Paper 192-2007

Latent Class Analysis in SAS®: Promise, Problems, and Programming

David M. Thompson, Dept. of Biostatistics and Epidemiology,
Univ. of Oklahoma Health Sciences Center, Oklahoma City, OK

ABSTRACT

Latent class analysis (LCA) is an important tool for marketing professionals who must characterize subgroups within large and heterogeneous populations. LCA is also of interest to clinical professionals who must place clients in diagnostic or prognostic categories when a gold standard for doing so is poorly defined.

Attempts to bring LCA into the SAS® mainstream are fairly recent. The paper discusses these efforts and demonstrates a SAS macro that combines PROC CATMOD with conventional DATA steps to perform LCA. The macro is demonstrated on data wherein four binary observed variables permit estimation of two hypothesized latent classes.

LCA is a categorical analog to factor analysis, and posits the existence of unobserved classes to explain the pattern of association observed in a multidimensional contingency table. LCA estimates two types of parameters: (1) latent class prevalences and (2) probabilities, conditional on class membership, of individuals' responses on each observed variable. The SAS macro estimates these parameters using a classic expectation-maximization (E-M) algorithm. Maximization steps specify a log-linear model in PROC CATMOD while expectation steps employ standard data step programming.

The presentation illustrates the usefulness of LCA and probes certain problems and limitations associated with constructing and interpreting LC models. LC parameter estimates are sensitive to their initial values, and the classic E-M approach does not estimate standard errors. Bootstrapping of standard errors and replicated analyses using a grid of initial estimates are among the approaches that can address these limitations.

Search keywords: categorical data; diagnosis

INTRODUCTION

Latent class analysis (LCA) is a categorical analog to factor analysis. Factor analysis defines unobserved factors to which to attribute the complex covariance structure of a multivariable sample. Similarly, LCA posits unobserved (latent) classes to explain complex associations in a multi-dimensional contingency table. The set of observed categorical variables are typically called "manifest indicators." LCA estimates two types of population parameters: (1) the prevalence of each latent class, the number of which the analyst must specify *a priori*; (2) the probabilities, conditional on latent class membership, that an individual demonstrates a specific response to an observed variable.

Its ability to detect an unobserved categorical structure makes LCA an important tool for marketing professionals who seek to identify subgroups within large and heterogenous populations. LCA is equally useful for clinical professionals who must place clients in diagnostic or prognostic categories when a gold standard for doing so is poorly defined.

Latent class analysis is unavailable in SAS. Investigators who wish to use SAS to perform latent class analysis must author algorithms in SAS' matrix language, PROC IML, or learn lesser used procedures. IML modules that perform latent class analysis include one by the author (Thompson, 2003) and latent class regression macros developed at the Johns Hopkins School of Public Health (Bandeen-Roche, Miglioretti, Zeger, & Rathouz, 1997). A group at Penn State's Methodology Center has produced a beta version of a module they call PROC LCA (Methodology Center, 2007). Other researchers have applied latent class models to assess diagnostic accuracy by using SAS PROC NLIN, which performs nonlinear regression using weighted least squares estimation (Engels, Sinclair, Biggar, Whitby, & Goedert, et al., 2000; Blick & Hagen, 2002).

This paper illustrates an approach to LCA that relies on IML, and another that attempts to use conventional PROC and DATA steps. The programs currently accommodate information on up to four binary manifest variables to estimate the latent structure of a single unobserved variable with two hypothesized classes.

Stouffer and Toby (1951) published a contingency table containing data on 216 observations that others have used (Goodman, 1974, 2002; McCutcheon, 1987, Table 1.2, p.10) to illustrate latent class analysis. The original research explored undergraduate students' responses to four stories that forced them to make ethical decisions when confronted with conflict between their roles as friends and their roles as members of larger social groups. The four

stimulus stories involve an auto-pedestrian accident to which the student is a witness and in which the faulty driver is a friend; a botched Broadway play that the student must review, and that is authored by a friend; a physical exam that will determine a friend's eligibility for insurance; and a secret meeting that yields information related to the future price of a stock that a friend owns. Respondents' decisions as to whether they would resolve role conflict in favor of a friend present a complex pattern of association and independence (Figure 1).

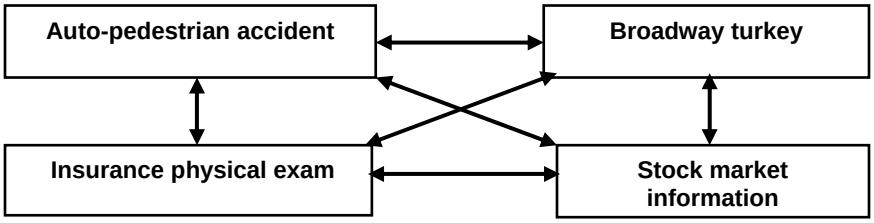

Figure 1. Observed responses to situations involving role conflict a complex pattern of association (after Stouffer & Toby, 1951).

Presented with these data in the form of a multidimensional contingency table, LCA arrives at a classification scheme (Figure 2) that "explains away" (Goodman, 2002, p.4) the complex associations that are observed among the four binary indicators. Once LCA constructs appropriate latent classes, the observed indicators are independent, conditional on a subject's class membership.

