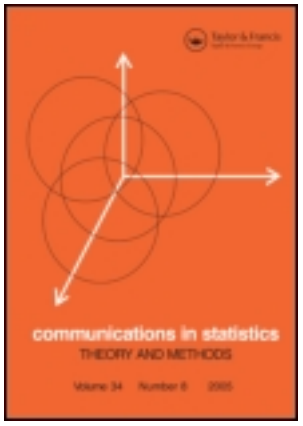

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Some Remarks on Latent Variable Models in Categorical Data Analysis

Alan Agresti [a] & Maria Kateri [b]
[a] Department of Statistics , University of Florida , Gainesville , Florida , USA
[b] Institute of Statistics , RWTH Aachen University , Aachen , Germany
Published online: 27 Jan 2014.

To cite this article: Alan Agresti & Maria Kateri (2014) Some Remarks on Latent Variable Models in Categorical Data Analysis, Communications in Statistics - Theory and Methods, 43:4, 801-814

To link to this article: http://dx.doi.org/10.1080/03610926.2013.814783

Communications in Statistics—Theory and Methods, 43: 801–814, 2014

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2013.814783

Taylor & Francis
Taylor & Francis Group

Some Remarks on Latent Variable Models in Categorical Data Analysis

ALAN AGRESTI[1] AND MARIA KATERI[2]

[1]Department of Statistics, University of Florida, Gainesville, Florida, USA
[2]Institute of Statistics, RWTH Aachen University, Aachen, Germany

We present an overview of some important and/or interesting contributions to the latent variable literature for the analysis of multivariate categorical responses, beginning with Lazarsfeld's introduction of latent class models. There is by now an enormous literature on latent variable models for categorical responses, especially in the context of including random effects in generalized linear mixed models, so this is necessarily a highly selective overview. Due to space considerations, we summarize the main ideas, suppressing details. As part of our presentation, we raise a couple of questions that may suggest future research work.

Keywords Association models; Latent class models; Local independence; Mixture models; Ordered categorical data; Random effects models; Rasch model.

Mathematics Subject Classification 62H25; 62H20; 62J99.

1. Introduction: The Lazarsfeld Latent Class Model

Latent variable models have by now a very long history. The development of methods for categorical data lagged behind that for continuous variables, and this is true also for latent variable methods. For continuous variables, research on methods such as factor analysis dates to the start of the twentieth century, with significant contributions by psychologists such as Charles Spearman. For categorical variables, landmark work was done by the sociologist Paul Lazarsfeld (1901–1976).

In particular, Lazarsfeld (1950a,b) introduced the basic *latent class model*, treating a contingency table as a finite mixture of unobserved tables generated under a conditional independence structure. For a set of categorical response variables $(Y_1, Y_2, \ldots, Y_T)$, the model assumes the existence of a latent categorical variable Z such that for each possible sequence $(a_1, \ldots, a_T)$ of response values and each category z of Z,

$$\begin{aligned} &P(Y_1 = a_1, \ldots, Y_T = a_T \,|\, Z = z) \\ &\quad = P(Y_1 = a_1 \,|\, Z = z) \cdots P(Y_T = a_T \,|\, Z = z). \end{aligned}$$

Received October 18, 2012; Accepted June 7, 2013
Address correspondence to Alan Agresti, Department of Statistics, University of Florida, Gainesville, FL 32611-8545, USA; E-mail: aa@stat.ufl.edu

The model did not receive much attention in terms of application or further methodological development until the publication of the text by Lazarsfeld and Henry (1968). Google Scholar reports that this book has now received more than 1500 citations. Lazarsfeld and Henry (1968, p. 22) stated that "The defining characteristic of the latent structure models is the *axiom of local independence*." That is, within a latent class, responses to different Y_t are independent. This principle was invoked by other authors about the same time (e.g., McDonald, 1967).

Lazarsfeld and Henry (1968) considered continuous as well as discrete latent variables, but the discrete case received the most attention in the years following publication of their book. More recently, the latent variable literature has commonly utilized continuous latent variables, and our discussion below will also consider such models; the unifying aspect of our presentation is its focus on categorical observed variables.

In reading this classic 1968 book, more than 50 years after its publication, one is likely to find quite striking the challenge provided by fitting this basic model. F. Mosteller had apparently suggested to Lazarsfeld the method of finding estimates by solving "accounting equations" that equate relative frequencies to corresponding marginal probabilities of various orders. The authors show how to solve these equations iteratively, using the "determinantal method." They note that the resulting estimates approximate the minimum chi-squared estimates that are in Neyman's best asymptotically normal (BAN) class of optimal estimates. Anderson (1954) showed asymptotic properties of such estimators, and he provided an appendix in their book dealing with their asymptotic properties.

