RC23391 (W0410-159) October 25, 2004
Mathematics

IBM Research Report

Latent Class Regression on Latent Factors

Yasuo Amemiya
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Melanie Wall, Jia Guo
University of Minnesota

Latent Class Regression on Latent Factors

Yasuo Amemiya, IBM T. J. Watson Research Center
Melanie M. Wall and Jia Guo, University of Minnesota

Abstract

There are two types of latent variable modeling often used in health sciences research; structural equation modeling with continuous factors, and latent class analysis with unordered-categorical latent segments. This paper develops a statistical methodology for a more general model with both continuous and categorical latent variables. Observed measurement types are allowed to include continuous and ordered categorical responses. Model fitting methods and associated statistical inference procedures are discussed. This methodology can be useful in health science applications, where a health condition outcome variable is a latent classification, and some of possible predictor variables are psychological/behavioral constructs. An example relating a underlying eating disorder condition to a physical appearance satisfaction construct is presented.

Keywords: latent class model, structural equation analysis, categorical response, latent health condition, psychological construct

1. Introduction

In the research of public health, psychology, and social sciences, it is very common to have variables or constructs that cannot be measured directly by a single observable variable but instead are hypothesized to be the driving force underlying a series of observed variables. The underlying or unobservable variables are called latent variables and a long literature exists and is evolving of methods for measuring them and examining relationships among them (for a brief history see Bartholomew and Knott, 1999). Just as observed variables can be characterized as continuous or categorical, so can latent variables. Latent class analysis (Lazarsfeld and Henry, 1968; Clogg, 1995; Hagenaars and McCutcheon, 2002) considers models for measuring latent categorical variables hypothesized to take on a finite number of values which partition the population into a finite number of discrete groups. Factor analysis or general latent factor (trait) analysis (Lawley and Maxwell, 1971; Moustaki and Knott, 2000) considers models for measuring continuous latent variables.

Many research questions propose to investigate the relationship between a categorical response variable and continuous predictors. Logistic regression is the obvious technique to use when both the outcome and the predictors are observed directly. But, when either the categorical outcome or the continuous predictors cannot be observed directly, different methods are needed. If the response variable is measured via a latent class model and the predictors are observable, the approach called latent class regression can be applied to analyze the relationship (Dayton and Macready, 1988, Bandeen-Roche, et al. 1997). Another case is where the predictors are latent factors and the categorical response variable is observable. This case may be considered as a kind of errors-in-variables problem (Fuller, 1987; Carroll, et al., 1995). In particular since the outcome is categorical, this would be a nonlinear errors-in-variables problem and has been extensively studied with several methods proposed (Burr, 1988; Carroll, et al., 1995). Following the above reasoning, the need remains to develop a model and method for including both categorical and continuous latent variables, which will be the focus of this paper.

To motivate the need for such an analysis technique, we consider an example from a project conducted to study adolescent nutrition and obesity called Project EAT (Neumark-Sztainer, et al., 2002). One research question of interest is whether an adolescent girl's body satisfaction could predict her eating disorders risk class. Body satisfaction had been hypothesized by the researchers to be a continuous latent factor measured by a battery of self report Likert items related to satisfaction with different parts of one's body (e.g. hips, shoulders, waist, etc.). The outcome variable of interest, eating disorders risk class, was hypothesized to be a categorical latent variable. The researchers hypothesized that there were different types of eating disorders risk related to girls engaging in purging vs. those engaging in restriction behaviors. A checklist of 9 unhealthy weight control behaviors was asked on the questionnaire. No absolute classification rule based on the checklist of 9 behaviors exits, but a latent class model can be used to measure the different latent classes of eating disorders risk. In section 6, this example will be examined further in particular the modeling of the relationship between the latent body satisfaction and eating disorders risk.

The paper is organized as follows. In section 2, we review the latent factor model, the latent class model, the latent class regression model and introduce notation. In section 3, based on the models in section 2, we propose a new model for the case when both the categorical response and continuous predictors are latent variables. In section 4, the maximum likelihood method is considered and two different computational algorithms are proposed. In particular, the Monte Carlo Expectation and Maximization (MCEM) algorithm is demonstrated with flexible assumptions of the distribution of the latent factors and the Gaussian quadrature approximation followed by quasi-Newton maximization method is proposed for the case when the latent factors are normally distributed (implementable in, e.g., SAS PROC NLMIXED). In section 5, we apply this model to the Project EAT data to analyze the relationship between body satisfaction and the characteristics of the latent classes for eating disorders risk formed in the data.