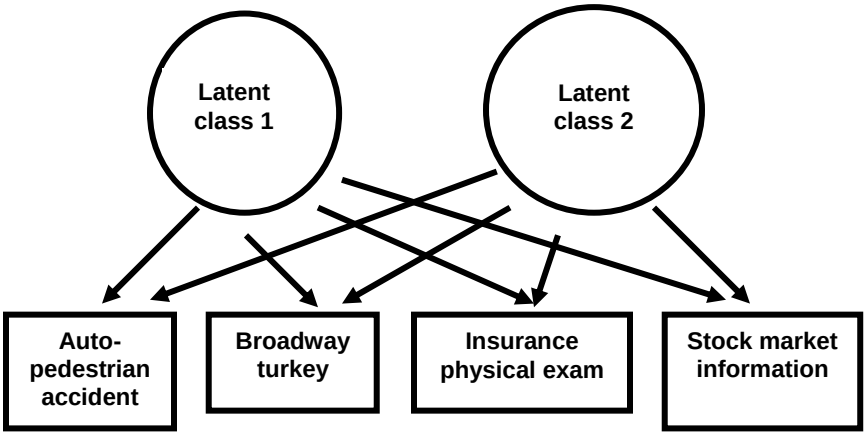

Figure 2. Conditional on latent class membership, observed variables are independent.

The identity or significance of the latent classes is informed by theoretical or substantive knowledge. Stouffer and Toby (1951) hypothesized that students resolved role conflict from either "universalistic" or "particularist" stances.

ASSUMPTIONS OF LATENT CLASS ANALYSIS

Latent class models are built on the assumptions of "exhaustiveness" and "local independence." Exhaustiveness refers to the assumption that every set of responses among the manifest indicators (e.g. A=i, B=j, C=k, D=l) is associated with membership in a latent class. Under this assumption, every subject is assigned provisionally to a category t of the unobserved or latent variable X. Further, the joint probability π^{ABCD} of any response profile (A=i, B=j, C=k, D=l) can be summed across the hypothesized latent classes.

$$\pi^{ABCD} = \Sigma_t \ \pi^{ABCDX}$$

Latent class prevalences, P(X=t), sum to one under the assumption of exhaustiveness.
Local independence is formally expressed in terms of joint probabilities (Goodman, 1974, equation 10, p. 217):

$$P(A=i, B=j, C=k, D=l , X=t) = P(A=i, B=j, C=k, D=l \mid X=t)$$

$$= P(A=i \mid X=t)\ P(B=j \mid X=t)\ P(C=k \mid X=t)\ P(D=l \mid X=t)\ P(X=t)$$

Under this assumption, observed responses are understood to be independent among members of the same latent class. Once latent class is assigned or accounted for, associations formerly observed among indicators are fully explained. In Goodman's original notation:

$$\pi^{ABCDX} = \pi^{A|X}\,\pi^{B|X}\,\pi^{C|X}\,\pi^{D|X}\,\pi^{X}$$

Under these two assumptions, joint probabilities are functions of the two kinds of parameters estimated in latent class models: (1) latent class prevalences and (2) conditional response probabilities. Conditional response probabilities or "conditional rating probabilities" (Uebersax & Grove, 1990, pp. 569-570) are the probabilities that a subject responds in a certain manner to a set of manifest indicators, given the subject's latent class membership.

MODEL FIT STATISTICS

The "goodness" of a latent class model's fit to observed data is conventionally evaluated by Pearson and likelihood ratio (LR) chi-square statistics. These statistics compare observed counts with the counts expected under the assumption of conditional independence. The statistics' associated degrees of freedom are equal to the number of non-redundant observed counts in the relevant contingency table, minus the number of LC parameters that the model estimates.

Akaike and Bayes information criteria, which are based on the likelihood ratio chi-square, and which account for the number of estimated parameters, are also calculable to compare models with differing restrictions on the parameters.

TWO PARAMETERIZATIONS OF THE LATENT CLASS MODEL

Goodman expressed the LC model's parameters directly in terms of conditional and marginal probabilities that relate to latent class membership. An alternative parameterization (Haberman, 1979; Espeland & Handelman, 1989; Heinen, 1996) expresses them in terms of a log-linear model where the manifest indicators (for example A, B, C, and D) are locally independent, given membership in the latent class X.

In place of Goodman's probabilistic parameterization of the joint probabilities, the loglinear parameterization defines log joint probabilities (Heinen, 1996, equation 2.15, p. 51):

$$\ln \pi^{ABCDX} = \lambda + \lambda_i^{A} + \lambda_j^{B} + \lambda_k^{C} + \lambda_l^{D} + \lambda_t^{X} + \lambda_{it}^{AX} + \lambda_{jt}^{BX} + \lambda_{kt}^{CX} + \lambda_{lt}^{DX}$$

$$\text{so } \pi^{ABCDX} = \exp(\lambda + \lambda_i^{A} + \lambda_j^{B} + \lambda_k^{C} + \lambda_l^{D} + \lambda_t^{X} + \lambda_{it}^{AX} + \lambda_{jt}^{BX} + \lambda_{kt}^{CX} + \lambda_{lt}^{DX})$$

Solutions for the loglinear parameters λ can be converted to more easily understood conditional probabilities and latent class prevalences (Heinen, 1996; Thompson, 2006a).

MAXIMUM LIKELIHOOD APPROACH TO LATENT CLASS ANALYSIS

Most statisticians credit Lazarsfeld and Henry (1968) with the origins of latent class analysis and Goodman (1974) with the computational breakthroughs that made it practical. Goodman's maximum likelihood approach (1974, pp. 216-218; Goodman, 2002; see also McCutcheon, 1987, pp. 21-27) remains the standard way to estimating parameters in the latent class model.

The likelihood function is built on the product of the joint probabilities of each observed response profile, given the assignment of the hypothesized latent class (X=t). These joint probabilities are summed across latent classes, so the likelihood function ultimately contains information from the indicators, which are observed, and from latent class membership, which is not (Dayton & Macready, 2002). Likelihood functions can, therefore, be extremely complex.

