LATENT VARIABLE MODELS FOR ANALYSIS OF SURVEY DATA

Paul B. Massell, U.S. Census Bureau
SRD, Room 3209-4, Washington, D.C. 20233

Key Words: Latent Class Models, Missing Data, Identifiability

Introduction to Latent Class Analysis[1]

Latent class analysis is a fairly advanced statistical topic that may be viewed as a part of categorical data analysis. It is advanced since it builds on material from categorical data analysis, such as log-linear models, and ideas about model building with social science data, and algorithms for treating missing data. It is a fairly new branch of statistics; the first papers were written about 1950 but significant use of latent class methods did not occur until the 1970's when good computational methods became available (ref: GOO, DEM). Building latent class models poses all the challenges of traditional categorical data modeling plus others. This paper is an attempt to provide an introduction to the fundamental mathematical and statistical ideas that underlie latent class analysis in a style typical of introductory mathematical statistics texts rather than the style of social science monographs in which most introductory treatments now appear. This style might appeal most to readers with a mathematics or statistical computing background. After defining the basic ideas of the subject, we give a list of simple latent class models and state what is known about their identifiability and other properties. These models have recently been used to model response error in important federal surveys (ref: BIE).

Missing Data Problems

Latent class models are often used for the analysis of categorical data when one of the first two "missing data" cases listed below applies. "Missing data" here means at least part of the set of "true values" of the relevant variables are missing. Using this notion of "missing data" there are three cases to consider (ref: VER, p.5)

1. Missing data may exist in **hidden (i.e., lurking) variables**. Such a variable might simplify the relationships among the observed variables (i.e., the indicators). The goal is to determine whether such a variable exists for some assumed relationship to the indicators. With categorical indicators, one often models the hidden variable as a latent class variable. One may test if the assumption of the indicators being conditionally independent given the latent class produces a good fit.

2. Data may be missing due to **measurement error**. Note **response error** is often the dominant component of this error and is sometimes used as a synonym (ref: SAR, p. 547). The **true values** of an observed variable may be missing since the observed values contain measurement error. For example, employment status may be clearly defined for everyone in some population; there exists a true value for this variable even if it is not observed. The value that actually appears in the response file can be different from the true value due to various reasons, e.g., intentional or unintentional respondent error or data processing errors. In a reinterview survey, there may exist two different responses for the same true value.

3. Perhaps the most common type of missing data is that of partial nonresponse; for this case variable values exist for some respondents but are missing for others. When there are missing data, parameter estimates are likely to be biased because data that are missing are often not missing at random. Traditionally these types of problems have been handled by *imputation* methods (ref: article by Rovine in VEC) rather than with latent variables.

Example of hidden type latent variable in survey data

This is a sketch of work by R.A. Johnson, on the measurement of Hispanic Ethnicity in the U.S. based on 1986 National Content Test data (ref: JOH). The author notes there are two major aspects of Hispanicity; viz., use of the Spanish language and Hispanic ancestry. On the questionnaire, there are separate questions for Spanish language usage and Hispanic ancestry. The author shows that responses from these two questions can be combined in various ways using latent class variables to model each aspect of Hispanicity. The author suggests that such models are superior to linear probability error models whenever the response categories are not known to be valid. These linear models require reinterview data and make assumptions that often are not realistic.

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

Definition: If a variable cannot, in theory, be directly measured or observed it is called **latent**. Such variables are sometimes called indirectly observed or unobserved. We can extend this standard definition, by defining a variable to be **latent w.r.t. to a given dataset**, if the variable does not exist in a given dataset either explicitly or implicitly. If a latent variable is categorical, its categories are called **latent classes**.
(note: Latent class analysis and latent trait analysis are special cases of latent structure analysis. In latent class analysis the latent variable is categorical; in latent trait analysis and in factor analysis, the latent variable is continuous. (ref: VEC, p.226))