2. Motivating Models

2.1 Latent factor analysis model

Suppose we have a data set with n individuals. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iP})^T$ be a P dimensional vector of continuous observations for each individual i ($i = 1, \ldots, n$). The factor analysis model (Lawley and Maxwell, 1971) assumes there are $Q\,(Q < P)$ latent factors $\mathbf{f}_i = (f_{i1}, \ldots, f_{iQ})^T$, which relate to $\mathbf{X}_i$ in the following way,

$$\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Gamma}\mathbf{f}_i + \boldsymbol{\epsilon}_i. \tag{1}$$

Here the P dimensional vector $\boldsymbol{\mu}$ and $P \times Q$ matrix $\boldsymbol{\Gamma}$ contain known and unknown scalars and $\boldsymbol{\epsilon}_i$ is a P dimensional vector of random error with $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$, $Var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$, and $\boldsymbol{\epsilon}_i$ is assumed independent of $\mathbf{f}_i$. Furthermore, $\boldsymbol{\Psi}$ is assumed to be diagonal, which implies along with the assumption that $\boldsymbol{\epsilon}_i$ and $\mathbf{f}_i$ are independent, that any correlations found between the elements in the observed vector $\mathbf{X}_i$ are due to their relationship with common $\mathbf{f}_i$ and not due to some spurious correlation between $\boldsymbol{\epsilon}_i$. As is, model (1) is not generally identifiable and so the following error-in-variables parametrization will be considered,

$$\mathbf{X}_i = \begin{pmatrix} \boldsymbol{\lambda}_0 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Lambda} \\ \mathbf{I} \end{pmatrix} \mathbf{f}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

where $\mathbf{I}$ is a $Q \times Q$ identity matrix, $\mathbf{0}$ is a Q dimensional vector of zero, $\boldsymbol{\lambda}_0$ and $\boldsymbol{\Lambda}$ are known or unknown scalars. This form allows straightforward interpretation of $\mathbf{f}_i$ and does not lose the generality of model (1). A measurement model of the form (2) possibly after reordering the elements of $\mathbf{X}_i$ and centering all qualities is widely used in practice (see e.g. Fuller, 1987; Jöreskog and Sorbom, 1996).

Estimation and inference for model (2) have been considered in many different ways relying on varying assumptions of the distributions of $\mathbf{f}_i$ and $\boldsymbol{\epsilon}_i$ (Bollen, 1989). Under the assumption that $\mathbf{f}_i$ and $\boldsymbol{\epsilon}_i$ are normally distributed, the maximum likelihood method is commonly used to estimate $\boldsymbol{\lambda}_0$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ because of its optimal properties. Furthermore, it has been shown (Anderson and Amemiya, 1988; Browne and Shapiro, 1988) that the estimators for model (2) using the maximum likelihood method assuming normality are asymptotically consistent even if the normality assumptions for $\mathbf{f}_i$ and $\boldsymbol{\epsilon}_i$ are violated. All estimation methods are based on the number of factors Q being fixed and known. A common technique for choosing the number of factors in the exploratory model (i.e. where the elements of $\boldsymbol{\Lambda}$ are freely estimated) is to examine a screen plot of the eigenvalues of $\mathbf{S}$, i.e. the covariance matrix for $\mathbf{X}$. The number of factors can be chosen to be the number of eigenvalues before the elbow in the plot.

More recently the latent factors model has been considered in the following more general form,

$$f(\mathbf{X}_i) = \int f(\mathbf{X}_i|\mathbf{f}_i) f(\mathbf{f}_i) d\mathbf{f}_i. \tag{3}$$

Assuming $f(\mathbf{X}_i|\mathbf{f}_i) \sim N\left(\begin{pmatrix} \boldsymbol{\lambda}_0 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Lambda} \\ \mathbf{I} \end{pmatrix} \mathbf{f}_i, \boldsymbol{\Psi}\right)$ and $f(\mathbf{f}_i) \sim N(\boldsymbol{\mu}_f, \boldsymbol{\Phi})$, this is just the normal factor analysis model (2). In the form (3), more flexible distributions can be assumed for $f(\mathbf{X}_i|\mathbf{f}_i)$, e.g. any exponential family (Moustaki and Knott, 2000; Sammel, et al., 1997, Dunson, 2000 and 2003).