A way around the difficulty of constructing a likelihood function from partly unobserved information involves estimating the missing information on latent class membership, then maximizing the likelihood for the provisional but complete 'data.' The approach involves alternating steps that first calculate the likelihood function's *expected* value, and then find the parameter values that *maximize* the function. The expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) first calculates the likelihood function's expected value, given observed counts for a complete (though provisional) contingency table, and given provisional values for the latent class model's parameter estimates. Next, it maximizes the likelihood function and, in so doing, updates the parameter estimates.

The E-M cycle must begin with the arbitrary assignment of each response profile (A=i, B=j, C=k, D=l) to a specific latent class X. This random assignment, which "fills in" missing information on latent class membership, constitutes the first *expectation step*. Goodman's seminal (1974) approach, which incorporates the EM algorithm, updates its parameter estimates in proportion to the observed marginal probabilities for each response profile. The algorithm repeats until expected and observed values converge (Uebersax & Grove, 1990, equations 3 and 4, p. 569; Goodman, 2002, p. 29) and the parameter estimates stabilize.

The E-M algorithm has a significant limitation in that it does not produce standard errors for the parameter estimates, so one cannot test hypotheses or estimate confidence intervals that relate to the parameters. Techniques proposed to overcome this limitation (Goodman, 1972; Louis, 1982) are complex and obtain estimates for the entire covariance matrix only for latent class models with certain kinds of manifest indicators.

SAS APPROACHES TO LCA USING THE E-M ALGORITHM

The author has explored two strategies to incorporate the EM algorithm into SAS approaches to LCA. This paper illustrates results from an algorithm that operationalizes the E-M algorithm in SAS-IML (Thompson 2003), and that reproduces Goodman's (1974) probabilistic parameterization of the latent class model. Another approach, details of which are available online (Thompson 2006c), employs conventional PROC and DATA steps to incorporate the E-M algorithm using SAS PROC CATMOD's loglinear modeling capabilities.

In their current forms, the approaches can employ information on four binary manifest variables (A,B,C, D) to estimate the structure of latent class model with a single latent variable (X) with two hypothesized classes (X=1,2). The approaches begin with raw data in the form of an unconditional contingency table that contains no information on latent class membership. The data from Stouffer and Toby (1951) are displayed below. They involve yes-no responses to four situations (A- Auto-pedestrian accident; B - Broadway turkey; C- Insurance physical exam; D - Stock market Information). The data take the form of sixteen response profiles with observed counts for each profile.

a	b	c	d	count
0	0	0	0	20
0	0	0	1	2
0	0	1	0	9
0	0	1	1	2
0	1	0	0	6
0	1	0	1	1
0	1	1	0	4
0	1	1	1	1
1	0	0	0	38
1	0	0	1	7
1	0	1	0	24
1	0	1	1	6
1	1	0	0	25
1	1	0	1	6
1	1	1	0	23
1	1	1	1	42

An initial step "augments" the unconditional contingency table and creates a complete (conditional) one by adding a column for latent class assignment (X=t). The step assigns each response profile, and its entire count of observations, to a specific latent class X. Assignment of profiles to latent classes is random, so different assignments constitute different initial estimates of latent class prevalences and conditional probabilities.

In simulation studies (Thompson, 2003), the IML approach has been superior in accuracy and speed, and this paper's results derive from the IML approach. Still, a strategy that takes advantage of PROC CATMOD's loglinear modeling capabilities (Thompson 2006a) promises flexibility for model building. The latter approach estimates loglinear LC model parameters in a series of "handoffs" between PROC steps that maximize the likelihood function and intervening DATA steps that convert the loglinear parameters to probabilistic ones. The DATA step also updates expected values and the likelihood function that is based on them.

In the algorithm's **maximization step**, PROC CATMOD produces updated parameter estimates by maximizing a likelihood function that is based on a loglinear model that features conditional independence.

```
ods output
  anova=mlr MaxLikelihood=iters estimates=mu covb=covb;
proc catmod data=two order=data;
  weight count;
  model a*b*c*d*x = _response_ / wls covb addcell=.1;
  loglin a b c d x a*x b*x c*x d*x;
run;
quit;
```

The PROC step's provisional estimates of loglinear LC model parameters, obtained above in the output datset MU, pass to an **expectation** step. This is a DATA step that transforms the loglinear parameters into probabilities, then uses the probabilities to update expected values for joint probabilities π^{ABCDX} and posterior probabilities for latent class membership $\pi^{X|ABCD}$.

PROC CATMOD can estimate loglinear model parameters using weighted least squares (WLS, as illustrated above), iterative proportional fitting, or Newton-Raphson estimation. PROC GENMOD is also potentially appropriate for this purpose if its options are set to estimate a loglinear model using Fisher scoring, in place of or in combination with Newton-Raphson estimation.

THE E-M ALGORITHM AND STANDARD ERRORS

The E-M algorithm has a significant limitation in that it does not produce standard errors for its parameter estimates. Both the IML-based approach featured here and the PROC CATMOD approach open the door to strategies that address this limitation. PROC CATMOD produces a covariance matrix V(λ) of its loglinear parameter estimates, which the sample code above preserves in the output dataset COVB. These loglinear covariance parameters are, in principle, convertible into probabilistic covariance parameter estimates. However, probabilistic parameters are nonlinear functions of their loglinear counterparts. For example, a latent class prevalence π_x is a nonlinear function:

$$\pi_x = \exp \lambda_t^X / \Sigma_x \exp \lambda_t^X$$

The complex calculations involved in producing probabilistic covariance estimates motivated the search for more practical approaches. Two alternatives involve (1) generating bootstrapped standard errors from an initial latent class solution and (2) generating multiple solutions by repetitively "seeding" the EM algorithm with randomly chosen initial estimates. Both approaches calculate multiple sets of LC parameter estimates, producing distributions for the estimates whose means and standard deviations are easily calculated. The distributions' standard deviations constitute standard errors for the respective parameter estimate.