Relating a Latent Class Variable to its Indicators

The calculation of a latent variable often is done as follows. We assume a latent variable X exists. We then construct **conditional response probabilities** that relate the indicators (also called responses or observed or manifest variables) to the initially unknown X. In simple models, we often have dichotomous (i.e. 2-valued) indicators for a latent variable X, which is also assumed to be dichotomous. If we assume we have three indicators for X, denoted A,B,C, then we would need to compute six conditional probabilities, viz. $P(A=1|X=i)$, $P(B=1|X=i)$, and $P(C=1|X=i)$ for $i=1,2$, as well as the **latent (class) probability** $P(X=1)$. Note that the omitted parameters (e.g. $P(X=2)$) can be computed simply from the seven listed parameters. However, without additional information, we cannot say if such a model is **identifiable** for the given observed data, i.e. whether there are unique solutions for the seven parameters.

Decomposition of a joint distribution (table) using a latent class variable

Suppose we have two indicators, A with I classes and B with J classes. Then we could form a relative frequency (I by J) table T in which each cell value is the observed fraction of the total responses for that cell. T is an estimate of the joint distribution of A and B, P_T. If there exists a variable X with K classes, such that for each class $k = 1,2,\ldots, K$, **the variables A and B are conditionally independent given X**, and if $p_X(k)$ denotes the probability that X has value k, let it can be shown that P_T can be decomposed as follows:

$$P_T = \sum_{k=1}^{K} p_X(k) \cdot outer(\vec{p}_{A|k} , \vec{p}_{B|k})$$

(ref: AGR, p. 164). $\vec{p}_{A|k}$, $\vec{p}_{B|k}$ denote the distributions of A and B given X=k. The outer product of two vectors is the matrix formed by multiplying the column version of the first by the row version of the second. This is the simplest case of a joint distribution expressed as a mixture of several simpler distributions. In general one does not know a priori if such a decomposition is possible. Even when theory suggests such a decomposition is possible for some K (the number of latent classes), computation may be required to find the 'best' value of K.

Parameter Estimation for Latent Class Models

Maximum Likelihood Estimation (MLE) of Parameters. In latent class analysis, one tries to find a decomposition for the observed frequency table T analogous to the decomposition for the theoretical joint distribution described above. The modeler may suspect that such a decomposition is possible based on subject matter knowledge. He may also have an idea of the number of latent classes K. To confirm his suspicions, he may build and test a model as follows.

Consider a model with one latent variable X with two classes (i.e. a dichotomous variable) and three associated indicators. At first glance, it seems that direct estimation of the response probabilities is impossible since we have no direct measurement of X. However, the EM algorithm (see below) does allow estimation of the proportions of the two latent classes as well as the six response probabilities listed above. To estimate the proportions for this X requires just one parameter, $P(X=1)$. Thus we have a total, so far, of seven parameters to be estimated. If conditional independence is assumed (see below), then these seven parameters fully specify the model. As is standard in latent class modeling, we use maximization of the likelihood function associated with the given data and model. It is often convenient to maximize the logarithm of the likelihood function but this leads to the same parameter solutions.

Before specifying the assumptions using equations, it is useful to state some general properties of the indicator-latent variable relationship.

(1) The latent variable is the variable which comes closest to being consistent with the information provided by the indicators. Using geometric language, we may view the indicators as the projections of some unknown physical object, and the latent variable, when unique, is the most likely shape of this object.
(2) If there is not sufficient information provided by the indicators, there may be many solutions for the latent variables, all equally likely as measured by the likelihood function.
(3) If all the indicators are flawed in the same way, i.e., all fail to measure some common aspect of the hidden

variable, then the estimate of the latent variable will also be flawed. Using more statistical language, we can say that if all the indicators have a common bias, so will the estimated latent variable. This happens since the best that the estimation process can do is to extract common information from the indicators; the latent variable information can only be as good as that revealed by its indicators.

The EM algorithm: a brief overview.