In the context of the 1950s and 1960s, the computational aspects of fitting the model were daunting. The authors stated (p. 13) that when $\{Y_t\}$ have >2 categories, the model has so many restrictions that its practical application seems doubtful. In fact, not long after that it was applied in such cases and also with more than one latent variable. In an email to the first author of this article on May 15, 2012, Neil Henry wrote, "While Lazarsfeld was many things, he was not a statistician. The people he had working on LSA with him were sociology students with mathematical abilities, but no interest in inferential statistics.... The papers, mostly unpublished, that I inherited in 1960 were full of these accounting equation solution techniques. Eventually I learned enough history to realize that he had adopted Karl Pearson's 'method of moments' technique of estimation. Maximum likelihood estimation was impossible (as a practical estimation technique) in the 40s and 50s, of course."

Not many years later, Goodman (1974) showed how to fit the basic Lazarsfeld latent class model for the case of a discrete latent variable, using maximum likelihood (ML). His algorithm was an early application of the EM algorithm, three years before the classic article by Dempster et al. (1977). The algorithm treats the data on Z as missing. The E (expectation) step in each iteration calculates pseudo-counts for the unobserved table using the working conditional distribution for $(Z \mid Y_1, \ldots, Y_T)$. The M (maximization) step treats pseudo counts as data and maximizes the pseudo-likelihood, by fitting the model that treats the responses as conditionally independent within each category of Z.

From properties of the EM algorithm, this method of fitting the model is computationally simple and stable, and each iteration increases the likelihood. However, convergence can be very slow. Some of the later literature on such models instead used a Newton–Raphson algorithm (e.g., Haberman, 1988). Now, some software packages for fitting such models (e.g., Latent GOLD) use EM at the

initial stage, then switch to Newton–Raphson to speed convergence. Regardless of the iterative method, problematic issues are that the log likelihood can have local maxima (especially as the number of latent classes increases), and as the model increases in complexity, identifiability becomes an issue.

2. Some Extensions of Latent Class and Latent Variable Models

The impact of Lazarsfeld's model and of the subsequent text by him with Henry has been substantial, particularly in the social sciences. The model has been applied frequently, and many books have been written since theirs about the model and its generalizations. A popular reference book is Hagenaars and McCutcheon (2002), and a more recent one is by Collins and Lanza (2010). Much of the methodological generalization was performed by statisticians having a social science orientation, such as Leo Goodman and Clifford Clogg. In this section, we'll give examples of such generalizations, some of which enter the much more general realms of mixture modeling and of generalized linear mixed models containing continuous random effects, for which the literature has exploded in the past quarter century.

2.1. *Using Goodman's Association and Correlation Models*

The standard latent class model treats both the observed variables and the latent variables as nominal scale, not taking into account any ordering that may exist among the categories. Goodman (1979) proposed a class of association models that provide structured form for associations between variables, when at least one of them is ordinal. An example is his *uniform association model*: for expected frequencies $\{\mu_{ij}\}$ in a two-way contingency table and for sets of equally-spaced scores $\{u_i\}$ and $\{v_j\}$ for the rows and columns, the model has the loglinear form

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j.$$

With $\{u_i = i\}$ and $\{v_j = j\}$, the model implies a common value of *local odds ratios*, $(\mu_{ij}\mu_{i+1,j+1})/(\mu_{i,j+1}\mu_{i+1,j}) = \exp(\beta)$, for pairs of adjacent rows and adjacent columns. Later work by Goodman and others showed that such association models fit well when there is an underlying bivariate normal distribution.

With ordinal categorical responses, it seems natural to have ordered latent classes also, such as by assuming ordinal structure (e.g., uniform association) between the latent variable and each ordinal response variable. For example, Agresti and Lang (1993) modeled agreement among many raters evaluating carcinoma with an exchangeble model having the same β between each ordinal variable and the latent variable. The model parameters describe two components of agreement: The strength of association between classifications by pairs of raters (governed by the size of β), and the degree of heterogeneity among the observers' marginal distributions. Strong agreement requires strong association and relatively minor heterogeneity. For the data analyzed, the model fit well with three latent classes: The first may reflect cases with rater agreement that carcinoma was present; the second may reflect strong disagreement, by which some raters thought carcinoma was present and some thought it was not; the third may reflect cases with rater agreement that carcinoma was not present.

In later research, Goodman popularized canonical correlation models as an alternative type of structure that can describe ordinal associations. For them, the

association term has linear-by-linear structure on the expected frequency scale, rather than its log. For two-way tables, Gilula (1984) related latent class models to canonical correlation models.