BOOTSTRAPPED STANDARD ERRORS

One straightforward approach is to produce a single LCA solution, to use its probabilistic parameter estimates to find posterior probabilities, and to use these to construct a single complete (conditional) contingency table. From this conditional table, efficient SAS code (Barker 2005) can construct numerous bootstrapped tables. Finally, we can subject each bootstrapped table to its own LC analysis, and collect the parameter estimates from each analysis. The estimates' means and standard deviations constitute, respectively, bootstrapped parameter estimates and standard errors.

The output below, produced in a final PROC REPORT step, summarizes solutions performed using the IML approach on 100 conditional contingency tables bootstrapped from a single initial solution.

		Latent Class X=t			
		X=1		X=2	
	LC Parameter	Mean	STD	Mean	STD
Conditional Probs	P(A=1|X=t)	0.9752	0.0316	0.7022	0.0556
	P(B=1|X=t)	0.9143	0.0685	0.2939	0.0772
	P(C=1|X=t)	0.8816	0.0825	0.3395	0.0520
	P(D=1|X=t)	0.7315	0.1083	0.1103	0.0411
Latent Class Prevalences		0.3300	0.0870	0.6700	0.0870

Number of solutions	LR Statistic			Number of iterations		
	Mean	Min	Max	Mean	Min	Max
100	10.64	2.47	42.93	31.02	7	91

STANDARD ERRORS OBTAINED THROUGH MULTIPLE SOLUTIONS

Standard errors can also be obtained by performing repeated latent class solutions and then collecting and summarizing the multiple estimates of latent class parameters. The approach's virtue is its explicit recognition of the potential complexity of the likelihood surface, and of latent class models' sensitivity to initial estimates. By "seeding" the EM algorithm's first expectation step with numerous random initial mappings of response profiles to latent classes, we should obtain a distribution of estimates that reflects the likelihood surface's relative maxima. Empirical standard errors drawn from this distribution should, therefore, illuminate the latent likelihood surface's complexity, and alert the investigator to complications in parameter estimation. The distribution of parameter estimates can help the investigator, better than conventional goodness of fit statistics or iteration histories, to judge the adequacy of parameter estimates related to a latent structure.

The output below, produced in a final PROC REPORT step, summarizes solutions performed using the IML approach on 100 conditional contingency tables. Each table was produced from a random initial assignment of response profiles to latent class X=1 or X=2. An intermediate DATA step screens the solutions and preserves only those whose LR statistics are no more than twice the median value for all solutions. This step rejects solutions that converge poorly, on the assumption that they resulted from unreasonable initial latent class assignments.

		Latent Class X=t			
		X=1		X=2	
	LC Parameter	Mean	STD	Mean	STD
Conditional Probs	P(A=1|X=t)	0.9909	0.0018	0.7113	0.0038
	P(B=1|X=t)	0.9330	0.0129	0.3256	0.0060
	P(C=1|X=t)	0.9210	0.0138	0.3499	0.0053
	P(D=1|X=t)	0.7597	0.0114	0.1291	0.0061
Latent Class Prevalences		0.2873	0.0117	0.7127	0.0117

Number of solutions	LR Statistic			Number of iterations		
	Mean	Min	Max	Mean	Min	Max
94	2.84	2.75	4.17	31.29	8	90

"FRACTURED" SAMPLING DISTRIBUTIONS AND THEIR CURE

Both bootstrapping and the production of multiple solutions from different starting points address the EM algorithm's inability to produce standard errors for parameter estimates. However, the approaches introduce another programming challenge. LCA's identification of latent class membership is flexible. LCA begins with an unconditional contingency table and "completes" the table by associating response profiles with categories of a latent variable (X=1,2 for example). The results, for each LCA solution, are *parameter vectors* that consist of estimates of an LC prevalence and a set of conditional probabilities for each postulated latent class.

However, latent classes have no intrinsic meaning or identity. Therefore, no criterion can guide the algorithm during repeated solutions to consistently identify the classes. The algorithm may attribute a vector of estimates (LC prevalence and a set of conditional probabilities) to the latent class X=1 in one solution, then attribute the corresponding vector of estimates to the class X=2 in the next. Consequently, when applied to many bootstrapped or repeated solutions for a specific set of data, LCA produces "fractured" sets of estimates. The estimates' sampling distributions can appear bimodal, as illustrated in figures 3 and 4, which summarize multiple estimates for the conditional probabilities PA=1| X=1 and P(A=1|X=2) for the Stouffer and Toby (1951).

