Biostatistics (2006), **7**, 1, *pp.* 145–163
doi:10.1093/biostatistics/kxi046
Advance Access publication on August 3, 2005

Latent class regression on latent factors

JIA GUO, MELANIE WALL*

Division of Biostatistics, School of Public Health, University of Minnesota, A460 Mayo Building, MMC 303, Minneapolis, MN 55455-0378, USA
melanie@biostat.umn.edu

YASUO AMEMIYA

IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Route 134, Yorktown Heights, NY 10598, USA

Summary

In the research of public health, psychology, and social sciences, many research questions investigate the relationship between a categorical outcome variable and continuous predictor variables. The focus of this paper is to develop a model to build this relationship when both the categorical outcome and the predictor variables are latent (i.e. not observable directly). This model extends the latent class regression model so that it can include regression on latent predictors. Maximum likelihood estimation is used and two numerical methods for performing it are described: the Monte Carlo expectation and maximization algorithm and Gaussian quadrature followed by quasi-Newton algorithm. A simulation study is carried out to examine the behavior of the model under different scenarios. A data example involving adolescent health is used for demonstration where the latent classes of eating disorders risk are predicted by the latent factor body satisfaction.

Keywords: Factor analysis; Latent class models; Monte Carlo EM.

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.

As a motivating example, we consider a study from behavioral public health that is interested in predictors of eating disorders in adolescent girls. As part of a large comprehensive study of adolescent nutrition and obesity (Project EAT, Neumark-Sztainer *et al.*, 2002), which collected self-report survey data from students in 7th and 10th grade at 31 Twin Cities schools in the 1998–1999 school year, one research question was whether a personal trait related to a girl's body satisfaction could predict her eating disorder risk class. Neither body satisfaction nor eating disorder risk can be measured directly (without error) with a single self-report questionnaire item but both can be considered as latent variables underlying a series of questionnaire items all of which may be measuring the latent variables with error. Body satisfaction

*To whom correspondence should be addressed.

is hypothesized by the researchers to be a continuous latent variable 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, is hypothesized to be a categorical latent variable representing different types of eating disorder risk related to girls engaging in purging vs. those engaging in restriction behaviors. A checklist of nine unhealthy weight control behaviors was asked on the questionnaire. No absolute classification rule based on the checklist of nine behaviors exists, but given a girl's particular (unobserved) eating disorder risk, the researchers would expect certain behaviors to show up more than others. Thus, as hypothesized, the researchers are interested in a regression of a categorical latent variable (eating disorder risk) on a continuous latent variable (body satisfaction) while controlling for other observed covariates.

A long literature exists and is evolving for latent variable models and methods (for a brief history, see Bartholomew and Knott, 1999). Categorical latent variable models, i.e. latent class analysis (Lazarsfeld and Henry, 1968; Clogg, 1995; Hagenaars and McCutcheon, 2002), are common in the health science literature including, e.g. Uebersax and Grove (1990), measuring distinct diagnostic categories given presence/absence of several symptoms; Flaherty (2002), measuring smoking initiation; and Croudace *et al.* (2003), studying typologies for nocturnal enuresis. Continuous latent variable models, i.e. factor analysis or general latent factor (trait) analysis (Lawley and Maxwell, 1971; Moustaki and Knott, 2000), are numerous in health science applications where hypothesized continuous latent factors are used, e.g. Bowling (1997), measuring quality of life; Neumark-Sztainer *et al.* (2003a,b), utilizing social cognitive theory of health behaviors; and Lee *et al.* (2003), measuring attitudes toward drinking alcohol.

But to directly address the research question relating eating disorder risk and body satisfaction where a latent class variable is regressed on a latent factor, we need a model for incorporating both types of latent variables simultaneously and a method for estimating the conditional distribution of the categorical latent class variable given the continuous latent factor. In Section 2, we propose this new model where both the categorical response and continuous predictors are latent variables. In Section 3, 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 (available in SAS PROC NLMIXED). In Section 4, we apply this model to the Project EAT data to analyze the relationship between eating disorder risk class and body satisfaction adjusted for other observed covariates and carry out the model assumption checking. In Section 5, a simulation study examining the behavior of the model with respect to sample size and the measurement reliability for the latent variables is shown, as well as a simulation study mimicking the data example setup to examine computational issues in this realistic situation. Discussion and future work are given in Section 6.

2. LATENT CLASS REGRESSION ON LATENT FACTORS MODEL

Suppose we have a data set with n independent individuals. For individual $i (i = 1, \ldots, n)$, 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 continuous latent variable $\mathbf{f}_i$. Let c_i be a categorical latent 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. The primary model of interest will be the conditional distribution of c_i given $\mathbf{f}_i$ and possibly additional observed covariates $\mathbf{W}_i$, where $\mathbf{W}_i$ represents R-dimensional observed covariates. Figure 1 (model D) provides a diagram representing the proposed model for all the observed and latent variables. Similar in spirit to the latent class regression model (model C in Figure 1) (Dayton and Macready, 1988; Bandeen-Roche *et al.*, 1997), the model introduced in this paper considers latent class regression but the regressors include latent variables as well as observed variables.

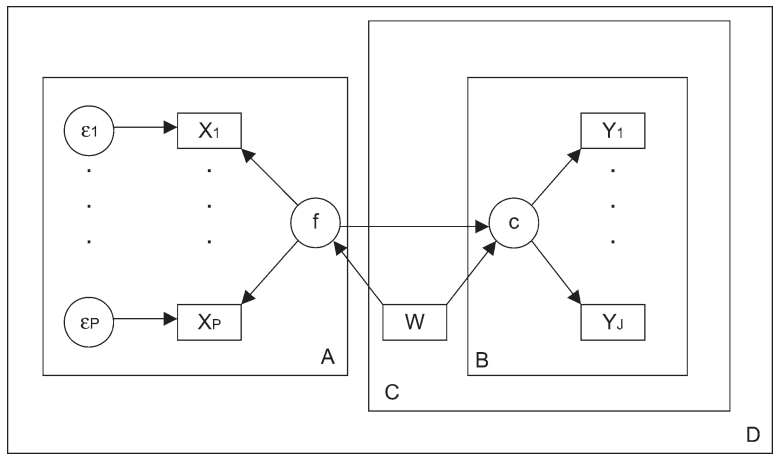

Fig. 1. Model diagram.