2.2. *Latent Mixture Model for Summarizing Goodness of Fit*

For any model, it is nearly always the case that we do not expect it to hold perfectly in the population of interest. Hence, for categorical data, with sufficiently large n, traditional goodness-of-fit statistics such as the Pearson and deviance chi-squared statistics will reject the model with high probability. This is the case even if, in practical terms, the lack of fit is minor and the model is actually adequate. As a way of addressing this and developing a measure to reflect model inadquacy, Rudas et al. (1994) proposed a mixture model. For a model for a contingency table with true probabilities $\boldsymbol{\pi}$ (of any fixed dimension), they expressed

$$\boldsymbol{\pi} = (1 - \rho)\boldsymbol{\pi}_1 + \rho\boldsymbol{\pi}_2,$$

where $\boldsymbol{\pi}_1$ are the model-based probabilities and $\boldsymbol{\pi}_2$ are unconstrained. This always holds for *some* values of ρ, in a set with upper limit 1. They proposed their index of lack of fit as the smallest such ρ possible for which this holds. That is, it is the fraction of the population that cannot be described by the model.

Their approach recognizes the late George Box's famous quote that "All models are wrong, but some are useful," with "useful" meaning that the minimal ρ is very close to 0. Note that such a mixture model contrasts with the standard latent class model in which both $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ satisfy a conditional independence structure. In fact, many of the innovative uses of latent variable models in the past quarter century have moved away from the previously unifying concept of local independence to more general forms of mixtures.

2.3. *Latent Mixing of Logistic Regression and Count Data Models*

As we'll discuss later, many applications of latent variable models for categorical data now involve some sort of mixing of ordinary models such as logistic regression models. In an early and innovative application of this type, Follman and Lambert (1989) analyzed the effect of a dosage of poison on the probability of death of a protozoan of a particular genus. In the application considered, there were two genuses expected, but they were unobserved. For $\pi_i(x) =$ the probability of death at log dose level x for genus type i, $i = 1, 2$, and $\rho =$ the probability a protozoan belongs to genus type 1, their model is

$$\pi(x) = \rho\pi_1(x) + (1 - \rho)\pi_2(x), \quad \text{where } \text{logit}[\pi_i(x)] = \alpha_i + \beta x.$$

The curve for $\pi(x)$ is a weighted average of two curves having the same logistic shapes but different intercepts. For the data they analyzed, the deviance decreased by 21.3 ($\text{df} = 2$) compared to using a single logistic regression curve. (Perhaps surprisingly, it decreases by only 1.7 when we instead assume a normal mixture of curves rather than a binary weighting, an issue we'll address in a later section.)

A few years later, Lambert (1992) proposed a mixture model for count response variables. Her *zero-inflated Poisson* (ZIP) regression model is useful in applications

in which some observations must be zero and others are zero just by chance (e.g., number of times went to a gym in the past week; some people will never go, whereas some weeks those who are members of a gym will fail to go). This can be regarded as a latent class model in which one class consists of the necessarily 0 responses and the other class consists of those subjects whose observations follow a standard parametric distribution such as the Poisson. This type of approach is quite useful in other settings in statistical modeling to allow for two (or more) types of subjects in a sample. A direct generalization of the ZIP model to deal with the overdispersion commonly enountered with count data is the *zero-inflated negative binomial* model (Greene, 1994). Another generalization deals with repeated measures of zero-inflated data (Min and Agresti, 2005).

2.4. *Item Response Models*

Our discussion so far has focused on models with discrete latent classes. However, since Lazarsfeld's original proposal of the latent class model, many latent variable models for categorical responses have used continuous latent variables. A large and early-to-develop literature of this type dealt with applications in which subjects had unobserved *latent traits*, such as their "ability" for the performance on an exam.

For a set of items, let y_{it} denote the response of subject i on item t. An important application is a set of questions on an exam, in which $y_{it} = 1$ denotes a correct response on question t. For a binary response, Rasch (1961) proposed a model having the form

$$\text{logit}[P(Y_{it} = 1 \mid u_i)] = u_i + \beta_t.$$

In the context of a set of questions on an exam, u_i is a latent ability measure for subject i. Rasch treated $\{u_i\}$ as fixed effects. To estimate $\{\beta_t\}$, he used the Fisherian approach of eliminating $\{u_i\}$ using conditional ML. He assumed the local independence structure by which, conditional on u_i, the T responses by subject i are independent.

Since Rasch's landmark work, a huge literature has evolved on such *item response models* (also referred to as latent trait models). It is increasingly common to treat u_i as an unobserved latent variable (a "random effect") rather than as a fixed effect. As discussed later, most such models assume normality for the random effect, but some authors have taken other approaches. For example, Tjur (1982) averaged over u_i in a nonparametric manner in obtaining a marginal distribution for estimating $\{\beta_t\}$. He showed that the Rasch model for T items implies a model for the observed 2^T table that afficionados of contingency table modeling will recognize as the *quasi-symmetry* (QS) model (e.g., Agresti, 2013, Sec. 11.7.1), which is a model that takes the complete symmetry structure for the T items but then permits marginal distributions to differ. In fact, he showed that ML estimates of $\{\beta_t\}$ for the QS model are identical to conditional ML estimates of $\{\beta_t\}$ for the Rasch model. Analogous results apply for nonparametric treatment of random effects in extensions of the Rasch model for ordered categories and corresponding ordinal QS loglinear models (e.g., Agresti, 1993).