2.2 Latent Class model

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^T$ denote the vector of the observed variables for each individual i, the basic finite mixture model (McLachlan and Peel, 2000) is,

$$f(\mathbf{Y}_i) = \sum_{k=1}^{K} f(c_i = k) f_k(\mathbf{Y}_i|c_i = k). \tag{4}$$

Here K is the number of latent classes, c_i is the latent class variable, $f_k(\mathbf{Y}_i|c_i = k)$ is the class-specific distribution for $\mathbf{Y}_i$, and $\pi_k \equiv f(c_i = k)$ denotes the marginal probability of belonging to latent class k such that $\sum_{k=1}^{K} \pi_k = 1$.

The special case of model (4) when all elements of $\mathbf{Y}_i$ are binary variables is often called the "latent class model" in the literature (Clogg, 1981 and

1995). Moreover, in the latent class model it is generally assumed that the elements of $\mathbf{Y}_i$ are mutually independent within classes. The latent class model is then,

$$f(\mathbf{Y}_i) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \pi_{j|k}^{Y_{ij}} (1 - \pi_{j|k})^{1-Y_{ij}}, \qquad (5)$$

where $\pi_{j|k} \equiv Pr(Y_{ij} = 1 | c_i = k)$ is the probability that $Y_{ij} = 1$ when the i^{th} individual is in the latent class k. Although it is possible to constrain certain parameters $\pi_{j|k}$ to zero, it is most common to allow them to be freely estimated by the data. Given a fixed number of classes K, maximum likelihood estimates can be obtained straightforward by the EM algorithm. It is common to fit models with different numbers of classes and compare them by BIC values and choose the model with the smallest BIC (Collins, et al., 1993).

2.3 Latent Class Regression model

Latent class regression models (Dayton and Macready, 1988; Bandeen-Roche, et al. 1997) build a relationship between a latent class variable and other observable continuous or categorical predictor variables of interest. Consider again the latent class model (5), it may be thought of as a measurement model relating $\mathbf{Y}_i$ to c_i. That is, a model for how c_i is measured by $\mathbf{Y}_i$. Now consider additional observed covariates $\mathbf{X}_i$ where it is of interest to investigate if $\mathbf{X}_i$ can directly effect the latent class membership, i.e. if $\mathbf{X}_i$ can be a good predictor for the latent classes. It will be assumed that $\mathbf{Y}_i$ is conditionally independent of $\mathbf{X}_i$ given c_i. That is, intuitively the model assumes the only reason that $\mathbf{X}_i$ is related to $\mathbf{Y}_i$ is because it is related to c_i. The following is the latent class regression model,

$$f(\mathbf{Y}_i|\mathbf{X}_i) = \sum_{k=1}^{K} f(c_i = k|\mathbf{X}_i) \prod_{j=1}^{J} \pi_{j|k}^{Y_{ij}} (1 - \pi_{j|k})^{1-Y_{ij}}. \qquad (6)$$

Bandeen-Roche, et al. 1997 assumed a generalized logit model for $f(c_i = k|\mathbf{X}_i)$ given by

$$Pr(c_i = k|\mathbf{X}_i) = exp(\mathbf{X}_i^{oT} \boldsymbol{\beta}_k) / \sum_{k=1}^{K} exp(\mathbf{X}_i^{oT} \boldsymbol{\beta}_k) \qquad k = 1, \cdots, K,$$

where $\mathbf{X}_i^o = (1, X_{i1}, \cdots, X_{iP})^T$ and $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{Pk})^T$ are $P+1$ dimensional vectors. In the above equation, the K^{th} class is taken as the reference class, i.e. $\boldsymbol{\beta}_K = (0, 0, \ldots, 0)^T$. Then, similar to the latent class model, given the number of classes K fixed, estimates can be obtained by the EM algorithm.

3. Latent Class Regression on Latent Factors

In the previous section, we have introduced three models that construct relationships between latent variables and observed variables. As an extension to these models, we will build on the idea of latent class regression and introduce a new model that considers the structural relationship between a latent class variable and latent factors.