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 that $\mathbf{Y}_i$ 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. Furthermore, we assume that $\mathbf{Y}_i$ is conditionally independent of $\mathbf{f}_i$ given c_i and $\mathbf{W}_i$ and likewise, $\mathbf{X}_i$ is conditionally independent of c_i given $\mathbf{f}_i$ and $\mathbf{W}_i$. Finally, we assume that the observed covariates only influence the model through their influence on the latent variables, i.e. we assume that $\mathbf{W}_i$ is conditionally independent of $\mathbf{Y}_i$ and $\mathbf{X}_i$ given c_i and $\mathbf{f}_i$. In theory, this last assumption could be weakened to allow for covariates to directly influence the observed variables rather than indirectly through the latent variables. But generally it is of interest to examine the effect that covariates have on the variables of interest, i.e. the latent variables, rather than their influence on the measurement itself. Hence, we introduce the following model for the joint distribution of the observed data $\mathbf{Y}_i$ and $\mathbf{X}_i$,

$$f(\mathbf{Y}_i, \mathbf{X}_i|\mathbf{W}_i) = \sum_{k=1}^{K} \int f(\mathbf{Y}_i|c_i = k)\, f(\mathbf{X}_i|\mathbf{f}_i)\, f(c_i = k|\mathbf{f}_i, \mathbf{W}_i)\, f(\mathbf{f}_i|\mathbf{W}_i)\, \mathrm{d}\mathbf{f}_i, \tag{2.1}$$

with specific parametric models specified as follows

$$f(\mathbf{Y}_i|c_i = k) = \prod_{j=1}^{J} \pi_{j|k}^{Y_{ij}} (1 - \pi_{j|k})^{1-Y_{ij}}, \tag{2.2}$$

$$f(\mathbf{X}_i|\mathbf{f}_i) \sim N_p\left(\begin{pmatrix} \boldsymbol{\lambda}_0 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Lambda} \\ \mathbf{I} \end{pmatrix} \mathbf{f}_i, \boldsymbol{\Psi}\right), \tag{2.3}$$

$$f(c_i = k|\mathbf{f}_i, \mathbf{W}_i) = \pi_k(\mathbf{f}_i, \mathbf{W}_i) = \frac{\exp(\alpha_k + \mathbf{f}_i{}^T \boldsymbol{\beta}_k + \mathbf{W}_i{}^T \boldsymbol{\gamma}_k)}{\sum_{k=1}^{K} \exp(\alpha_k + \mathbf{f}_i{}^T \boldsymbol{\beta}_k + \mathbf{W}_i{}^T \boldsymbol{\gamma}_k)}, \tag{2.4}$$

$$f(\mathbf{f}_i|\mathbf{W}_i) \sim F(\boldsymbol{\Gamma}\mathbf{W}_i, \boldsymbol{\Phi}). \tag{2.5}$$

A latent class model with conditional independence of measurement within class (Clogg, 1981,1995; McLachlan and Peel, 2000) is assumed for the relationship between $\mathbf{Y}_i$ and c_i (model B in Figure 1)

with $\pi_{j|k} = \Pr(Y_{ij} = 1|c_i = k)$ representing the probability that $Y_{ij} = 1$ when the ith individual is in the latent class k. A confirmatory factor analysis model is used for the relationship between $\mathbf{X}_i$ and $\mathbf{f}_i$ (model A in Figure 1) in the errors-in-variables parameterization (see e.g. Fuller, 1987; Jöreskog and Sorborm, 1996), where $\mathbf{I}$ is a $Q \times Q$ identity matrix, $\mathbf{0}$ is a Q-dimensional vector of zeros, and $\boldsymbol{\lambda}_0$ and $\boldsymbol{\Lambda}$ are known or unknown scalars. We specify $\mathbf{X}_i = (\boldsymbol{\lambda}_0', \mathbf{0}')' + (\boldsymbol{\Lambda}', \mathbf{I})'\mathbf{f}_i + \boldsymbol{\epsilon}_i$, where the random error $\boldsymbol{\epsilon}_i$ is a P-dimensional vector with $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$, $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Psi}$, 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 $\boldsymbol{\epsilon}_i$. Note that like the latent class regression model (model C in Figure 1), we use the generalized logit link for the probabilities of the latent classes, i.e. $\log(\frac{\pi_k(\mathbf{f}_i, \mathbf{W}_i)}{\pi_K(\mathbf{f}_i, \mathbf{W}_i)}) = \alpha_k + \mathbf{f}_i{}^T\boldsymbol{\beta}_k + \mathbf{W}_i{}^T\boldsymbol{\gamma}_k$, where $\alpha_K = 0$, $\boldsymbol{\beta}_K = (0, 0, \ldots, 0)^T$, and $\boldsymbol{\gamma}_K = (0, 0, \ldots, 0)^T$, indicating class K as the reference class. The parameter $\boldsymbol{\beta}_k$ is a Q-dimensional vector and $\boldsymbol{\gamma}_k$ a R-dimensional vector relating the latent factors and observed covariates (respectively) to the odds of being in a particular latent class versus the reference class adjusted for one another. Finally, some Q-dimensional distribution F for latent factors is specified where $\boldsymbol{\Gamma}$ is a $Q \times R$ matrix of scalars such that $E(\mathbf{f}_i|\mathbf{W}_i) = \boldsymbol{\Gamma}\mathbf{W}_i$ and $\boldsymbol{\Phi}$ represents other unknown parameters specifying the distribution.