In an alternative nonparametric approach, Lindsay et al. (1991) assumed that u_i can take a finite number q of ordered numerical values,

$$P(U = a_k) = \rho_k, \quad k = 1, \ldots, q,$$

for unknown q, $\{a_k\}$ and $\{\rho_k\}$. For this *Rasch mixture model*, we might expect to get an increasingly precise estimate of the actual mixture distribution, using more and more mass points $\{a_k\}$, as the sample size n increases. However, Lindsay et al. (1991) showed that the likelihood function increases in q but reaches a maximum when $q = (T + 1)/2$. Regardless of the sample size, the fitted discrete mixture distribution need not give a good representation of the actual one. Note also that such a model differs from the ordinary latent class model, since it assumes structure for $P(Y_{it} = 1 \mid u_i)$, whereas the ordinary latent class model assumes no structure for $P(Y_t = y_t \mid Z = z)$.

In the 1980s, other research was more in the vein of extending ordinary factor analysis to categorical responses. Examples are Bartholomew's (1980) development of factor analysis for categorical data and his 1984 work on latent variable models for ordered categorical data. These and earlier work such as by Christoffersen (1975) and Muthén (1978) were influential in psychometrics and education. For a recent survey of such work, see Bartholomew et al. (2011).

2.5. *Generalized Linear Mixed Models*

In most applications, such as those addressed by item response theory, it is more realistic to assume a continuous latent variable than a discrete one. Using a discrete latent variable, such as Lindsay et al. (1991) suggested, provides merely a rough approximation for this more realistic structure. Although these days it seems quite natural to insert into a generalized linear model a set of random effects having a normal distribution or some other continuous distribution, it took some time for such an approach to catch on. Pierce and Sands (1975) may have been the first to propose using logistic regression with a random intercept term that was assumed to be normally distributed, in an unpublished technical report. According to Pierce (2011 oral communication with first author), the article received a lukewarm reception from referees of a major journal and the authors never bothered to revise and resubmit, but this paper has received numerous citations since then.

Pierce and Sands used Gauss–Hermite quadrature to integrate out the random effects in order to approximate the likelihood function. This is still a practical and effective approach for generalized linear mixed models (GLMM) with simple random effects structure. In a later highly influential article, Breslow and Clayton (1993) developed penalized quasi-likelihood (PQL) as a simple alternative to Gauss–Hermite quadrature for more complex random effects structure for which quadrature is impractical. However, PQL can be highly biased for a categorical response with large variance component (Lin, 1997), and literature since then has focused on other approaches. Ways of efficiently fitting such models is still a relevant research topic, as models with more complex random effects structure are proposed (e.g., multilevel models). Zipunnikov and Booth (submitted) suggested that higher-order Laplace approximations work better in practice than some methods that, in theory, produce ML but may be very slow to do so (such as Monte Carlo EM). The Bayes approach with diffuse priors is also used to approximate ML (e.g., with MCMC), but it is still unclear how well this works in models having a large number of parameters and a large number of random effects.

Another still relevant question for current research concerns the importance of the choice of distribution assumed for random effects in such models. In the GLMM context, perhaps the most important extension of the basic latent class structure

of conditional independence within levels of an unobserved variable, let y_{it} denote observation t in cluster i, $t = 1, \ldots, T_i$, with random effects $\boldsymbol{u}_i$ for cluster i. For $\mu_{it} = E(Y_{it} \mid \boldsymbol{u}_i)$, a GLMM has the form

$$g(\mu_{it}) = \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \boldsymbol{z}_{it}^T \boldsymbol{u}_i$$

for a link function $g(\cdot)$ and fixed effects $\boldsymbol{\beta}$. Typically such models make the assumption that $\boldsymbol{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ for unknown "variance components." Alternative assumptions for the random effects include a nonparametric approach (e.g., Aitkin, 1999) and a mixture of normal distributions (Molenberghs et al., 2010).