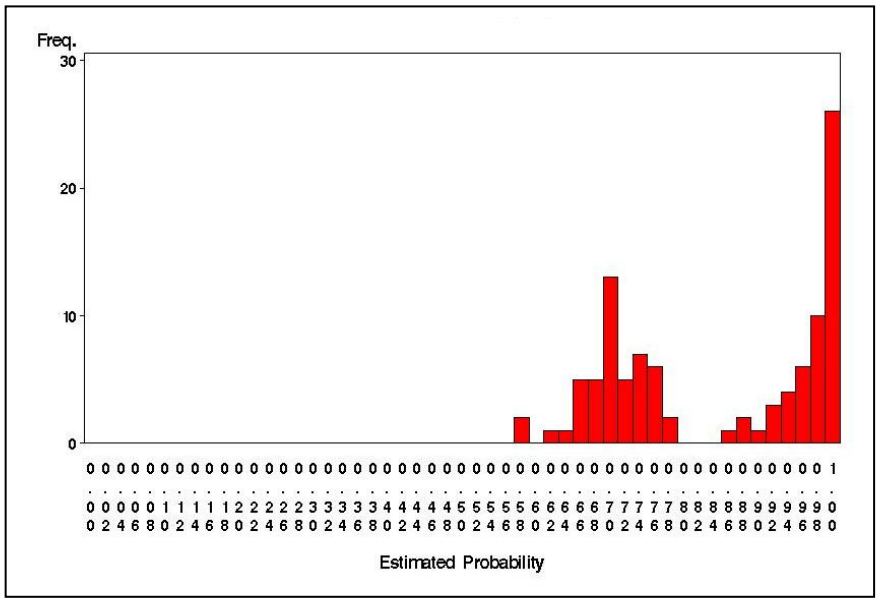

Figure 3. Distribution of multiple estimates of conditional probability P(A=1|X=1) .

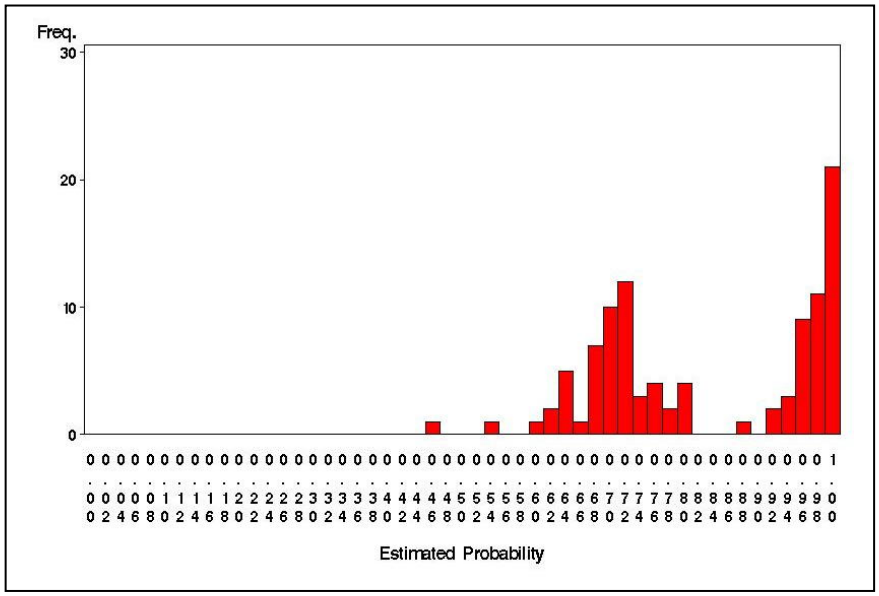

Figure 4. Distribution of multiple estimates of conditional probability P(A=1|X=2)

The distributions' bimodal appearance is an artifact, as explained previously, of the fact that the algorithm attributes a conditional probability to the latent class X=1 for some solutions, but to the latent class X=2 for others. In simulation studies, solutions produced satisfactory goodness of fit statistics within individual replications (Thompson, 2003). That is, within each individual solution, the algorithm calculated a parameter vector for latent class X=1 whose members' values were logical and meaningful in relation to those of the vector it calculated for class X=2.

The verification, through simulation studies, of the coherency of individual solutions motivated the search for a way to consistently assign vectors to the appropriate latent class. The present strategy reflects vectors into the appropriate region or "half-plane" of a parameter space that maps individual estimates of, for example, P(A|X=1) versus P(A|X=2). Figure 5 illustrates how multiple solutions for a conditional probability naturally fracture into two half-planes that are defined by the simultaneous paired estimates for (P(A=1|x=1) and P(A=1|X=2).

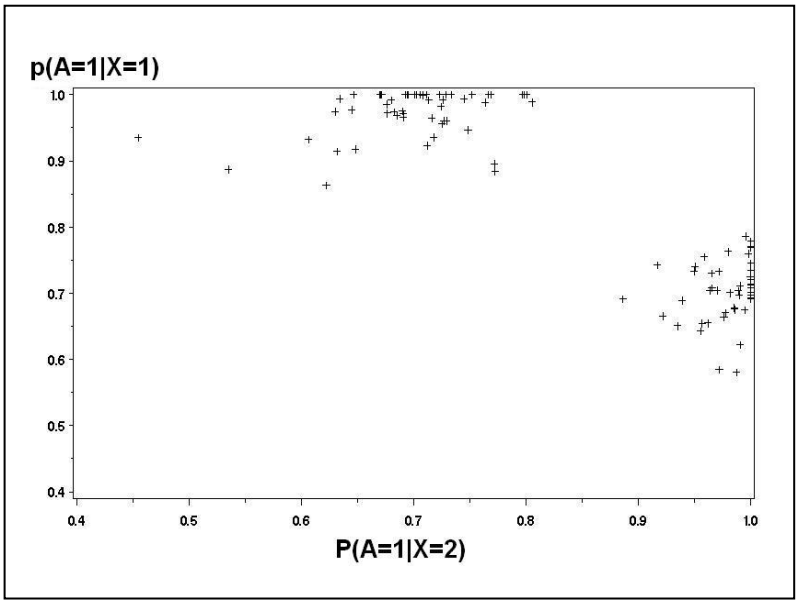

Figure 5. Raw or "unreflected" estimates for P(A=1|X=1) and (PA=1|X=2) from multiple LCA solutions congregate in two half-planes.