All estimations of the model (2.1)–(2.5) will be based on the number of factors Q and then the number of classes K being fixed and known. While for many real problems, the Q and K will be considered given based on subject matter theory, it is useful to consider how to choose them based on data or at least to give some guidance. 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 scree plot of the correlation matrix of $\mathbf{X}$. The number of factors can be chosen to be the number of eigenvalues before the elbow in the plot. Similarly, it is common in latent class analysis to fit models with different numbers of classes and compare them by Bayesian information criterion (BIC) values and choose the model with the smallest BIC (Collins *et al.*, 1993). These techniques will be used for the data example in Section 4.

3. MAXIMUM LIKELIHOOD ESTIMATION

Given the parametric model (2.1)–(2.5) 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 $\boldsymbol{\theta} = (\{\pi_{j|k}\}, \boldsymbol{\lambda}_0, \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \{\alpha_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k\}, \boldsymbol{\Gamma}, \boldsymbol{\Phi})$ be the vector of parameters relating $\mathbf{Z}_i$ with $\mathbf{d}_i$ and $\mathbf{W}_i$. Thus, the likelihood function for the model (2.1)–(2.5) can be written as

$$L_o = \prod_{i=1}^{n} f(\mathbf{Z}_i|\mathbf{W}_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \int f(\mathbf{Z}_i, \mathbf{d}_i|\mathbf{W}_i; \boldsymbol{\theta})\, \mathrm{d}\mathbf{d}_i, \tag{3.1}$$

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: MCEM algorithm and Gaussian quadrature followed by quasi-Newton algorithm.

3.1 *MCEM algorithm*

It is natural to consider the latent variables, $\mathbf{d}_i$, as missing data and implement the expectation and maximization (EM) algorithm for maximizing (3.1). Since it is hard to maximize the observed data likelihood L_o directly, we construct the complete data likelihood and apply the EM algorithm to maximize it.

The complete data likelihood is

$$L_{\mathrm{c}} = \prod_{i=1}^{n} f(\mathbf{Z}_i, \mathbf{d}_i | \mathbf{W}_i; \boldsymbol{\theta}).$$

The MCEM algorithm will iterate between the E-step and M-step until the parameter estimates converge according to some criteria. We monitor the convergence of the EM algorithm by plotting $\boldsymbol{\theta}_l$ vs. 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)

$$I_{\mathbf{Z}}(\boldsymbol{\theta}) = E_{\mathbf{d}}\left(-\frac{\partial^2 L_{\mathrm{c}}(\mathbf{Z}, \mathbf{d}|\mathbf{W}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) - \mathrm{Var}_{\mathbf{d}}\left(\frac{\partial L_{\mathrm{c}}(\mathbf{Z}, \mathbf{d}|\mathbf{W}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right),$$

evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$.

Details of the MCEM algorithm are described in Appendix A.

3.2 *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 associated with the latent class regressed on latent factors model. Because the latent classes are discrete, it can be written as

$$L_{\mathrm{o}} = \prod_{i=1}^{n} \sum_{k=1}^{K} \int f(\mathbf{Y}_i | c_i = k) f(\mathbf{X}_i | \mathbf{f}_i) f(c_i = k | \mathbf{f}_i, \mathbf{W}_i) f(\mathbf{f}_i | \mathbf{W}_i) \, \mathrm{d}\mathbf{f}_i.$$

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 approximated 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 B gives the 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 five times longer using MCEM. It should also be noted that for increasing numbers of factors, the integration in both methods may be computationally prohibitive.

4. EXAMPLE

Continuing with the Project EAT data set described in Section 1, for each individual i ($i = 1, \ldots, 1905$), 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. Furthermore, let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i9})^T$ be the nine 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?').

4.1 *Exploratory data analysis*

Assume that underlying the observed responses $\mathbf{Y}_i$ ($i = 1, \ldots, 1905$) 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 nine observed indicators asking which unhealthy weight control behaviors had been used within the past year. Table 1 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 (59.0%), a class doing just the restricting behaviors (i.e. eating very little and skipping meals) (35.1%), and a high-risk class having a high probability of doing everything (5.9%).

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 0.77. The eigenvalues of the correlation matrix are (3.682, 0.485, 0.373, 0.262, 0.197), which indicates that one dimension is well described by these variables, providing empirical support for the one-factor model. Thus, we will consider the body satisfaction, a one-dimensional continuous latent factor f_i underlying the observed $\mathbf{X}_i$.

As a further data exploration, we consider nine separate logistic regressions of the nine different binary outcomes in $\mathbf{Y}$ on the body satisfaction score (i.e. $\hat{\mathbf{f}}_i$ in Section 4.3), obtained from the factor analysis model for $\mathbf{X}$, and three other covariates, age, body mass index (BMI), and social economic status (SES). Note, age was self-reported, BMI was collected by a trained staff member, and SES was measured by self-reported highest parental education level. The parameter estimates and 95% confidence intervals for the nine separate regressions are shown in Table 2. We note that in all cases, the body satisfaction score is significantly negatively related with the unhealthy weight control behavior outcomes but that the

Table 1. *Estimated $\pi_{j|k}$ (probability of saying yes to the variable j given that the individual is in latent class k) under latent class models with different K values*

To control weight	Marginal	2-Class		3-Class			4-Class			
		0	1	0	1	2	0	1	2	3
Fasted	17.7	2.4	39.8	2.2	32.0	65.2	1.8	14.4	67.9	52.8
Ate little	43.5	8.7	93.5	7.3	89.6	97.7	1.3	73.2	100.0	95.8
Diet pills	6.3	1.2	13.6	1.2	7.2	37.9	0.8	6.3	14.3	56.0
Vomit	6.0	0.1	14.5	0.2	6.3	46.3	0.0	3.3	21.3	56.6
Laxatives	1.6	0.2	3.7	0.2	0.0	18.8	0.2	0.0	2.7	39.3
Diuretics	1.2	0.0	2.8	0.0	0.1	11.4	0.0	0.1	0.8	28.9
Food substitutes	8.9	2.1	18.8	2.0	13.5	37.6	1.5	8.9	24.4	49.5
Skipped meals	44.0	11.3	91.0	9.3	89.8	89.4	3.5	73.2	100.0	75.3
Smoked more cigs	9.2	2.5	19.0	2.4	12.8	42.3	2.0	7.9	28.5	41.7
Percent in each class	100	58.9	41.1	59.0	35.1	5.9	47.7	34.0	16.4	1.9
BIC		9843.2		9780.4			9815.6			