Let c_i be a latent class variable with K categories and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^T$ be a J dimensional observed vector with binary elements, which is used to measure c_i via the latent class model (5). Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iP})^T$ be a P dimensional observed vector with continuous elements used to measure a Q dimensional latent factor $\mathbf{f}_i$ via the latent factor model (2) or more generally (3).

Similar in spirit to latent class regression model (6), our new model considers regressing the latent class c_i on the latent factors $\mathbf{f}_i$. One of the fundamental assumptions of this new model is that $\mathbf{Y}_i$ is conditionally independent of $\mathbf{X}_i$ given the latent variables c_i and $\mathbf{f}_i$. This means that the model assumes $\mathbf{Y}_i$ and $\mathbf{X}_i$ are only related because the variables they are measuring are related. This is a natural assumption when modeling relationships between variables measured with error, i.e. we want to model the relationship between the underlying variables, not the ones with error. Thus we consider the joint distribution of $\mathbf{Y}_i$ and $\mathbf{X}_i$ to be

$$f(\mathbf{Y}_i, \mathbf{X}_i) = \qquad (7)$$

$$\sum_{k=1}^{K} \int f(\mathbf{Y}_i|c_i = k) f(\mathbf{X}_i|\mathbf{f}_i) f(c_i = k|\mathbf{f}_i) f(\mathbf{f}_i) d\,\mathbf{f}_i \qquad (8)$$

where

(latent class model)

$$f(\mathbf{Y}_i|c_i = k) = \prod_{j=1}^{J} \pi_{j|k}^{Y_{ij}} (1 - \pi_{j|k})^{1-Y_{ij}},$$

(latent factor model)

$$f(\mathbf{X}_i|\mathbf{f}_i) \sim N_p \left(\begin{pmatrix} \boldsymbol{\lambda}_0 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Lambda} \\ \mathbf{I} \end{pmatrix} \mathbf{f}_i, \boldsymbol{\Psi} \right),$$

(generalized logit model)

$$f(c_i = k|\mathbf{f}_i) = \pi_k(\mathbf{f}_i) = \frac{exp(\mathbf{f}_i^{oT} \boldsymbol{\beta}_k)}{\sum_{k=1}^{K} exp(\mathbf{f}_i^{oT} \boldsymbol{\beta}_k)},$$

(specified latent factor distribution)

$$f(\mathbf{f}_i) \sim F(\boldsymbol{\Phi}).$$

The parameters $\pi_{j|k}$, $\boldsymbol{\lambda}_0$, $\boldsymbol{\lambda}_1$ and $\boldsymbol{\Psi}$ are all as defined before in section 2. While most commonly the distribution $F(\boldsymbol{\Phi})$ for the underlying factors will be normal, it is not necessary so we present it here to be any specified distribution with unknown parameters $\boldsymbol{\Phi}$. Note that like the latent class regression model in section 2.3, we use the generalized logit link, i.e. $\log\left(\frac{\pi_k(\mathbf{f}_i)}{\pi_K(\mathbf{f}_i)}\right) = \mathbf{f}_i^{oT}\boldsymbol{\beta}_k$, where $\mathbf{f}_i^o = (1, \mathbf{f}_i)^T$, $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \cdots, \beta_{Qk})^T$ and $\boldsymbol{\beta}_K = (0, 0, \cdots, 0)^T$ indicating class K as the reference class. The parameter $\boldsymbol{\beta}_k$ is a $Q+1$ dimensional vector relating the latent factors to the probability of being in a particular latent class. The big difference between this model and that in section 2.3 is that here $\mathbf{f}_i$ is not observed directly. In this model, the parameters of interest are the vector $\boldsymbol{\beta}_k$'s, which describe the relationship between the latent factors and the latent classes.