The present strategy cures the fracturing of estimates by congregating them in the same "half-plane." To arrive at a decision tool that consistently reflects estimates into the appropriate "half-plane," a DATA step first identifies the most informative indicator for a given latent class solution. The most informative indicator is the one whose conditional probabilities, given latent class membership, differ the most. The DATA step illustrated below screens a single LCA solution (in the SAS dataset PIONE) to determine the indicator that best distinguishes among latent class membership. The DATA step operates on the initial LCA solution when the bootstrapping approach is used. When multiple solutions are calculated, the DATA step operates on the first solution with an acceptable LR statistic.

```
data direction;
  set pione;
  array xx1 [4] pa_x1 pb_x1 pc_x1 pd_x1;
  array xx2 [4] pa_x2 pb_x2 pc_x2 pd_x2;
  array dd  [4];
  do i=1 to 4;
    dd[i]=abs(xx1[i]-xx2[i]);
    crit=max (of dd1-dd4);
   if crit=dd[i] then icrit=i;
   if i=4 then output;
  end;
run;
```

Solutions in this key indicator's less populated half-plane are reflected into its more populated half plane. Once determined for the key indicator, the direction of reflection is applied to each member of the vector of estimates (the conditional probabilities and the LC prevalence) for a given solution. This strategy consistently maps all of the many vectors of estimates to the same latent classes. Figure 6 illustrates the results of this reflection for the pair of conditional probabilities P(A=1|X=1) and (PA=1|X=2).

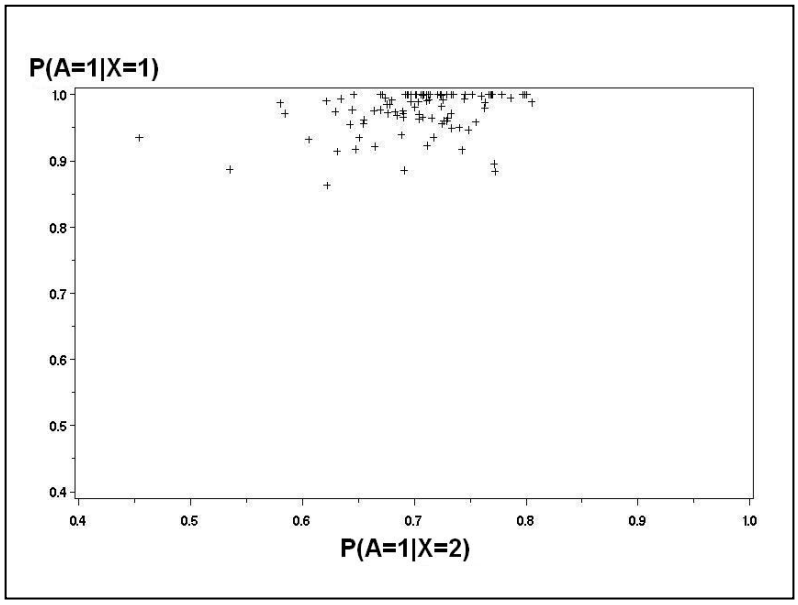

Figure 6. Systematic reflection into a single half-plane cures the "fracturing" of estimates for the conditional probabilities P(A=1|X=1) and (PA=1|X=2).

After reflection (Figures 7 and 8), solutions are consistently categorized so the bimodal sampling distributions resolve to ones whose means and standard errors are appropriate.

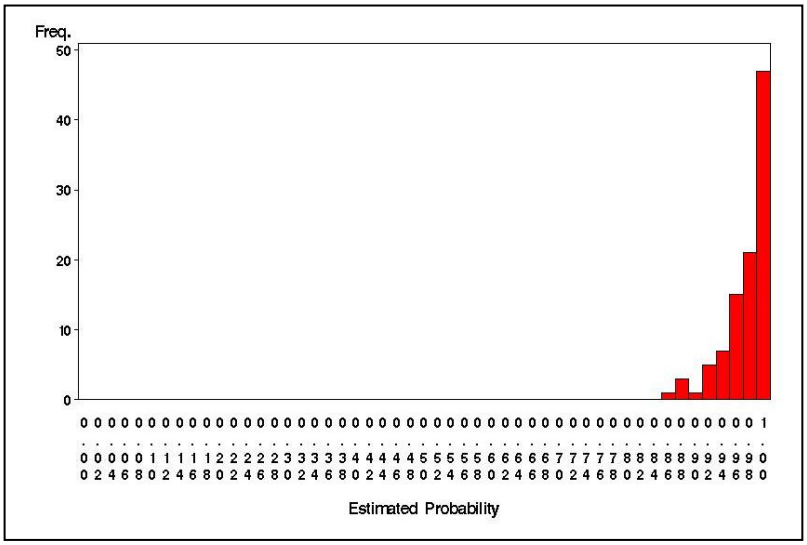

Figure 7. Distribution of multiple estimates, after reflection, for conditional probability P(A=1|X=1)