Table 2. *Logistic regressions for nine binary outcomes on body satisfaction score, age, BMI, and SES (expected value is the mean of the coefficient estimates from 100 simulated data sets; see section 5.2)*

Response	Predictor	Estimate	95% CI	Expected value
Fasted	Body satisfaction score	−0.5521	(−0.6769, −0.4300)	−0.6318
	Age	0.1634	(0.0847, 0.2437)	0.1314
	BMI	−0.0100	(−0.0402, 0.0193)	0.0109
	SES	−0.0117	(−0.1047, 0.0813)	−0.0248
Ate little	Body satisfaction score	−0.6873	(−0.7901, −0.5868)	−0.7597
	Age	0.1110	(0.0491, 0.1734)	0.1011
	BMI	0.0443	(0.0195, 0.0694)	0.0490
	SES	−0.1631	(−0.2385, −0.0881)	−0.1520
Diet pills	Body satisfaction score	−0.7411	(−0.9480, −0.5423)	−0.5630
	Age	0.1322	(0.0082, 0.2607)	0.1189
	BMI	0.0420	(0.0006, 0.0813)	0.0099
	SES	−0.0773	(−0.2267, 0.0710)	−0.0193
Vomit	Body satisfaction score	−0.9045	(−1.1231, −0.6954)	−0.7275
	Age	0.0205	(−0.1010, 0.1451)	0.1921†
	BMI	−0.0179	(−0.0657, 0.0269)	−0.0048
	SES	−0.0017	(−0.1514, 0.1480)	0.0303
Laxative	Body satisfaction score	−0.6543	(−1.0500, −0.2876)	−0.7519
	Age	0.1601	(−0.0739, 0.4124)	0.2277
	BMI	0.0587	(−0.0154, 0.1244)	−0.0184†
	SES	−0.1524	(−0.4440, 0.1301)	0.0923
Diuretics	Body satisfaction score	−0.4932	(−0.9254, −0.0831)	−0.5908
	Age	−0.0697	(−0.3265, 0.1971)	0.1691
	BMI	−0.0029	(−0.1097, 0.0868)	−0.0149
	SES	−0.2052	(−0.5433, 0.1191)	0.0124
Food substitute	Body satisfaction score	−0.5988	(−0.7682, −0.4339)	−0.5257
	Age	0.0250	(−0.0761, 0.1281)	0.1057
	BMI	0.0182	(−0.0199, 0.0545)	0.0091
	SES	0.1798	(0.0554, 0.3060)	−0.0191†
Skipped meals	Body satisfaction score	−0.7733	(−0.8791, −0.6702)	−0.7360
	Age	0.0938	(0.0314, 0.1566)	0.0966
	BMI	0.0405	(0.0154, 0.0658)	0.0545
	SES	−0.1307	(−0.2067,−0.0551)	−0.1602
Smoked	Body satisfaction score	−0.6413	(−0.8095, −0.4781)	−0.5279
	Age	0.2764	(0.1670, 0.3901)	0.1300†
	BMI	−0.0224	(−0.0629, 0.0160)	0.0034
	SES	−0.0868	(−0.2110, 0.0367)	−0.0001

†The expected values are out of the 95% CIs of the observed coefficients.

other covariates vary in their significance with different outcomes. While the values in Table 2 provide a means for examining the association of each type of behavior with some measure related to body satisfaction (despite the measurement error in the body satisfaction score not being taken into account), the research question of interest was not related to any one particular unhealthy weight control behavior. On the contrary, the research question was how ‘groupings’ or ‘clusters’ of these behaviors were related to

body satisfaction (and other covariates). These 'groupings' will thus be modeled via a latent class model as described above and regressed on a latent factor model for the body satisfaction variables along with other covariates.

4.2 *Model fitting*

For many latent variable models, the log likelihood is relatively flat and may have more than one local maximum (McHugh, 1956; Habermann, 1977). Choosing good starting values for the parameters in model (2.1)–(2.5) is important. Assuming $\mathbf{f}_i$ is normally distributed as $N(\mathbf{0}, \boldsymbol{\Phi})$, the starting values for $\boldsymbol{\lambda}_0$, $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ are chosen as their estimates from the latent factor model for $\mathbf{X}_i$. For $\{\pi_{j|k}\}$, we use their estimates from the latent class model for $\mathbf{Y}_i$ as the starting values. For α_k, $\boldsymbol{\beta}_k$, and $\boldsymbol{\gamma}_k$ values, we obtain the predicted values $\hat{c}_i$ and $\hat{\mathbf{f}}_i$ for c_i and $\mathbf{f}_i$ (explained below) from the two measurement models, respectively, and then fit the generalized logit model for $\hat{c}_i$ on $\hat{\mathbf{f}}_i$ and $\mathbf{W}_i$ to obtain the parameter estimates. They are used as starting values for α_k, $\boldsymbol{\beta}_k$, and $\boldsymbol{\gamma}_k$ values. Under the normality assumptions of the distributions of $\mathbf{X}_i|\mathbf{f}_i$ and $\mathbf{f}_i$, the best unbiased predictor of $\mathbf{f}_i$ is $E(\mathbf{f}_i|\mathbf{X}_i) = \boldsymbol{\Phi}\begin{pmatrix}\boldsymbol{\Lambda}\\ \mathbf{I}\end{pmatrix}^T\left(\begin{pmatrix}\boldsymbol{\Lambda}\\ \mathbf{I}\end{pmatrix}\boldsymbol{\Phi}\begin{pmatrix}\boldsymbol{\Lambda}\\ \mathbf{I}\end{pmatrix}^T + \boldsymbol{\Psi}\right)^{-1}\left(\mathbf{X}_i - \begin{pmatrix}\boldsymbol{\lambda}_0\\ \mathbf{0}\end{pmatrix}\right)$. We plug the parameter estimates $\hat{\boldsymbol{\lambda}}_0$, $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Phi}}$, and $\hat{\boldsymbol{\Psi}}$, which are obtained from the latent factor model, into this conditional expectation to get its estimate, i.e. $\hat{\mathbf{f}}_i$. Similarly, from the latent class model, we can estimate the probability that subject i is in one latent class or another. The posterior probability of subject i belonging to latent class k is $\Pr(c_i = k|\mathbf{Y}_i) = \frac{\pi_k \prod_{j=1}^{J} \pi_{j|k}^{Y_{ij}} (1-\pi_{j|k})^{1-Y_{ij}}}{f(\mathbf{Y}_i)}$ by Bayes theorem. The parameter estimates from the latent class model can be plugged in and we can classify subject i to a latent class for which the posterior probability is greatest, i.e. $\hat{c}_i = \{k\text{: } \max(\Pr(c_i = k|\mathbf{Y}_i), k = 1, \ldots, K)\}$.