The EM algorithm is a general scheme for treating a wide variety of applications in which there are missing data. The general idea is to start with some initial estimate of the missing data and using that estimate, estimate the optimal model parameter values. The optimization involves maximizing the likelihood function; i.e., MLE. One can now try to "improve the missing data" by calculating the expected value of the missing data, given (i) the current parameter estimates and (ii) the non-missing data. One continues this alternation of improvement in parameter estimates and improvement in missing data estimates until the procedure converges. The EM theory (ref: DEM) guarantees convergence to a local maximum under weak conditions (ref: VER, p.66)

The name of the algorithm derives from the fact that the step in which the expected values for the missing data values are calculated is called the **Expectation step** (the 'E' step) whereas the step in which the optimal parameter values are estimated is called the **Maximization Step** (the 'M' step).

The EM algorithm for the latent class problem can be implemented easily (see example EM2 below). The EM algorithm is both conceptually and computationally very simple (ref: VER, p.66). Generally, random starting values are good enough. The accuracy of the parameter estimates can be (roughly) assessed by examining their standard errors (see section on information matrices below).

The measured data is in general a compression and/or a distortion of the "theoretical" data; i.e., values of the variables that represent the true values. An example of compression is a measured variable that is the sum of two or more latent variables. An example of distortion, is a measured variable that is the noisy version of a single latent variable. Creating theoretical data from the observed data and the model (with the current estimates of parameters) can be done in various ways. In the EM algorithm, theoretical data is created (i.e., imputed) using the expectation operator.

Example EM1: Using the EM algorithm to distribute compressed data. Suppose we use a multinomial distribution to model a possibly unfair die that is thrown 100 times. Suppose our model for the 6 possible outcomes is as follows: $\vec{p}_X$ = (p/12, p/6, p/4, p/4, p/6, p/12) , i.e. we suspect that the low and high values don't occur as often as the middle values. Let xi= true number of times that the value i arose in 100 throws. Suppose we have compressed data ; e.g. we have direct estimates for x=1,2,3 but only the sum of values for x=4,5,6. Thus we observe (y1,y2,y3,yup) where y1=x1; y2=x2, y3=x3 but yup=x4+x5+x6. (e.g. (y1,y2,y3,yup)=(15, 20,10,55)). If the current estimate of p is p0, what are the expected values of the xi's ? The EM expectation operator here uses the observed data for values that are directly observed; thus E(x1) = y1, E(x2) = y2, E(x3) =y3. But the EM expectation operator requires use of **both the observed data and the model** to estimate x4,x5,and x6. In this case, the EM expectation operator simply leads to the familiar formula for the expected value of a multinomial variable: Then E(x4) = ((p/4)/ pup) * yup; where pup = p/4 + p/6 + p/12.Similar formulas hold for E(x5) , and E(x6). Thus the operator uses the model to distribute the **compressed** data, yup, to x4, x5, x6.

Example EM2: Using the EM algorithm to estimate the latent class sizes and conditional response probabilities. Assume the latent class variable X has two classes denoted 1 and 2. Let pi = P(X=1). Let A and B be two indicators of X; assume A and B have the same classes as X. Let eps = P(A=2| X=1). Assume eps represents this misclassification error for B as well. Assume also that the response errors for A and B are independent, given the value of X; i.e. A and B are **conditionally independent** given X. It is common to use a random number generator to choose starting values for pi and eps. With these starting values (a.k.a. initial 'estimates') for pi and eps we can form the **complete table** 'X by A by B' and compute the expected frequency value for each cell since each such expected value can be expressed in terms of pi, eps, and the observed data in the **incomplete table** 'A by B.' This is the expectation step. Using these estimates for the frequencies for each cell in the complete table, one can form the likelihood equation and estimate the values of pi and eps that maximize the log of the likelihood function, LogL. This process converges (in general) for any given starting values of pi and eps. However, different starting values may lead to different final values. It is for this reason that it is necessary to run the program many times to become convinced that one has covered a large portion of the parameter space and that the computed global maximum of LogL is close to the true global maximum.