The ordinary multivariate normal assumption has the advantage of natural use in multivariate cases (e.g., because of the variety of possible correlation structures) and for multilevel models. But what if we assume normality and the actual distribution is quite different? Most literature shows relatively little effect in bias and efficiency of model parameter estimates in using an incorrect random effects distribution (e.g., Neuhaus et al., 1992). In addition, the accuracy of the predicted random effects does not seem to be much affected by such violations (McCulloch and Neuhaus, 2011); different distributions can yield quite different predicted values (that resemble in distribution the assumed shape) but have similar MSE performance in how close they fall to the actual random effects. However, when var($\boldsymbol{u}_i$) depends on covariates, between-cluster effects may be quite sensitive to misspecification of the distribution of $\boldsymbol{u}_i$ (Heagerty and Zeger, 2000). One reason for this is the implied diminution of marginal effects relative to conditional effects, the diminution being greater for cases with larger variance for the random effects.

There is at least one case where misspecification of shape can be relevant. Agresti et al. (2004) found a significant efficiency loss for estimating parameters in the logistic random intercept model when normality is assumed for a random intercept but the true distribution is a two-point mixture. This is especially true when var(u_i) and T are large. Such binary latent variables with large variance are natural in applications in which a population has extreme polarization, such as modeling responses to several items dealing with opinions about whether abortion should be legal in various situations.

2.6. *Latent Transition Analysis*

Latent class analysis does not handle dynamic latent variables that change systematically over time. Discrete dynamic latent variables are mostly analyzed by Markov models that predict the probability of movement between the categories of the latent variable between successive time points. Latent transition analysis is an extension of latent class analysis in a longitudinal framework that allows the modeling of more complex situations, involving static and dynamic latent variables, for which their stage-sequential change is investigated (Graham et al., 1991; Collins and Wugalter, 1992).

For details and references about latent transition models, see Collins and Lanza (2010). For an overview and a comparative study of latent class models, Markov models, latent Markov models, and latent transition analysis applied for modelling a stage-sequential development; see Kaplan (2008). Cho et al. (2010) proposed a latent transition analysis model with a mixture Rasch model as the measurement model that permits within-class variability on the latent variable.

3. Standard Categorical-Response Models Generated by Latent Variable Models

Many standard models for categorical response variables can be motivated by latent variable models. This is useful to know, even if in a particular application we are not explicitly accounting for latent variables.

Examples are models for binary data such as the probit model and the logistic regression model. Early uses of such models were in dose–response studies, for which a *tolerance distribution* relates to an unobserved latent variable. Specifically, in dose–response studies, a tolerance distribution with cumulative distribution function (cdf) F for the dosage x that induces a "success" response implies the model

$$G^{-1}[\pi(x)] = \alpha + \beta x$$

for standardized cdf G corresponding to F. That is, G^{-1} becomes the link function for the model. The choice $G = \Phi$ (standard normal) gives the probit model (Bliss, 1935), whereas $G =$ standard logistic gives the logit link (Berkson, 1944). Related latent variable models that can induce such binary regression models are the *threshold model* and the *utility model*; see Agresti (2013, Sec. 7.1.1) for details.

Standard models for ordinal response variables also result from latent variable models, as shown by Anderson and Philips (1981). Suppose the underlying response y^* has mean relating to explanatory variables by $E(y^*) = \boldsymbol{\beta}^T\boldsymbol{x}$, and the random errors come from a distribution having standardized cdf G. Suppose also that there are thresholds (cutpoints) $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$ such that the observed ordinal response y falls in category j of c outcome categories if $\alpha_{j-1} < y^* \leq \alpha_j$. Then,

$$P(y \leq j \,|\, \boldsymbol{x}) = P(y^* \leq \alpha_j \,|\, \boldsymbol{x}) = G(\alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}).$$

This implies that the model for y is $G^{-1}[P(y \leq j \,|\, \boldsymbol{x})] = \alpha_j - \boldsymbol{\beta}^T\boldsymbol{x}$. That is, the appropriate link function is the inverse of the standardized cdf for the errors for the latent variable model. So, the cumulative logit model with the same effects for each cumulative probability (i.e., the model with the so-called *proportional odds* property) applies when G is logistic, and the cumulative probit applies when G is normal. That is, the cumulative logit (probit) model fits well when an ordinary linear regression model holds for an underlying logistic (normal) response.

This derivation suggests that such models are designed to detect shifts in *location* (center), not *dispersion* (spread), at different settings of the explanatory variables. In fact, the standard ordinal models that have the same effect parameters for each cumulative probability imply that the conditional distributions of y at different settings of explanatory variables are stochastically ordered. When this is badly violated, simple models such as the proportional odds version of the cumulative logit model tend not to fit well.

4. Brief Summary of Other Latent Variable Modeling

Since the seminal book by Lazarsfeld and Henry, the literature on latent class and latent variable modeling for categorical variables has continually expanded at a rapid rate, both for new methodological developments and for applications of models. Here we'll briefly mention some other advances of each type and mention

some useful software. In each subsection, we'll first list articles dealing with latent classes and then articles dealing with latent variables.