Consider the parametric model (2.1)–(2.5), where $P = 5$, $Q = 1$, $J = 9$, and $K = 2, 3, 4$ for the example data set. Note that although the 3-latent class outcome model was chosen based on exploratory data analysis, the 2- and 4-class cases are shown as comparison for the full latent class regression on latent factor model. Covariates included in the model and treated to be measured without error, i.e. the $\mathbf{W}_i$, were age, BMI, and SES. Table 3 shows the parameter estimates for different models, the standard errors of estimates, and the *P*-values for each parameter where the 'low' eating disorders risk class is treated as the reference class 0. The BIC values indicate that the model with 3-class outcome fits the data better than the others.

The parameter estimates for $\{\pi_{j|k}\}$ (not shown) are very similar to the corresponding estimates from the simple latent class model (Table 1). This similarity is due to the separation in model (2.1)–(2.5) between the parameters in the measurement models from those in the structural or 'regression' part of the model. Furthermore, the estimates for the factor loadings range between 0.8921 and 1.0356, and the estimated variance of body satisfaction is 1.2691, which are similar to the results of just fitting the one-factor model to the $\mathbf{X}$. The intercepts (α_k's) represent the log odds of being in class k rather than class 0 for a girl with body satisfaction, age, BMI, and SES at 0. The estimates of the log ORs for class 1 ($\beta_1 = -0.7442$) and class 2 ($\beta_2 = -1.5400$) 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 adjusted for the other covariates. 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. In addition, it is found that as age increases there is a significantly higher likelihood to be in the eating disorder risk class 2 rather than risk class 0, whereas age did not distinguish class 1 from class 0; higher BMI girls and lower SES girls are significantly more likely to be in class 1 but not more likely to be in class 2 than class 0.

Additionally, for the 3-class situation, tests of the contrasts for the slopes between classes 1 and 2 show that there is significant difference between the two classes, which indicates that body satisfaction, age, BMI, and SES provide enough information to distinguish between the two classes. For the 4-class

Table 3. *Estimation results for Project EAT data*

Model	Class	Parameter	Estimate	Standard error	*P*-value	Contrast	*P*-value
2-Class,	Class 1	Intercept α_1	−1.5993	0.4048	<0.0001		
BIC: 18116		Body satisfaction β_1	−0.9685	0.0689	<0.0001		
		Age γ_{11}	0.1395	0.0377	0.0002		
		BMI γ_{21}	0.0477	0.0159	0.0028		
		SES γ_{31}	−0.1779	0.0454	<0.0001		
3-Class,	Class 1	Intercept α_1	−1.6389	0.4608	0.0004	Class 1 vs. 2	
BIC: 18044		Body satisfaction β_1	−0.7442	0.0836	<0.0001	β_1 vs. β_2	<0.0001
		Age γ_{11}	0.0389	0.0456	0.3933	γ_{11} vs. γ_{12}	0.0003
		BMI γ_{21}	0.0900	0.0194	<0.0001	γ_{21} vs. γ_{22}	0.0407
		SES γ_{31}	−0.2828	0.0538	<0.0001	γ_{31} vs. γ_{32}	0.0041
	Class 2	Intercept α_2	−1.7009	0.7279	<0.0001		
		Body satisfaction β_2	−1.5400	0.1378	<0.0001		
		Age γ_{12}	0.3299	0.0714	<0.0001		
		BMI γ_{22}	0.0375	0.0254	0.1395		
		SES γ_{32}	−0.0382	0.0785	0.6270		
4-Class,	Class 1	Intercept α_1	−1.7463	0.5250	0.0009		
BIC: 18056		Body satisfaction β_1	−0.6952	0.1019	<0.0001		
		Age γ_{11}	−0.0218	0.0541	0.6987		
		BMI γ_{21}	0.1134	0.0245	<0.0001		
		SES γ_{31}	−0.3252	0.0644	<0.0001		
	Class 2	Intercept α_2	−3.3758	0.8353	<0.0001	Class 2 vs. 3	
		Body satisfaction β_2	−1.3453	0.1850	<0.0001	β_2 vs. β_3	0.4756
		Age γ_{12}	0.3723	0.0900	<0.0001	γ_{12} vs. γ_{13}	0.2200
		BMI γ_{22}	0.03616	0.0304	0.2349	γ_{22} vs. γ_{23}	0.4559
		SES γ_{32}	−0.0550	0.0904	0.5425	γ_{32} vs. γ_{33}	0.2627
	Class 3	Intercept α_3	−4.9110	1.3661	0.0003		
		Body satisfaction β_3	−1.6891	0.3653	<.0001		
		Age γ_{13}	0.0929	0.1703	0.5857		
		BMI γ_{23}	0.0899	0.0555	0.1044		
		SES γ_{33}	−0.3004	0.1761	0.0882		

case, the contrasts between classes 2 and 3 indicate that all the four predictors cannot distinguish the two classes (Table 3) which provides evidence against the 4-class model in addition to the worse BIC.