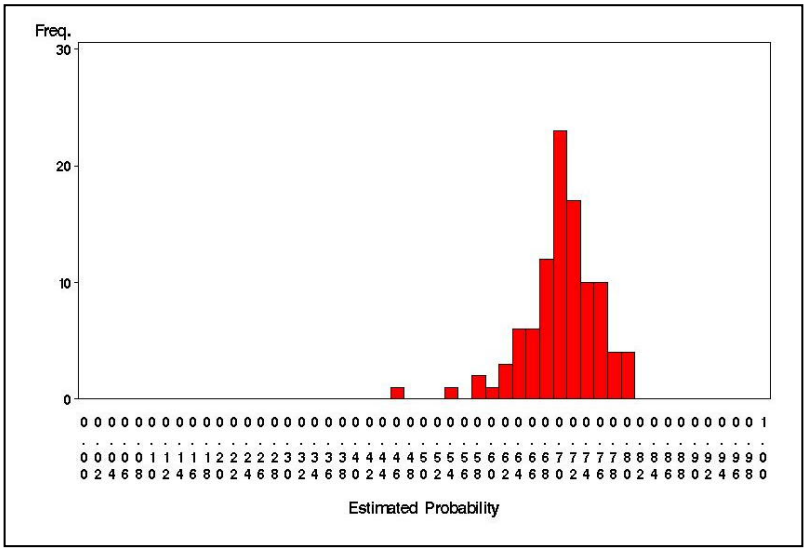

Figure 8. Distribution of multiple estimates, after reflection, for conditional probability P(A=1|X=2)

LIMITATIONS OF LATENT CLASS ANALYSIS

LCA can detect an unobserved structure only if each of its classes is large enough to be discernible. When membership in a latent class is rare, LCA may be unable to distinguish the class' low prevalence from zero. Detecting rare latent classes is increasingly difficult as sample size decreases or the number of latent classes increases.

LCA's ability to detect latent structure is similarly limited when the analyst chooses manifest indicators that are inappropriate or inadequate. LCA cannot arrive at precise parameter estimates when indicators' diagnostic accuracies are low, that is, when their conditional probabilities (e.g. P(A=1|X=t) are close to 0.5 (Albert, McShane, & Shih, 2001, p.618). Individual indicators are most informative when their conditional probabilities differ from one another.

LCA assumes that all indicators are locally independent. However, in practice certain indicators may be correlated for reasons other than latent class membership (Vacek, 1985; Hagenaars, 1988). Responses to questions about symptoms, for example, may correlate more highly among patients with a severe form of a disease than among those without the disease or with milder disease.

The presence of a pair of correlated indicators can lead to overestimation of both indicators' informativeness (Vacek, 1985; Albert, McShane, & Shih, 2001, p. 618). Vacek (1985) suggests further that a positive covariance between responses on two indicators, conditional on membership in one latent class, leads to overestimation of the other class' prevalence. Pepe and Janes (2005, 2006) issue strong warnings about interpreting latent class parameters when conditional independence does not hold.

Conventional goodness of fit statistics, which are based on the assumption of local independence, can warn the analyst about violations of conditional independence. However, these statistics have important limitations. They are asymptotically distributed as χ^2 random variables, but may not dependably follow the distribution when calculated from multi-dimensional contingency tables where sparse data produce small expected cell counts (Formann & Kohlmann, 1996, p. 195).

Moreover, simulation studies (Thompson, 2003) illustrate that the LR statistic remains relatively small even when indicators are relatively uninformative and no strong latent structure exists. The LR statistic, which focuses on conditional independence, may not sufficiently alert the analyst to equally important reasons for lack of fit, as when the model is struggling to characterize a latent structure that is weak or non-existent.

ASSESSING CONDITIONAL DEPENDENCE

The presence of uncharacterized or unappreciated local dependence leads to poor model fit. Confronted with evidence that the model fits poorly, a naïve analyst reasonably attributes the problem to errors in the hypothesized latent structure. One response is to posit additional latent classes. When a substantive reason to add latent classes does not exist, models thus created are, at the very least, more difficult to interpret than the more parsimonious models originally hypothesized. Additionally, these models are likely to capitalize on idiosyncrasies in the sample and, therefore, to have limited applicability to the population from which the sample originated.

Detection of violations of local independence relies on the availability and validity of goodness of fit statistics. Previous discussion revealed that these statistics may not be reliably distributed as χ^2 random variables in small samples. Accordingly, the algorithm described here performs a diagnostic procedure advocated by Garrett and Zeger (2000) for models with binary indicators. It compares observed and model-predicted log odds ratios for marginal tables formed according to pairs of indicators. Because the expected log odds ratio has an asymptotic standard error equal to $(1/a + 1/b + 1/c + 1/d)^{1/2}$, the difference can be standardized, and the resulting z-score used as an informal screening tool. A large residual arouses suspicion that the corresponding pair of indicators is locally dependent. Garrett and Zeger further refine the technique by including a “continuity correction,” adding 0.5 to any zero cell that persists in a two-way frequency table (2000, p. 1062).

The output reproduced below from a PROC REPORT step summarizes the checks for evidence of pairwise residual dependence. None of the reported z-statistics are sufficiently large to arouse suspicion that residual conditional dependence is unaccounted for by the latent class model.

Pairs of indicators with Z>2 may display residual dependence.

Indicators		Log odds			
		Expected	Observed	ASE	z
a	b	0.2993	0.7270	0.3557	1.2024
a	c	0.3630	0.7953	0.3557	1.2154
a	d	0.7847	0.5312	0.4796	-0.5285
b	c	0.6534	0.5586	0.2760	-0.3435
b	d	1.2871	1.3876	0.3430	0.2929
c	d	1.6395	1.6994	0.3685	0.1626

When conditional dependence is detected, a reasonable response is to model it explicitly, and account for it in the estimation of LC parameters. The IML-based algorithm does not contain this extension. However, coexisting with PROC CATMOD's limitations is a strong advantage, its flexibility in entertaining more complex loglinear models. In place of the LOGLIN statement syntax that assumes local independence among all indicators:

```
loglin a b c d x a*x b*x c*x d*x;
```

conditional dependence between, for example, indicators C and D, can be explicitly included by posing:

```
loglin a b c d x a*x b*x c*x d*x c*d c*d*x;
```

PROC CATMOD’s flexibility in loglinear modeling potentially permits estimation of models with more indicators, and of models whose indicators demonstrate some local dependence after positing a latent structure. The primary challenge to these expansions lies in the writing of more flexible DATA steps to calculate the expected values of the probabilistic parameters.