4.1. *Methodological*

- Clogg (1981) proposed latent structure models for the analysis of mobility tables and examined their relationship to some earlier mobility models. He introduced the quasi-latent structure and noted that this new model is similar to the "mover-stayer" model, but differs from it by positing two latent classes of movers instead of a single one.
- de Leeuw and van der Heijden (1991) presented correspondence analysis and latent class models and discussed their relationship. They pointed out that Good (1965) seems to have been the first to express the correspondence analysis model as a latent class model.
- Vermunt (2003) proposed multilevel latent class models, relaxing the assumption of local independence, and he also considered complex sampling designs.
- Vermunt et al. (2008) proposed the use of latent class analysis for the multiple imputation of incomplete categorical data, as an alternative to loglinear analysis. The uncertainty about the unknown model parameters is reflected in the imputations by a nonparametric bootstrap procedure.
- Formann (1992) proposed a linear logistic latent class analysis for polytomous responses that is a Rasch-type model but constrains the unknown class sizes and the latent response probabilities. Fitting is again achieved by the EM algorithm.
- Anderson and Vermunt (2000) noted that the Goodman association model arises when the observed $\{Y_t\}$ are conditionally independent given a latent Z that is conditionally normal (given observed variables).
- Gueorguieva and Agresti (2001) proposed a probit model for joint modeling of clustered binary and continuous responses, based on underlying joint normality.
- The text by Skrondal and Rabe-Hesketh (2004) presented many models having greater complexity than we've had space to discuss in this article. Skrondal and Rabe-Hesketh (2007) provided a survey of latent variable modelling, including an extended literature review.

4.2. *Applications*

A large variety of latent class and latent variable model applications appear in quite diverse fields. However, they are quite common for applications in the social sciences, education, and psychology, such as described in Rost and Langeheine (1997). Some more recent applications follow.

- DeSantis et al. (2008) developed a penalized latent class model for correlated high-dimensional ordinal data and applied it in a study of schwannoma, a peripheral nerve sheath tumor, that included 3 clinical subtypes, 7 binary, and 16 ordinal histological measures.
- Reboussin and Ialongo (2010) modeled drug use among students who suffer from attention deficit hyperactivity disorder (ADHD), using (1) a longitudinal latent transition model with latent classes for stages of marijuana

use that describes probability of transitioning between stages, and (2) a cross-sectional latent class model that constructs ADHD subtypes and describes the influence of subtypes on transition rates.

- Moustaki and Steele (2005), in a demographic application, explored women's fertility preferences and family planning behavior in Bangladesh. They proposed a latent variable model with continuous latent variables for manifest variables that are a mixture of categorical and survival outcomes, possibly censored. Covariate effects, both on the manifest and the latent variables, are incorporated into the model.
- Lin et al. (2008), in an aging study, modeled repeated transitions between independence and disability states of daily living using multivariate latent variables. A state-specific latent variable represents an individual's tendency to remain in a state, and accounts for correlation among repeated sojourns in the same state. Correlation among sojourns across states is accounted for by correlation between different latent variables.

4.3. *Software, and Internet Resources*

- *MLLSA* software: This is the first latent class program written by C. Clogg (1977).
- *Latent GOLD* software: Written by J. Vermunt and J. Magidson and marketed by Statistical Innovations, this software fits a wide variety of mixture models, including latent class models, nonparametric mixtures of logistic regression models, Rasch mixture models, zero-inflated models, multilevel models, and models with continuous latent variables.
- *TWOMISS* software: This program, by Albanese and Knott (1992), fits one- or two- factor logit-probit latent variable models to binary data when observations may be missing.
- In `R`: The `LCA` package performs a latent class analysis with k classes. Latent transition analysis (LTA) can be performed by the package `LTA`. Estimation can use ML (applying the EM algorithm) or Bayesian methods (applying MCMC methods). The package `poLCA` deals with the estimation of latent class models and latent class regression models for polytomous outcome variables, using EM and Newton–Raphson algorithms for ML estimation. `FlexMix` provides a general framework for finite mixture models and latent class regression (Leisch, 2004). It uses the EM algorithm and provides the E-step and all data handling, while the M-step can be supplied by the user to easily define new models. Existing drivers implement mixtures of standard linear models, generalized linear models and model-based clustering. Furthermore, `mmlcr` is appropriate for mixed-mode latent class regression models, for which the manifest variables can be of mixed types, including longitudinal or single response, normal or censored-normal or categorical or Poisson. The function *lmer* (linear mixed effects in R) in the R package *Matrix* can be used to fit generalized linear mixed models. See also the *lme4* package and the function glmmML in the *glmmML* package. These use adaptive Gauss–Hermite quadrature. The function *glmmPQL* in the MASS library can fit GLMMs using penalized quasi-likelihood. The R package *MCMCglmm* can fit them with Markov Chain Monte Carlo methods.
- Useful websites dealing with various aspects of latent variable modeling for categorical variables include:

```
www.people.vcu.edu/~nhenry/LSA50.htm (Neil Henry reminiscences)
statisticalinnovations.com/products/aboutlc.html (Latent GOLD)
www.stata.com/meeting/2nasug/lclass.pdf (Stata)
www.msu.edu/~chunghw/downloads.html R: LCA, LTA and LCPA
userwww.service.emory.edu/~dlinzer/poLCA/ (R: poLCA)
cran.r-project.org/web/packages/flexmix/index.html (R: flexmix)
www.stat.rutgers.edu/home/buyske/software.html (R: mmlcr)
cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf (R: lme4)
support.sas.com/kb/30/623.html (SAS)
www.statmodel.com (Mplus)
spitswww.uvt.nl/~vermunt (LEM et al.)
www.john-uebersax.com/stat (overview)
faculty.chass.ncsu.edu/garson/PA765/latclass.htm (overview)
```