4. Maximum Likelihood Estimation

Given the parametric model (7) and the *i.i.d.* data $(\mathbf{Y}_i, \mathbf{X}_i)$, for $i = 1, \ldots, n$, estimation of the model parameters can proceed via the maximum likelihood method. Let $\mathbf{Z}_i = (\mathbf{Y}_i, \mathbf{X}_i)$, $\mathbf{d}_i = (c_i, \mathbf{f}_i)$, and $\boldsymbol{\theta} = (\{\pi_{j|k}\}, \boldsymbol{\lambda}_0, \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \{\boldsymbol{\beta}_k\}, \boldsymbol{\Phi})$ is the vector of parameters relating $\mathbf{Z}_i$ and $\mathbf{d}_i$. Thus the likelihood function for the model (7) can be written as

$$L_o = \prod_{i=1}^{n} f(\mathbf{Z}_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \int f(\mathbf{Z}_i, \mathbf{d}_i; \boldsymbol{\theta}) d\,\mathbf{d}_i, \qquad (9)$$

where the notation for the integral over $\mathbf{d}_i$ is taken very generally to include the continuous integral for $\mathbf{f}_i$ and the summation over c_i. This likelihood function is hard to maximize due to the integration of the latent variables for which there is no closed form solution. Hence two numerical methods for performing the full maximum likelihood are described in this section: Monte Carlo Expectation and Maximization algorithm (MCEM) and Gaussian quadrature followed by quasi-Newton algorithm.

4.1 MCEM Algorithm

It is natural to consider the latent variables, $\mathbf{d}_i$, as missing data and implement the EM algorithm for maximizing (9). Instead of working with the observed data likelihood L_o, we consider the complete data likelihood

$$L_c = \prod_{i=1}^{n} f(\mathbf{Z}_i, \mathbf{d}_i; \boldsymbol{\theta}).$$

The E-step obtains the expectation of the log complete data likelihood given the observed data and the current parameter estimates, i.e. $\boldsymbol{\theta}_l$.

$$\begin{aligned} E(\log L_c | \mathbf{Z}_1, \ldots, \mathbf{Z}_n; \boldsymbol{\theta}_l) &= \\ \sum_{i=1}^{n} \int \log f(\mathbf{Z}_i, \mathbf{d}_i; \boldsymbol{\theta}) f(\mathbf{d}_i | \mathbf{Z}_i; \boldsymbol{\theta}_l) d\,\mathbf{d}_i & \\ &\equiv g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}; \mathbf{Z}). \end{aligned}$$

For the latent class regression on latent factor model (7), unfortunately we do not have a closed form for $f(\mathbf{d}_i | \mathbf{Z}_i; \boldsymbol{\theta}_l)$ and consequently we do not have a closed form solution for the integral in $g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}; \mathbf{Z})$. Hence, we propose to use the Monte Carlo method to obtained an approximation to $g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}; \mathbf{Z})$. First note that

$$\begin{aligned} g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}; \mathbf{Z}) &= \sum_{i=1}^{n} \int \log f(\mathbf{Z}_i, \mathbf{d}_i; \boldsymbol{\theta}) f(\mathbf{d}_i | \mathbf{Z}_i; \boldsymbol{\theta}_l) d\,\mathbf{d}_i \\ \sum_{i=1}^{n} & \frac{\int \log f(\mathbf{Z}_i, \mathbf{d}_i; \boldsymbol{\theta}) f(\mathbf{Z}_i | \mathbf{d}_i; \boldsymbol{\theta}_l)}{\int f(\mathbf{Z}_i | \mathbf{d}_i; \boldsymbol{\theta}_l) f(\mathbf{d}_i; \boldsymbol{\theta}_l) d\,\mathbf{d}_i} f(\mathbf{d}_i; \boldsymbol{\theta}_l) d\,\mathbf{d}_i. \end{aligned}$$

Hence, given the current $\boldsymbol{\theta}_l$, a Monto Carlo sample $(\mathbf{d}_i^1, \ldots, \mathbf{d}_i^M)$ is generated from $f(\mathbf{d}_i; \boldsymbol{\theta}_l)$, and the expectation can be approximated by an average

$$\begin{aligned} g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}; \mathbf{Z}) &\approx \sum_{i=1}^{n} \frac{1}{M} \sum_{m=1}^{M} [\log f(\mathbf{Z}_i, \mathbf{d}_i^m; \boldsymbol{\theta})\; W_i^m] \\ &\equiv g_{\boldsymbol{\theta}_l}^{MC}(\boldsymbol{\theta}; \mathbf{Z}), \end{aligned}$$

where

$$W_i^m = \frac{f(\mathbf{Z}_i | \mathbf{d}_i; \boldsymbol{\theta}_l)}{\frac{1}{M}\sum_{m=1}^{M} f(\mathbf{Z}_i | \mathbf{d}_i^m; \boldsymbol{\theta}_l)}.$$