CONCLUSIONS

Its ability to detect unobserved categories makes latent class analysis a promising analog to factor analysis for those seeking to probe complex patterns of association among a collection of categorical indicators. LCA has important limitations. Although strategies exist to obtain standard errors for parameter estimates derived from the E-M algorithm, the analyst must interpret estimates cautiously. The complexities of the likelihood function and likelihood surface make the procedure sensitive to initial estimates. This sensitivity promotes a strategy of obtaining multiple solutions from different starting points, and of monitoring the LR statistics and iteration histories for evidence of problems. Estimates should be interpreted cautiously if evidence exists of residual conditional dependence. Although diagnostic tests can detect evidence of conditional dependence, more flexible programming is necessary to build and test more complex models that account for residual dependence among indicators.

REFERENCES

Albert, P.S., McShane, L.M., & Shih, J.H. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors. *Biometrics, 57*, 610-619.

Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L. & Rathouz, P.J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association, 92,* 1375-1386.

Barker, N. (2005). A practical introduction to the bootstrap using the SAS System. Oxford Pharmaceutical Sciences, Wallingford, UK. Retrieved 2/7/07 from the World Wide Web: www.lexjansen.com/phuse/2005/pk/pk02.pdf

Blick, J., & Hagen, P.T. (2002). The use of agreement measures and latent class models to assess the reliability of classifying thermally marked otoliths. *Fishery Bulletin, 100,* 1-10.

Dayton, C.M., & Macready, G.B. (2002). Use of categorical and continuous covariates in latent class analysis. In J.A. Hagenaars & A.L. McCutcheon (Eds), *Applied Latent Class Analysis* (pp. 213-233). Cambridge, UK: Cambridge University.

Dempster, A.P, Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Engels, E.A., Sinclair, M.D., Biggar, R.J., Whitby, D., Ebbesen, P., Goedert, J.J., & Gastwirth, J.L. (2000). Latent class analysis of human herpesvirus 8 assay performance and infection prevalence in sub-Saharan Africa and Malta. *International Journal of Cancer, 88,* 1003-1008.

Espeland, M.A., & Handelman, S.L. (1989). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics, 45,* 587-599.

Formann, A.K., & Kohlmann, T. (1996). Latent class analysis in medical research. *Statistical Methods in Medical Research, 5*, 179-211.

Goodman, L.A. (1972). A general model for the analysis of surveys. *American Journal of Sociology, 77*, 1035-1086.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 61, 215-231.

Goodman, L.A. (2002). Latent class analysis: The empirical study of latent types, latent variables, and latent structures. In: J.A. Hagenaars & A.L. McCutcheon (Eds.). Applied Latent Class Analysis. Cambridge: Cambridge University.

Haberman, S.J. (1979). *Analysis of qualitative data. Vol. 2: New developments*. New York: Academic Press.

Heinen, T. (1996). Latent class and discrete latent trait models: Similarities and differences. Thousand Oaks, CA: Sage.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*, Boston: Houghton Mifflin.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, B 44,* 226-233.

The Methodology Center, Penn State University. Latent Class Analysis. Proc LCA. Retrieved 2/6/2007 from the World Wide Web: http://methcenter.psu.edu/lca/

McCutcheon, A. L. (1987) *Latent class analysis*. Thousand Oaks, CA: Sage.

Pepe, M.S., & Janes, H. (2005). Insights into latent class analysis. *UW Biostatistics Working Paper Series.* Working Paper 236. Retrieved February 28, 2007 from the World Wide Web: http://www.bepress.com/uwbiostat/paper236

Pepe, M.S., & Janes, H. (2006). Insights into latent class analysis of diagnostic test performance. *Biostatistics,* in press.

Stouffer, S.A., & Toby, J. (1951). Role conflict and personality. *American Journal of Sociology, 56*, 395-406.

Thompson, D. M. (2003). Comparing SAS-based applications of latent class analysis using simulated patient classification data. Unpublished doctoral dissertation, University of Oklahoma, Oklahoma City.

Thompson, D.M. (2006a). Performing latent class analysis using the CATMOD procedure. Contributed paper in the Statistics and Data Analysis section, 31st Annual SAS Users Group International (SUGI) Conference. Retrieved 5/18/2006 from the World Wide Web: http://support.sas.com/usergroups/sugi/ sugi31/winners.html#stat

Thompson, D.M. (2006b). Approaches to Obtaining Standard Errors for Parameter Estimates in Latent Class Analysis. Proceedings of Joint Statistical Meetings, American Statistical Association, Seattle, WA, August , 2006.

Thompson, D.M. (2006c). Research on latent class analysis. Retrieved 2/26/2007 from the World Wide Web: http://moon.ouhsc.edu/dthompso/latent%20variable%20research/lvr.htm

Uebersax, J.S., & Grove, W.M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine, 9*,559-572.

Vacek, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics, 41*, 959-968.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Name	David M. Thompson
Enterprise	Dept. of Biostatistics and Epidemiology, Univ. of Oklahoma Health Sciences Center
Address	CHB 321
City, State, ZIP	Oklahoma City, OK, 73190
Work Phone:	405-271-2229, ext. 48054
Fax:	405-271-2068
E-mail:	dave-thompson@ouhsc.edu
Web:	http://moon.ouhsc.edu/dthompso/