." However, "consistency" here means that the estimated first R moments of the distribution of ξ approach the first R moments that define the equivalence classes in $\bar{\Gamma}$. In general, the asymptotic behavior of the estimates of these moments will follow standard results for maximum likelihood estimates. However, the estimates of λ are more complicated since many of the "likelihood" estimates fall on the boundary of the parameter space (i.e., are either 0 or 1). In this section we will examine the large sample properties of the estimates $\hat{\lambda}_{kjl}$.

To examine the large sample properties of the estimates $\hat{\lambda}_{kjl}$ we need the following definitions

$\boldsymbol{\theta}$ = vector of λ values and all raw moments of G_0 of degree $\leq R$
θ_i = i-th element of $\boldsymbol{\theta}$
$\hat{\theta}_i$ = i-th element of the maximum likelihood estimate of $\boldsymbol{\theta}$
I_p = index set of length p
B_I = event that $\hat{\theta}_m$, $m \in I_p$, is set to a boundary value (e.g. zero) in the estimation process
w_m = 0 if $m \notin I_p$
= 1 if $m \in I_p$
$f_i(\boldsymbol{x}, \boldsymbol{\theta})$ = $h(\boldsymbol{x}, \lambda, \boldsymbol{g}_i)$, where g_i satisfy the constraints on the sample moments of degree $\leq R$

$$Y_m(\boldsymbol{\theta}) = n^{-1/2} \sum_{i=1}^{n} \frac{\partial \log}{\partial \theta_m} f_i(\boldsymbol{x}, \boldsymbol{\theta})$$

$$\varphi(I) = (w_1\theta_1, \ldots, w_p\theta_p)^T$$

$$\boldsymbol{Y}(\boldsymbol{\theta}) = (Y_1(\boldsymbol{\theta}), \ldots, Y_p(\boldsymbol{\theta}))^T$$

$$C = \left\{ -\frac{\partial^2 \log f_1(\boldsymbol{x}, \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \right\}$$

$C_{12}(I)$ = matrix formed by striking out the rows of C corresponding to each $m \in I$
$C_{22}(I)$ = matrix formed by striking out the columns of $C_{11}(I)$ corresponding to each $m \in I$
$C_{11}(I)$ = matrix formed by striking out the rows and columns of C corresponding to each $m \notin I$
$Z_1(I)$ = vector formed from the vector $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ by striking out entries corresponding to $m \notin I$
$\boldsymbol{Z}_2(I)$ = vector formed from the vector $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ formed by striking out all entries corresponding to each $m \in I$.

Define vectors $\boldsymbol{Y}_1(I)$ and $\boldsymbol{Y}_2(I)$ from $\boldsymbol{Y}(\boldsymbol{\theta})$ similar to the definition of $\boldsymbol{Z}_1(I)$ and $\boldsymbol{Z}_2(I)$ and $\hat{\boldsymbol{Y}}_1(I)$ and $\hat{\boldsymbol{Y}}_2(I)$ defined similarly from $\boldsymbol{Y}(\hat{\boldsymbol{\theta}})$.

LEMMA 5.1. *Conditional on the event $B(I)$, the vector $n^{-1/2}C_{22}(I)\boldsymbol{Z}_2(I)$ is distributed asymptotically as a normal random vector with mean $C_{12}^T(I)\lambda_2(I)$ and variance $C_{22}(I)$.*

PROOF. (This proof follows the logic presented in Chant (1974)) Since $f(\boldsymbol{x}, \boldsymbol{\theta})$ represents a discrete density function, we have

$$E\boldsymbol{Y}(\boldsymbol{\theta}) = 0, \qquad E\boldsymbol{Y}(\boldsymbol{\theta})\boldsymbol{Y}^T(\boldsymbol{\theta}) = C.$$

Thus, by the central limit theorem $n^{-1/2}\boldsymbol{Y}_1(I)$ is asymptotically distributed as a normal random variabnle with mean 0 and covariance matrix $C_{11}(I)$. We may expand the vector $\boldsymbol{Y}(\hat{\boldsymbol{\theta}})$ as

$$\boldsymbol{Y}(\hat{\theta}) = \boldsymbol{Y}(\theta) - C\boldsymbol{Z} + \boldsymbol{Z}^T R_n \boldsymbol{Z}, \tag{5.1}$$

where from Theorem 4.1, $\boldsymbol{Z}^T\boldsymbol{R}_n\boldsymbol{Z}$ goes to zero in probability (see Lehman (1983), p. 415). Under the event $B(I)$, $\hat{\theta}_m$, $m \in I$, is set to zero. Thus

$$\hat{\boldsymbol{Y}}_1(I) < 0 \quad \text{and} \quad \hat{\boldsymbol{Y}}_2(I) = 0. \tag{5.2}$$