Degrees of Freedom

There are various definitions of degrees of freedom. We use the term as applied to contingency tables [ref: AGR, p. 176].

Let N be the number of cells in the contingency table. The dimension of the table is not relevant for the discussion that follows; i.e., if the table is d-way, the results are independent of d. Assume a **sampling model SM** for the table. The two most commonly used are: Poisson and multinomial.

The Poisson has one parameter per (table) cell and there are no constraints on the N parameters; therefore there are N independent parameters. The multinomial also has one parameter per cell but there is one constraint on the N parameters; viz. that the sum of cell counts = n (= total sample size) is fixed; therefore there are N-1 independent parameters. Note that the parameters in SM refer only to the manifest (i.e. directly observed) variables.

Notation: **Let NP(SM) = number of independent parameters in the sampling model.**

Let Model G = model for a given contingency table.

Note that Model G typically imposes constraints on the parameters of SM in an indirect way. One usually uses a log-linear model for the table which uses a parameterization of the table which is equivalent to the simple parameterization in which the SM is given. One then often assumes that certain interaction terms in this new parameterization of SM are zero. However, when G contains latent variables it must contain parameters for the latent variable itself and its interactions with the manifest variables.

Notation: **Let NP(G) = number of independent parameters in model G**

When G contains latent variables it is possible that NP(G) > NP(SM).

Definition: The Degrees of Freedom (DoF) for testing a model G

Using the definitions of NP(SM) and NP(G) above, we define **DoF= NP(SM) - NP(G).**

Remark: The DoF is an integer, and if it is positive, there are theorems that show that it represents the DoF for test statistics that are measures of goodness of fit of G and that are asymptotically chi-squared. The two most commonly used statistics for testing goodness of fit in LCA are the Pearson chi-square statistic and the likelihood-ratio chi-square statistic [ref; VER.,p.19, AGR., p47-8].

Equivalent definition: The DoF for testing G = (number of independent data cell values) - (number of parameters in G that we need to estimate)

Identifiability

One goal of modeling is to estimate the size of the latent classes and the conditional response probabilities relating manifest variables to these latent classes. This goal is achieved most easily when there is a unique solution to the EM estimation algorithm that is run against the data. Even when there is a unique solution, it may not be easy to find it. Since the estimation often involves non-linear functions of the parameters, the possibility of local maxima for the loglikelihood function LogL arises. As is frequently the case in non-linear problems, one must explore the parameter space thoroughly before one can be convinced that one has found a local maximum that is also a global maximum. (ref: LEM, §10.2)

Multiple Solutions

Frequently, models with latent variables are not **globally identifiable**. This means that there are two or more sets of values for the parameters that yield the same globally maximal log-likelihood value. If there are just a few such maximal sets of parameter values, it may be possible to rule out all but one set on various grounds. For example, the modeler commonly has some a priori knowledge of the acceptable (i.e. sensible) ranges for some or all of the parameters. If all but one of the parameter sets have some unacceptable value, then the solution is the unique set of estimates which are all acceptable

Information Matrix

There is another way to determine **local** identifiability. One can compute numerically the Hessian matrix H (the matrix of 2nd order partial derivatives) of LogL with respect to the model parameters as follows.

$$H \equiv \partial^2 LogL / \partial\beta_i \; \partial\beta_j$$

Then define the **expected information matrix** as the expected value of H:

$$Info \equiv -E(H)$$

It can be shown that:

$$Info = N \sum_{k=1}^{N} (1/\pi_k)\cdot(\partial\pi_k/\partial\beta_i)\;(\partial\pi_k/\partial\beta_j)$$

where N is the number of cells in the (manifest) variable table and π_k is the probability associated with cell k in the incomplete (manifest) table (ref: VER, p.317).