Note that the same Monte Carlo sample is used to evaluate the integral in the denominator of the weights in the expectation. Now we note that $\log f(\mathbf{Z}_i, \mathbf{d}_i^m; \boldsymbol{\theta})$ can be factorized into four parts corresponding to the four parts of the model (7), i.e.,

$$\begin{aligned} \log f(\mathbf{Z}_i, \mathbf{d}_i^m; \boldsymbol{\theta}) &= \log f(\mathbf{Y}_i | c_i^m; \{\pi_{j|k}\}) \\ &+ \log f(\mathbf{X}_i | \mathbf{f}_i^m; \boldsymbol{\lambda}_0, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) \\ &+ \log f(c_i^m | \mathbf{f}_i^m; \{\boldsymbol{\beta}_k\}) + \log f(\mathbf{f}_i^m; \boldsymbol{\Phi}). \end{aligned}$$

The M-step is to maximize $g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}; \mathbf{Z})$ with respect to $\boldsymbol{\theta}$. Based on (10), we see that $g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}; \mathbf{Z})$ has four parts with distinct parameters associated with each.

Hence, we can maximize each component separately as a straightforward weighted regression (weighted by W_i^m) in order to obtain $\boldsymbol{\theta}_{l+1}$.

The MCEM algorithm will iterate between the E-step and M-step until the parameter estimates converge according to some criteria. In order to decrease the Monte Carlo error at the E-step, a large M should be used. But it has been pointed out that it is inefficient to choose a large M when $\boldsymbol{\theta}_l$ is far from the ML estimate (Wei and Tanner, 1990; Booth and Hobert, 1999). Following recommendation, it is preferable to start with a small M and increase it for each iteration to $M_l = M_0 + Tl$, where M_l is the sample size for the Monte Carlo step at the l^{th} iteration and M_0 and T are positive constants. We monitor the convergence of the EM algorithm by plotting $\boldsymbol{\theta}_l$ versus the iteration l.

Standard Error Estimates of the parameter estimates from MCEM can be obtained by inverting the information matrix of the log likelihood function based on the observed data. We apply Louis' formula (Louis, 1982)

$$\begin{aligned} I_{\mathbf{Z}}(\boldsymbol{\theta}) &= E_{\mathbf{d}}\left(-\frac{\partial^2 L_c(\mathbf{Z},\mathbf{d};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right) \\ &- Var_{\mathbf{d}}\left(\frac{\partial L_c(\mathbf{Z},\mathbf{d};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right), \end{aligned}$$

evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. The expectation and variance are taken with respect to the conditional distribution of the latent variable $\mathbf{d} = (\mathbf{d}_1, \ldots, \mathbf{d}_n)^T$ given the observed data $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^T$ and parameter $\boldsymbol{\theta}$. These conditional expectations are difficult to evaluate because as before in the EM algorithm the conditional distribution of $\mathbf{d}$ given $\mathbf{Z}$ is unavailable. Hence we cannot sample from the conditional distribution or get the closed forms for the expectations. By a similar method described for the parameter estimation, we can switch the conditional expectation to a weighted unconditional expectation with respect to $\mathbf{d}$ and then use the Monte Carlo method to approximate the expectation. Here a large M is used as it is unnecessary to iterate.

4.2 Gaussian quadrature with quasi-Newton algorithm

We note that the MCEM algorithm introduced above is flexible with regard to the assumptions of the distribution of the latent factors. That is, it was not necessary to assume $\mathbf{f}_i$ as normally distributed. Consider again the likelihood function L_o associated with model (7). Because the latent classes are discrete, it can be written as

$$\begin{aligned} L_o &= \prod_{i=1}^{n}\sum_{k=1}^{K}\int f(\mathbf{Y}_i|c_i=k)f(\mathbf{X}_i|\mathbf{f}_i) \\ &\qquad f(c_i=k|\mathbf{f}_i)f(\mathbf{f}_i)d\,\mathbf{f}_i. \end{aligned}$$

We note that the observed data likelihood is a function of the integral of the latent factors $\mathbf{f}_i$. In the special case, when the $\mathbf{f}_i$ is normally distributed, this can be approximated by adaptive Gaussian quadrature method (Golub and Welsch 1969, or Table 25.10 of Abramowitz and Stegun 1972). Then given a closed form approximation to the integral involving the normal factors $\mathbf{f}_i$, the observed likelihood can then be approximate in a closed form. With the closed form approximation for the likelihood, the maximization of it can be carried out through a quasi-Newton algorithm.