Thus (5.2) implies that asymptotically

$$\begin{aligned} T(I) = \boldsymbol{Y}_1(I) - C_1(I)\boldsymbol{Z}_1(I) - C_{12}(I)\boldsymbol{Z}_2(I) < 0 \quad &\text{and} \\ \boldsymbol{Y}_2(I) - C_{12}^T(I)\boldsymbol{Z}_1(I) - C_{22}(I)\boldsymbol{Z}_2(I) = 0. \end{aligned}$$

Hence,

$$C_{22}(I)\boldsymbol{Z}_2(I) = \boldsymbol{Y}_2(I) - C_{12}^T(I)\boldsymbol{Z}_1(I).$$

As a result, the marginal distribution of $C_{22}(I)\boldsymbol{Z}_2(I)$ is asymptotically a normal with mean $C_{12}^T(I)\lambda_2(I)$ and covariance $C_{22}(I)$, since $\varphi_1(I) = 0$ by definition. Hence the result.

To remove the conditional component on $B(I)$, arguments similar to those given in Self and Liang (1987) are required.

6. Conclusion

In this paper we have demonstrated consistency for the estimates of the structural parameters of a "fuzzy partition" model and for the moments of the distribution of individual "nuisance" parameters. The proof is obtained by defining the metric so that identifiable "packets" of information increase with sample size. The fact that no explicit distribution need be specified for the individual nuisance parameters makes the model extremely flexible in representing high dimensional discrete response data. Given recent rapid increases in computational power this general strategy promises to be an increasingly fruitful area for the development of new methods of modeling large multivariate data sets where heterogeneity is present. The results of this paper provide a statistical justification for modeling this type of data using fuzzy partitions.

REFERENCES

Berkman, L., Singer, B. and Manton, K. G. (1989). Black/white differences in health status and mortality among the elderly, *Demography*, **26**, 661–678.

Blazer, D., Woodbury, M. A., Hughes, D., George, L. K., Manton, K. G., Bachar, J. R. and Fowler, N. (1989). A statistical analysis of the classification of depression in a mixed community and clinical sample, *Journal of Affective Disorders*, **16**, 11–20.

Chant, D. (1974). On asymptotic tests of composite hypotheses in nonstandard conditions, *Biometrika*, **61**, 291–298.

Clive, J., Woodbury, M. A. and Siegler, I. C. (1983). Fuzzy and crisp set-theoretic-based classification of health and disease, *Journal of Medical Systems*, **7**, 317–331.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Ann. Math. Statist.*, **27**, 887–906.

Klir, G. J. and Folger, T. A. (1988). *Fuzzy Sets, Uncertainty, and Information*, Prentice Hall, New Jersey.

Lehman, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.

Maddox, I. J. (1970). *Elements of Functional Analysis*, Cambridge University Press, Cambridge, U.K.

Manton, K. G., Liu, K. and Cornelius, E. S. (1985). An analysis of the heterogeneity of U.S. nursing home populations, *Journal of Gerontology*, **40**, 34–46.

Manton, K. G., Stallard, E., Woodbury, M. A. and Yashin, A. I. (1987). Grade of membership techniques for studying complex event history processes with unobserved covariates, *Sociological Methodology*, 309–346, Jossey-Bass, San Francisco, California.

Moran, P. A. P. (1971). Maximum likelihood estimation in non-standard conditions, *Proceedings of the Cambridge Philosophical Society*, **70**, 441–450.

Neyman, J. and Scott, E. L. (1948). Consistent estimators based on partially consistent observations, *Econometrica*, **16**, 1–32.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *J. Amer. Statist. Assoc.*, **82**, 605–610.

Singer, B. (1989). Grade of membership representations: Concepts and problems, *Probability Statistics and Mathematics Papers in Honor of Samuel Karlin* (eds. T. W. Anderson, K. B. Athreya and D. L. Iglehart), Academic Press, New York.

Vertrees, J. and Manton, K. G. (1986). A multivariate approach for classifying hospitals and computing blended payment rates, *Medical Care*, **24**, 283–300.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.*, **20**, 595–601.

Wolfowitz, J. (1949). On Wald's proof of the consistency of the maximum likelihood estimate, *Ann. Math. Statist.*, **20**, 601–602.

Woodbury, M. A. and Clive, J. (1974). Clinical pure types as a fuzzy partition, *Journal of Cybernetics* **4**, 111–121.

Woodbury, M. A. and Manton, K. G. (1982). A new procedure for analysis of medical classification, *Methods of Information in Medicine*, **21**, 210–220.

Woodbury, M. A., Clive, J. and Garson, A., Jr. (1978). Mathematical typology: A grade of membership technique for obtaining disease definition, *Computers and Biomedical Research*, **11**, 277–298.