5. Summary and Future Challenges

Latent variable models have a long and substantial history for categorical data analysis, of which we've been able to discuss only a few highlights in this article. Many of the methods most commonly used by statisticians for categorical data analysis have latent variable justifications.

In the future, research for latent variable modeling is likely to be driven by the same challenges that statisticians face in an increasing number of applications. In particular, how does one deal with large data sets with huge numbers of variables? The Follman and Lambert (1989) model mentioned in Sec. 2.3 is a good example of a simplistic starting point. Suppose a population consists of a mixture of two genetic types, but instead of merely observing a single predictor such as dosage of a drug we've observed a very large number of predictors, of which very few may be related to the response of interest. The challenge is compounded for the Bayesian approach, in which it often unclear how to choose priors to yield an "objective Bayes" analysis when the number of parameters is huge.

As we apply latent variable models as researchers or as methodologists or as practitioners, however, we should not forget the dangers of *reification*—acting as if an assumed latent variable truly measures the characteristic of interest (Gould, 1981). Their use, especially in initial research studies of a particular question, is tentative in nature, and often valuable merely for suggesting models to use in follow-up studies. Related to this, anyone who has conducted research with latent variable models realizes the potential for misuses as methodologists with limited understanding of them apply them in practice. So, besides the research challenge of developing methods for ever more complex settings such as "big data," there is the more mundane challenge of teaching and explaining the methods and their limitations to decrease the frequency of such misuses.

References

Agresti, A. (1993). Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonal parameters. *Scand. J. Statist.* 20:63–71.

Agresti, A. (2013). *Categorical Data Analysis*. 3rd ed. New York: Wiley.

Agresti, A., Lang, J. (1993). Quasi-symmetric latent class models, with application to rater agreement. *Biometrics* 49:131–139.

Agresti, A., Caffo, B., Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Comput. Statist. Data An.* 47:639–653.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 55:218–234.

Albanese, M. T., Knott, M. (1992). TWOMISS: a computer program for fitting a one- or two-factor logit-probit latent variable model to binary data when observations may be missing. Technical Report, Statistics Department, London School of Economics and Political Science.

Anderson, J. A., Philips, P. R. (1981). Regression, discrimination, and measurement models for ordered categorical variables. *Appl. Statist.* 30:22–31.

Anderson, C. J., Vermunt, J. K. (2000). Log-multiplicative models as latent variable models for nominal and/or ordinal data. *Sociol. Methodol.* 30:81–121.

Anderson, T. W. (1954). On estimation of parameters in latent structure analysis. *Psychometrika* 19:1–10.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *J. Roy. Stat. Soc. B* 42: 293–321.

Bartholomew, D., Knott, M., Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. 3rd ed. New York: Wiley, Ch. 4–6.

Berkson, J. (1944). Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.* 39:357–365.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Ann. Appl. Biol.* 22:134–167.

Breslow, N., Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* 88:9–25.

Cho, S.-J., Cohen, A. S., Kim, S.-H., Bottge, B. (2010). Latent transition analysis with a mixture item response theory measurement model. *Appl. Psychol. Measure.* 34:483–504.

Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika* 40:5–32.

Clogg, C. C. (1981). Latent structure models of mobility. *Amer. J. Sociol.* 86:836–868.

Collins, L. M., Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis*. Hoboken, NJ: Wiley.

Collins, L. M., Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavi. Rese.* 27:131–137.

de Leeuw, J., van der Heijden, P. G. M. (1991). Reduced rank models for contingency tables. *Biometrika* 78:229–232.

Dempster, A. P., Laird, N. M., Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39:1–38.

DeSantis, S. M., Houseman, E. A., Coull, B. A., Stemmer-Rachamimov, A., Betensky, R. A. (2008). A penalized latent class model for ordinal data. *Biostatistics* 9:249–262.