In fact, this method of Gaussian quadrature approximation followed by quasi-Newton maximization can be implemented using the "general" likelihood function in PROC NLMIXED in SAS. Appendix A gives code demonstrating how this can be done.

Although detailed investigation of the computational speed and accuracy of this method as compared to MCEM is beyond the scope of the current paper, the estimation for the example considered herein takes 5 times longer using MCEM. It should also be noted that for increasing numbers of factors, the integration in both methods may be computationally prohibitive.

5. Example

Project EAT (Neumark-Sztainer, et al., 2002) is a comprehensive study of adolescent nutrition and obesity. Self-report survey data were collected from students in 7^{th} and 10^{th} grade at 31 Twin Cities schools in the 1998-99 school year. A portion of the survey data collected for the 2113 girls in Project EAT is used in this current example. The particular relationship of interest in this study is that between body satisfaction (a latent factor) and eating disorders risk (a latent class).

In the Project EAT data set, for each individual i ($i = 1, \ldots, 2113$) let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i9})^T$ be the 9 dichotomous questionnaire items indicating the self-reported use of unhealthy weight control behaviors within the past year (i.e. "Have you done any of the following things in order to lose weight or keep from gaining weight during the past year: fasting, eating very little food, taking diet pills, making myself vomit, using laxatives, using diuretics, using food substitute, skipping meals, smoking more cigarettes?"). Furthermore, let $\mathbf{X}_i = (X_{i1}, \ldots, X_{i5})^T$ indicate the five items measuring body satisfaction (i.e., "How satisfied are you with your: body shape, waist, hips, thigh, stomach?"). Each element was measured on a 5 point Likert scale where the anchors were 1="very dissatisfied" and 5="very satisfied". Despite the discrete nature of these Likert responses, we will treat $\mathbf{X}_i$ as a continuous variable in this data analysis and center each element to mean zero.

5.1 Exploratory data analysis

Assume that underlying the observed responses $\mathbf{Y}_i$ is a latent class variable c_i with categories representing different typologies of eating disorders risk. In practice, we will not know the "correct" number of latent classes in the model. The number of latent classes K needs to be investigated before fitting the relationship between latent variables. Here we present the exploratory latent class analysis of the 9 observed indicators asking which unhealthy weight control behaviors had been used within the past year. Table 3 shows the estimated latent class model parameters and associated BIC values. The 3-class model shows the best BIC fit value. Examining the $\{\pi_{j|k}\}$ for the 3-class model leads to a class of girls who are basically not doing any of the behaviors (56.4%), a class who are doing just the restricting behaviors (i.e. eating very little and skipping meals) (35.2%), and a high risk class who have high probability of doing everything (8.4%). The 4-class model reveals a classification worth discussing. In the 4-class model, the high risk group has been split into girls who are more likely to be using external substances (i.e. diet pills, laxatives, diuretics, food substances) to lose weight separate from girls who are restricting food intake and vomiting as well as smoking cigarettes to lose weight. This 4-class model did not match the researchers' theory as well as the 3-class model and since the 3-class model empirically fitted the best, it will be used for the latent class regressed on a latent factor model.

Now we explore the observed body satisfaction variables $\mathbf{X}_i$ as measurements of a latent factor f_i. The researchers hypothesize that these questions are measuring one dimension of body satisfaction. The correlations between the variables in $\mathbf{X}_i$ range between 0.57 and 0.75. The eigenvalues of the covariance matrix are (3.703, 0.476, 0.362, 0.261, 0.198), which indicates that 1-dimension is well described by these variables providing empirical support for the 1-factor model. Thus, we will consider the body satisfaction, a 1-dimensional continuous latent factor f_i underlying the observed $\mathbf{X}_i$.