4.3 *Model checking*

In the development of model (2.1), two implicit assumptions are

$$f(\mathbf{Y}_i|c_i, \mathbf{f}_i, \mathbf{W}_i) = f(\mathbf{Y}_i|c_i), \tag{4.1}$$

$$f(\mathbf{X}_i|c_i, \mathbf{f}_i, \mathbf{W}_i) = f(\mathbf{X}_i|\mathbf{f}_i). \tag{4.2}$$

These assumptions can be referred to as nondifferential measurement assumptions (Carroll *et al.*, 1995). Diagnostics for assessing nondifferential measurement have been proposed in situations where the violation is caused by an observed variable, but here our assumptions involve things we cannot observe. So, we

propose checking the assumption (4.1) where $\hat{\mathbf{f}}_i$ replaces $\mathbf{f}_i$ and likewise checking assumption (4.2) where $\hat{c}_i$ replaces c_i.

In order to check assumption (4.1) for the Project EAT data with the model including three latent classes regressed one latent factor and covariates, we follow the idea of the diagnostic approach for latent class regression model proposed by Bandeen-Roche *et al.* (1997). First, we fit the latent class regression model for $\mathbf{Y}_i$ on $\hat{\mathbf{f}}_i$ and $\mathbf{W}_i$ to get the estimated probabilities $(\hat{\theta}_{i1}, \hat{\theta}_{i2}, \hat{\theta}_{i3})$ of being in each of the three classes and randomly assign subject i to a class with probabilities $(\hat{\theta}_{i1}, \hat{\theta}_{i2}, \hat{\theta}_{i3})$. Within each estimated class, we summarize the 2^9 possible observed patterns describing unhealthy weight control behaviors self-report pattern. Because of sparseness of some patterns in the observed data, we merge some patterns together to construct new nominal response variables. Then we regress the resulting polytomous response patterns on the factor score and the observed covariates using a generalized logit link, separately for each class. If assumption (4.1) holds, we would expect no effect of the body satisfaction score, age, BMI, and SES on the new nominal variable within each class.

For assumption (4.2), it can be shown that the assumption can be written in terms of the measurement errors in the factor analysis model, i.e. $f(\boldsymbol{\epsilon}_i|c_i, \mathbf{f}_i, \mathbf{W}_i) = f(\boldsymbol{\epsilon}_i|\mathbf{f}_i)$. Furthermore, under the assumption that $\boldsymbol{\epsilon}_i$ is independent of $\mathbf{f}_i$ in the factor analysis model, it can be simplified as $f(\boldsymbol{\epsilon}_i|c_i, \mathbf{W}_i) = f(\boldsymbol{\epsilon}_i)$. So, checking assumption (4.2) is equivalent to checking whether the measurement errors in the latent factor analysis model differ depending on the latent classes and the observed covariates (age, BMI, and SES). To check this, the residuals from the factor analysis model $\hat{\boldsymbol{\epsilon}}_i = \mathbf{X}_i - \left(\hat{\boldsymbol{\lambda}}_{\mathbf{0}}\right) - \left(\hat{\underset{\mathbf{I}}{\boldsymbol{\Lambda}}}\right)\hat{\mathbf{f}}_i$ and $\hat{c}_i$ are used. We fit the multivariate linear regression model for $\hat{\boldsymbol{\epsilon}}_i$ on $\hat{c}_i$ and $\mathbf{W}_i$. If assumption (4.2) holds, we would expect no effect of the eating disorder risk class and the three observed covariates on the residuals from body satisfaction factor model.

Table 4 presents the results for checking assumption (4.1). We merge the patterns in each estimated class according to the number of unhealthy weight control behaviors and create a new polytomous variable, i.e. response group. In each estimated class, one response group is set as the reference group and the log ORs of the body satisfaction score, age, BMI, and SES are fixed as 0; we find that for the other two response groups, almost all log OR estimates obtained from the generalized logit model are not significantly different from 0, which implies that there is no significant effect of body satisfaction score, age, BMI, and SES on unhealthy weight control behavior patterns given a particular class. For assumption (4.2), the F tests (Wilks lambda) for each covariate ($\hat{c}$, age, BMI, and SES) are significant ($P < 0.001$) for testing whether these four covariates affect the multivariate residuals. Thus, there is evidence that assumption (4.2) may be violated.

Table 4. *Model assumption (4.1) checking for Project EAT data*

Class	Response group	Group size	Factor score	*P*-value	Age	*P*-value	BMI	*P*-value	SES	*P*-value
0	0 Behavior	829	0		0		0		0	
	1 Behavior	113	−0.4836	0.1609	0.2034	0.2611	−0.1255	0.1832	−0.1229	0.5582
	2–3 Behaviors	13	0.0054	0.9664	0.0947	0.1334	0.0537	0.0352	−0.0425	0.5738
1	0–1 Behavior	172	0		0		0		0	
	2 Behaviors	303	−0.0063	0.9585	0.1274	0.0518	−0.0229	0.3268	0.0643	0.4397
	3–5 Behaviors	192	−0.0521	0.6367	0.0160	0.7841	−0.0309	0.1459	0.0803	0.2874
2	2–3 Behaviors	108	0		0		0		0	
	4 Behaviors	94	0.0624	0.7420	−0.0572	0.6017	0.0091	0.8516	0.0145	0.9017
	5–9 Behaviors	81	0.2076	0.2510	−0.1133	0.2807	0.0557	0.1166	0.0660	0.5610

5. Simulation studies

In this section, we present two simulation studies. The first is designed to examine the behavior of the estimators given data generated from model (2.1)–(2.5) under scenarios with different amounts of measurement error, that is, with differing strength of relationship between the observed and latent variables as well as different sample sizes. The second generates data mimicking the motivating data example and compares results from the 'true' model with those obtained from the observed data. In addition, computational issues are examined comparing the original and a much smaller sample size.