Follman, D. A., Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *J. Amer. Statist. Assoc.* 84:295–300.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *J. Amer. Statist. Assoc.* 87:476–486.

Gilula, Z. (1979). Singular value decomposition of probability matrices: Probabilistic aspects of latent dichotomous variables. *Biometrika* 66:339–344.

Gilula, Z. (1984). On some similarities between canonical correlation models and latent class models for two-way contingency tables. *Biometrika* 71:523–529.

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

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

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* 74:537–552.

Gould, S. J. (1981). *The Mismeasure of Man*. New York: W. W. Norton & Company.

Graham, J. W., Collins, L. M., Wugalter, S. E. Chung, N. J., Hansen, N. B. (1991). Modeling transitions in latent stage-sequential processes: a substance use prevention example. *J. Consult. Clin. Psychol.* 59:48–57.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Technical report.

Gueorguieva, R., Agresti, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous responses. *J. Amer. Statist. Assoc.* 96:1102–1112.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociolog. Methodol.* 18:193–211.

Hagenaars, J., McCutcheon, A., eds. (2002). *Applied Latent Class Analysis*. New York: Cambridge University Press.

Heagerty, P. J., Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statist. Sci.* 15:1–19.

Kaplan, D. (2008). An overview of Markov chain methods for the study of stage-sequential developmental processes. *Develop. Psychol.* 44:457–467.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34:1–14.

Lazarsfeld, P. F. (1950a). The logical and mathematical foundation of latent structure analysis. In: Stouffer, S. A., Guttmann, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., Clausen, J. A., eds. *Studies in Social Psychology in World War II, Vol. IV, Measurement and Prediction*. Princeton, NJ: University Press, Ch. 10, pp. 342–412.

Lazarsfeld, P. F. (1950b). The interpretation and computation of some latent structures. In: Stouffer, S. A., Guttmann, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., Clausen, J. A., eds. *Studies in Social Psychology in World War II, vol. IV, Measurement and Prediction*. Princeton, NJ: Princeton University Press, Ch. 11, pp. 413–472.

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

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *J. Statist. Software* 11:1–18.

Lin, H., Guo, Z., Peduzzi, P. N., Gill, T. M., Allore, H. G. (2008). A semiparametric transition model with latent traits for longitudinal multistate data. *Biometrics* 64:1032–1042.

Lin, X. (1997). Variance component testing in generalized linear models with random effects. *Biometrika* 84:309–326.

Lindsay, B., Clogg, C., Grego, J. (1991). Semi-parametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc.* 86:96–107.

McCulloch, C. E., Neuhaus, J. M. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics* 67:270–279.

McDonald, R. P. (1967). *Nonlinear Factor Analysis*. Psychometric Monograph 15. Bowling Green, OH: Psychometric Society.

Min, Y., Agresti, A. (2005). Random effects models for repeated measures of zero-inflated count data. *Statist. Modell.* 5:1–19.

Molenberghs, G., Verbeke, G., Demetrio, C. G. B., Vieira, A. M. C. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statist. Sci.* 25:325–347.

Moustaki, I., Steele, F. (2005). Latent variable models for mixed categorical and survival responses, with an application to fertility preferences and family planning in Bangladesh. *Statist. Modell* 5:327–342.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika* 43:551–560.

Neuhaus, J. M., Hauck, W. W., Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* 79:755–762.

Pierce, D. A., Sands, B. R. (1975). Extra-Bernoulli variation in regression of binary data. Technical Report, Statistics Dept., Oregon State University.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In: Neyman, J., ed. *Proc. 4th Berkeley Symposium on Mathematics, Statistics, and Probability*, Berkeley, CA: University of California Press, Vol. 4, pp. 321–333.

Reboussin, B. A., Ialongo, N. S. (2010). Latent transition models with latent class predictors: attention deficity hyperactivity disorder subtypes and high school marijuana use. *J. Roy. Statist. Soc. A* 173:145–164.

Rost, J., Langeheine, R., eds. (1997). *Applications of Latent Trait and Latent Class Models in the Social Sciences*. Munster, New York, Munchen, Berlin: Waxmann.

Rudas, T., Clogg, C. C., Lindsay, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *J. Roy. Statist. Soc. B* 56:23–639.

Skrondal, A., Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC.

Skrondal, A., Rabe-Hesketh, S. (2007). Latent variable modelling: a survey. *Scand. J. Statist.* 34:712–745.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scand. J. Statist.* 9:23–30.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociol. Methodol.* 33:213–239.

Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociol. Methodol.* 38:369–397.

Zipunnikov, V., Booth, J. Closed form GLM cumulants and GLMM fitting with a SQUAR-EM-LA_2 algorithm. Submitted.