5.2 Model Fitting

Consider the parametric model (7), where $P = 5$, $Q = 1$, $J = 9$ and $K = 2, 3$ for the example data set. Note, although the 3 latent class outcome model was chosen based on exploratory data analysis, the 2 latent class outcome is shown for comparison. Table 1 shows the parameter estimates and their standard errors for different models, where the "low" eating disorders risk class is treated as the reference class 0. The AIC and BIC values indicate the model with 3-class outcome fits the data better. The estimates of the log ORs for class 1 (β_{11}) and class 2 (β_{12}) are negative and statistically significant, which are interpreted as the effect of a 1-unit increase in body satisfaction on the log odds of being in class $k (k = 1, 2)$ rather than class 0. It makes sense that these are negative since as a girl's satisfaction with body increases, she would be less likely to be in one of the high eating disorders risk classes. The intercepts represent the log odds of being in class k rather than class 0 for a girl with body satisfaction at the center of the scale, i.e. 0, since the elements of $\mathbf{X}_i$ have been centered. These are related to the overall prevalence in each class where we see that the prevalence in class 2, the high risk class, is small.

Table 1: Estimation results for Project EAT data

Model	Para.	Est.	SE	AIC	BIC
2-class	β_{01}	-0.400	0.065	20180	20350
	β_{11}	-1.018	0.064		
3-class	β_{01}	-0.952	0.092	20062	20294
	β_{02}	-1.201	0.108		
	β_{11}	-0.921	0.080		
	β_{12}	-1.147	0.093		

References

Abramowitz, M. and Stegun, I.A. (1972). *Handbook of Mathematical Functions*, New York: Dover Publications, Inc.

Anderson T.W., Amemiya Y. (1998). The asymptotic normal distribution of estimators in factor analysis under general conditions. *The Annals of Statistics*, 16, 759-771.

Bandeen-Roche, K., Miglioretti, D., Zeger, S., Rathouz, P (1997). *JASA*, 92, No. 440,1375-1386.

Bartholomew, D.J., and Knott, M. (1999). Latent Variable Models and Factor Analysis, 2nd. ed., Kendall's Library of Statistics

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York, Wiley.

Booth, J.G. and Hobert, J.H. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society* Series B 62: 265-285.

Browne, M.W., Shapiro A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology*, 41, 193-208.

Burr, D. (1988). On Errors-in-Variables in Binary Regression-Berkson Case. *JASA*, 83,739-743

Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995), *Measurement Error in Nonlinear Models*, Chapman & Hall, London.

Clogg, C.C. (1981). New developments in latent structure analysis. In D.J. Jackson and E.F. Borgotta (eds.) *Factor Analysis and Measurement in Sociological Research*. Beverly Hills: Sage.

Clogg, C.C. (1995). Latent class models. In G. Arminger, C.C. Clogg, and M.E. Sobel (eds.) *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum Press, pp. 311-59.

Collins, L.M., Fidler, P.L., Wugalter, S.E., Long, J. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, 28, 375-389.

Dayton, M., and Macready, G. (1988). Concomitant-variable latent class models. *JASA*, 83, 173-178.

Dunson, D.B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society* Series B, 62, 355-366.

Dunson, D.B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *JASA*, 98, No. 463 555-563.

Fuller, W. (1987). *Measurement Error Models*. New York: John Wiley and Son.

Jöreskog, K.G. and Sorborm, D. (1996). *LISREL8: User's Reference Guide*, Chicago: Scientific Software International.

Hagenaars, J.A. and McCutcheon, A.L. (eds.) (2002) *Applied Latent Class Analysis Models*. Cambridge University Press.

Golub, G.H., and Welsch, J.H. (1969). Calculation of Gaussian Quadrature Rules. *Mathematical Computing*, 23, 221 -230.

Lawley, D.N., and Maxwell, A.E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. London: Butterworths.

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

Louis, T. (1982). Finding the observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society*. Series B, Vol.44, No.2.

McLachlan, G.J., and Peel, G. (2000). *Finite Mixture Models*. New York, Wiley.

Moustaki, I. and Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65, 391-411.

Neumark-Sztainer, D., Story, M., Hannan, P., and Moe, J. (2002). Overweight status and eating patterns among adolescents: Where do youth stand in comparison to the Healthy People 2010 Objectives? *Am J Public Health*, 92: 844-51.

Sammel, M., Ryan, L., and Legler, J. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society* Series B, 59, 667-678.

Wei, G.C.G. and Tanner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *JASA*, 85, 699-704.