5.1 *Simulation examining influence of reliability*

We first investigate the effect that sample size and reliability of the measurement for both $\mathbf{f}_i$ and c_i have on the inference for the structural parameters in model (2.1)–(2.5). Reliability refers to the amount of measurement error there is in the latent class model and latent factor model. Consider the model with $P = 4$ dimensional observed $\mathbf{X}_i$ variable measuring $Q = 1$ latent factor and $J = 5$ dimensional observed $\mathbf{Y}_i$ variable measuring $K = 2$ latent classes. Furthermore, assume that the underlying factor f_i is normally distributed. Specifically, consider

$$
\begin{gathered}
f_i \overset{\text{i.i.d.}}{\sim} N(0,1), \\
\boldsymbol{\epsilon}_{\mathbf{i}} \overset{\text{i.i.d.}}{\sim} N_4(0, 0.5 I_{4\times 4}), \\
\begin{pmatrix} X_{i1} \\ X_{i2} \\ X_{i3} \\ X_{i4} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{14} \end{pmatrix} f_i + \boldsymbol{\epsilon}_{\mathbf{i}}, \\
c_i \,|\, f_i \sim \text{Bernoulli}(\text{logit}^{-1}(0.5 + 1 \times f_i)), \\
Y_{ij} | c_i \sim \text{Bernoulli}(\pi_{j|k}), \quad k = 1, 2, \quad j = 1, \ldots, 5.
\end{gathered}
\tag{5.1}
$$

Consider two different sets of factor loadings for the latent factor. First, we generate the data with $(\lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{14})^T = (0.4, 0.4, 0.5, 0.6)^T$, which implies that each of the four variables in X_i has low reliability (<0.6) for measuring f_i and then we choose $(\lambda_{11}, \lambda_{12}, \lambda_{13}, \lambda_{14})^T = (1.5, 1.6, 1.0, 1.7)^T$ so that each has high reliability (>0.9) for measuring f_i. Furthermore, we consider two different sets of values for $\pi_{j|1}$ and $\pi_{j|2}$, where $j = 1, \ldots, 5$, which describe the way that the observed data Y_{ij} are related to the latent classes. One case considers where $\pi_{j|1} = 0.1$ and $\pi_{j|2} = 0.8$ for all j. This implies that when a person is in class 1, he has a low probability, i.e. 0.1, of responding 'yes' to each of the five questions, and when being in class 2, he has a high probability, i.e. 0.8, of responding 'yes' to each of the five questions. We refer to this as the parallel probabilities case since the probabilities for each response Y_{ij} in two latent classes are the same with big differences. The other case we consider is where three of the five observed variables have $\pi_{j|1} = 0.1$ and $\pi_{j|2} = 0.8$ ($j = 1, 2, 3$) but the other two variables (i.e. $j = 4, 5$) have $\pi_{j|1} = 0.2$ and $\pi_{j|2} = 0.3$. Note that the probabilities of being 1 for the three corresponding elements of $\mathbf{Y}_i$ are quite different (0.1 vs. 0.8) and the other two are similar (0.2 vs. 0.3). We refer to this as the non-parallel case and would expect it to be a less precise measurement model for c_i. Finally, we consider two sample sizes, $n = 200$ and 2000. So, in total we have $2 \times 2 \times 2 = 8$ different scenarios. For each scenario, we generate 1000 data sets; to each of the simulated data sets in each of the eight different scenarios, we perform maximum likelihood as described in Section 3. Specifically, because the factor is normally distributed, the adaptive Gauss quadrature method for approximating the observed data likelihood and quasi-Newton optimization as the optimization technique is implemented via PROC NLMIXED in SAS.

Table 5. *Simulation results*

	$\{\pi_{j\|1}\}$ $\{\pi_{j\|2}\}$	Sample size	True value	Reliability for measuring f					
				Low (<0.6)			High (>0.9)		
				Mean	Standard error	Coverage probability	Mean	Standard error	Coverage probability
α	Parallel	200	0.5	0.5103	0.1897	0.9560	0.5067	0.1826	0.9570
		2000	0.5	0.5005	0.0589	0.9390	0.5002	0.0571	0.9370
	Nonparallel	200	0.5	0.5186	0.2559	0.9310	0.5129	0.2482	0.9290
		2000	0.5	0.5002	0.0746	0.9540	0.4998	0.0720	0.9480
β	Parallel	200	1	1.0360	0.2820	0.9450	1.0186	0.2191	0.9510
		2000	1	1.0028	0.0823	0.9520	1.0016	0.0666	0.9430
	Nonparallel	200	1	1.0482	0.3148	0.9420	1.0296	0.2491	0.9530
		2000	1	1.0033	0.0867	0.9570	1.0019	0.0705	0.9530

The simulation results are shown in Table 5. From this table, we find that as expected when the sample size increases from 200 to 2000, the bias and the standard errors are decreasing. We see that in all scenarios for both α (the intercept) and β (the slope), the high reliability for f case shows smaller standard errors than the respective low reliability case. Likewise, since the parallel case can be considered more precise than the nonparallel case, we find smaller standard errors for both α and β for the parallel case as compared to the nonparallel case. What is more interesting is the relative impact of each of these has on the estimates of α and β. The impact that the reliability for f has on the efficiency of β is greater than the impact it has on α. On the other hand, the impact that the parallel vs. nonparallel has on the efficiency of α is much more substantial than that on β. This suggests that the slope of the relationship between c and f is more sensitive to the model for f, whereas the intercept is more sensitive to the model for c. The coverage probability in Table 5 is the number of 95% confidence intervals obtained for each data set based on estimates and standard error estimates from PROC NLMIXED that covered the true value over 1000. In all scenarios, the coverage probabilities are close to 95%.