The **expected information matrix** is sometimes called Fisher's information matrix. The information matrix, Info = - H, without the expected value, is sometimes called the **observed information**. (ref: LIT, p.85, a Bayesian argument suggests use of the observed information, a frequentist argument suggests use of the expected information). Info is positive semi-definite, in theory. A theorem of Goodman (ref:GOO) states a **sufficient** condition for **local identifiability** is that Info is positive definite (ref: VER., p. 69). In general, **global identifiability** can be determined only by comparison of local maxima of LogL over the parameter space.

Standard Errors

For asymptotically large samples, it is known that the parameter covariance matrix is the inverse of Info; i.e. Cov = $(\text{Info})^{-1}$. This implies that the standard error of the ith parameter is the square root of the ith diagonal element of Cov. For finite samples, the Cramer-Rao inequality shows that $(\text{Info})^{-1}$ is a lower bound for Cov.

Algorithms for Computing Parameter Estimates and Standard Errors

In addition to the EM algorithm used in LEM, there are several other algorithms that can handle missing data. Two commonly used algorithms for ML parameter estimation are both examples of gradient search algorithms. They are Fisher scoring and Newton-Raphson. They differ only in that Fisher scoring uses the expected information (i.e. the Fisher information) matrix whereas Newton-Raphson uses the observed information matrix. When all variables of a log-linear model are manifest (i.e. observed), the two algorithms are equivalent because in that case the observed and expected information matrices are equal. (ref: VER, p.65, AGR, p.114). The main disadvantage of these methods compared to the EM algorithm when the model includes latent variables, is that they need starting values close to the solution to converge (ref: VER, p.65). The EM algorithm, although not as fast as these algorithms near a solution, does converge to (at least) a local maximum under relatively weak conditions, even with bad starting values; in fact, in general, random starting values are good enough (ref: VER, p.66). (As implemented in LEM, after the EM algorithm converges, the expected information matrix is computed and inverted to find the standard errors (ref: VER,p.66).)

Examples of Models run on LEM

General properties of solutions to latent class models.

a) In order to have a chance of creating identifiable models, one needs to impose constraints on the parameters. One usually assumes that one or more easy to state properties hold, such as conditional independence or identical response probabilities across subpopulations. Such properties, if there is some reason to believe that they hold, we call **a reasonable assumption set** or **a reasonable constraint set.**

b) Specification of independence in LEM
If two variables appear in a log-linear model, we express that by including at least the main effect parameters for both variables. If we wish to assume that the variables are independent we do that simply by omitting any interaction terms involving those two variables. Thus "independence" is the relationship that requires the fewest number of parameters to specify.

c) Switching of latent classes
Often when one runs a program that allows latent class models (e.g. LEM), one finds that on one run, a certain latent class, e.g. the largest one, is assigned one code whereas on the next run it is assigned some other code. This happens because there simply is no information to determine a specific set of codes for the latent classes. This variability of assignment of codes is not a serious problem if there is only one population being modeled since if the relative sizes of the latent classes are well separated, the modeler will probably be able to provide a useful interpretation to each latent class. However when 2 or more subpopulations are being modeled, switching of latent classes codes within the subpopulations **may** lead to multiple solutions (see example M5 below).

d) In the table below we present a summary of facts for five simple latent class response error models. See (ref:MAS) for details on these and other models.

Open questions

There is much need for more results on identification of latent class models. Are proofs of global identifiability possible for certain types of models ? For which models is computational exploration of the parameter space the only way to determine identifiability ? A precise analysis of how identification relates to the data space would also be helpful (e.g. loss of identifiability for Hui-Walter models when data for two or more subpopulations are not sufficiently separated).

References:

Introductory books on Categorical Data and Log-linear Models

AGR: Agresti, Alan, Categorical Data Analysis; Wiley, 1990.
CHR: Christensen, Ronald, Log-linear models and Logistic Regression, 2nd ed., Springer, 1997

Latent Class website: http://ourworld.compuserve.com/homepages/jsuebersax/index.htm
(includes the page:LCA Frequently Asked Questions)

Advanced Books:

LDA: Bijleveld, C, van der Kamp, L, (eds.) Longitudinal Data Analysis, Designs, Models, and Methods, Sage Pubs, 1998

SAR: Sarndal, Swensson, Wretman, Model Assisted Survey Sampling, Springer, 1992

VEC: von Eye, A, Clogg, C., Latent Variables Analysis; Applications for Developmental Res., Sage Pubs, 1994

VER: Vermunt, Jeroen K., Log-Linear Models for Event Histories, Sage Pubs, 1997

Articles: Theory, Applications of LCA; EM algorithm

BIE, Biemer, Paul, Bushery, John. A Markov Latent Class Analysis of Classification Error in the CPS, (RTI report to U.S. Census Bureau, March 1999)

DEM, Dempster, A.P., Laird, N.M., Rubin, D.B., Maximum Likelihood from Incomplete Data via the EM Algorithm, J. Roy. Stat. Soc.Ser. B 39, 1-38, 1977.

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

JOH, Measurement of Hispanic Ethnicity in the U.S. Census: An Evaluation Based on Latent-Class Analysis, JASA, March 1990, Vol. 85, No. 409

Hui, S.L., Walter, S.D., Estimating the Error Rates of Diagnostic Tests, Biometrics, 36, 1980, p.167-171

LAI, Laird, Nan, The EM Algorithm, in Handbook of Statistics, by C.R. Rao, Vol . 9

LEM: The manual for LEM: A general program for the analysis of categorical data; by Jeroen K. Vermunt, Sept. 1997

LIT, Little, R.J.A., Rubin, D.B., Statistical Analysis with Missing Data, Wiley, 1987

MAS, Massell, Paul B., Identifiability of Simple Latent Class Response Error Models: Computational Experiments, (informal paper)

SIN, Sinclair, M D., Gastwirth, J. L., On Procedures for Evaluating the Effectiveness of Reinterview Survey Methods: Application to Labor Force Data, JASA, Sept. 1996, Vol. 91, p.961-969.

Simple Latent Class Response Error Models and their Identifiability
(all models have 1 latent variable X and 2 or 3 indicators A,B,(C); all are dichotomous)
(DoF refers to degrees of freedom for testing; recall DoF=(#(cells)-1 - #(ind parms))
(A|X denotes 'A given X'; CRP denotes 'conditional response probability')

Model #	Verbal Model Description	Symbolic Model Description	Data : Table for Indicators	Selected Output; Conclusions

M1	2 indicators with independent and equal CRP's	X , A|X ,
		B|X = A|X,
		(note: DoF = 0)
			A by B:
			[300 25 50 400]
				P(X=1) not unique;
				Not Identifiable;

M2	3 indicators with ind. CRP's	{X,XA,XB,XC} (note: DoF=0)	A by B by C : [100 2 4 6 8 10 12 300]	CRP's not unique; Not Identifiable
M3	3 indicators with ind. and equal CRP's	X , A|X, B|X = A|X C|X = A|X; (note: DoF=4)	A by B by C: [100 2 4 6 8 10 12 300]	Appears to be Identifiable (even when small data values are set to 0)

M4	2 indicators and 1 grouping variable G;
	G defines 2 subpops;
	(Hui-Walter method)
		X|G, A|XG, B|XG, for each indicator, CRP's for two subpops are equal	G by A by B
			
			[150 10 20 250
			150 5 3 150]
				Identifiable for this data; note: method assumes subpops have different latent class probs

M5	2 indicators and G; Given a latent class and a subpop, CRP's for both indicators are equal	X|G A|XG B|XG= A|XG	G by A by B [100 1 2 300 190 3 4 380]	Not identifiable but solution space is interesting