5.2 *Simulation mimicking Project EAT data*

While the 3-class model converged and provided estimates and standard error for the Project EAT data, it is of interest to examine for possible computational difficulties which may arise in this realistic setup. A simulation study is carried out where 100 data sets are simulated with 1905 observations per data set from the 3-class estimated model, in which the age, BMI, and SES are kept the same as in the data example. We fit the latent class regression on latent factors model with true number of classes and factors to each data set using PROC NLMIXED. Similarly, we also do the simulation study for 100 data sets with 200 observations per set. Table 6 shows the simulation results. From these results, as expected, we find the bias of the parameter estimates and the standard errors decrease when the sample size increases. For the scenario with 1905 observations per data set, the structural parameters (α, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ values) estimates are quite stable with the mean of the estimates and the standard errors very close to the corresponding ones estimated from the data example. But for the smaller sample size scenario, the estimates were not as well behaved. For 11 data sets, the standard error of the estimates was not provided by NLMIXED due to most $\hat{\pi}_{j|k}$ values on the boundary. The CPU time for fitting one data set of size $n = 1905$ was between 24 and 30 min, while for the small data set ($n = 200$) the CPU time was between 4 and 7 min.

Table 6. *Simulation results from Project EAT estimated model*

Parameter	True value	Sample size	Mean	Standard error
α_1	−1.6388	200	−1.8092	1.3602
		1905	−1.6179	0.4194
β_1	−0.7442	200	−0.8331	0.2698
		1905	−0.7535	0.0778
γ_{11}	0.0389	200	0.0589	0.1497
		1905	0.0412	0.0415
γ_{21}	0.0900	200	0.0953	0.0633
		1905	0.0895	0.0159
γ_{31}	−0.2828	200	−0.3626	0.2010
		1905	−0.2912	0.0512
α_2	−4.0254	200	−4.6713	2.6490
		1905	−3.9036	0.7324
β_2	−1.5400	200	−1.8328	0.5262
		1905	−1.5631	0.1323
γ_{12}	0.3299	200	0.4307	0.2915
		1905	0.3428	0.0706
γ_{22}	0.0376	200	0.0271	0.1049
		1905	0.0286	0.0266
γ_{32}	−0.0382	200	−0.0771	0.3028
		1905	−0.0463	0.0777

Using the simulated data with $n = 1905$ per set, we also fit the nine logistic regressions for the nine simulated outcome behaviors on the body satisfaction score, age, BMI, and SES, similar to what was done in the exploratory analysis in Section 4.1. The expected values of regression coefficients are estimated by the average of the estimates over 100 data sets, which are shown in the fourth column of Table 2. All nine expected values of coefficients associated with the body satisfaction are in the 95% confidence intervals of the observed coefficients, which suggests that the estimated latent class regressed on latent factor model reasonably captures relationships in the data between the individual unhealthy weight control variables and the individual body satisfaction variables. Furthermore, we notice that some expected values associated with the observed covariates (age, BMI, and SES) are out of corresponding 95% confidence intervals. This suggests that the model may not be adequately capturing a few relationships between the observed covariates and the individual unhealthy weight control behavior or body satisfaction variables perhaps due to differential measurement.

6. Discussion

This paper proposes a new model for fitting the relationship between a latent class outcome and latent factor predictors. It is demonstrated that maximum likelihood estimation is possible by the MCEM algorithm, or in the case where the factors are normally distributed and Q is small by Gaussian quadrature and quasi-Newton. The model presented in this paper is a natural extension of the latent class regression model and more generally is an extension of structural equation modeling. Structural equation modeling focuses on the relationships among latent variables, but almost exclusively continuous latent variables. Here, we present a method for examining the relationship among latent variables that are of mixed types.

Previous models including both continuous and categorical latent variables have been considered in the literature. Muthen and Shedden (1999) considered a model where continuous latent factors are the random coefficients in a growth model and are regressed on latent classes underlying a different set of observed variables. Both their model and the one proposed in (2.1)–(2.5) can be considered subclasses of a more general model

$$f(\mathbf{Y}_i, \mathbf{X}_i|\mathbf{W}_i) = \sum_{k=1}^{K} \int f(\mathbf{Y}_i|c_i = k)\, f(\mathbf{X}_i|\mathbf{f}_i)\, f(c_i = k, \mathbf{f}_i|\mathbf{W}_i)\, \mathrm{d}\mathbf{f}_i. \tag{6.1}$$

By modeling $f(c_i, \mathbf{f}_i|\mathbf{W}_i) \equiv f(\mathbf{f}_i|c_i, \mathbf{W}_i)\, f(c_i|\mathbf{W}_i)$ with a normal mixture, the model becomes (as introduced in Muthen and Shedden, 1999) an extension of the Gaussian finite mixture model. By modeling the joint distribution $f(c_i, \mathbf{f}_i|\mathbf{W}_i) \equiv f(c_i|\mathbf{f}_i, \mathbf{W}_i)\, f(\mathbf{f}_i|\mathbf{W}_i)$ with a generalized logit for $f(c_i|\mathbf{f}_i, \mathbf{W}_i)$ and some parametric distribution for $f(\mathbf{f}_i|\mathbf{W}_i)$, we obtain the latent class regression on latent factor model presented in this paper. Note that graphically the difference in the two models amounts to changing the direction of the arrow between c and $\mathbf{f}$ in Figure 1. By contrast, the difference in computing the maximum likelihood estimates for the two models is substantial. Recall, in Section 3, that as a result of conditioning on the continuous latent factor in model (2.1)–(2.5), it is necessary to integrate with respect to that factor in the likelihood and that this requires some form of numerical integration. On the other hand, the Gaussian finite mixture model which results from focusing on the conditional distribution in the other direction, i.e. $f(\mathbf{f}_i|c_i, \mathbf{W}_i)$, has a straightforward closed-form (no integration) likelihood. Which model to choose will depend on what the research question is interested in as an outcome vs. predictor.

Similar to a common practice in structural equation modeling, we recommend that the latent class regressed on latent factor model be built in steps. In particular, the measurement models (i.e. the latent class and the latent factor models) can be examined first to assess the fit of different numbers of classes or factors. Then once these measurement models are settled upon, the 'structural model', i.e. the relationship among the latent variables, can be modeled simultaneously with the measurement models. This is how we presented the method for the Project EAT example and feel this allows for appropriately focused model checking. Additionally, a diagnostic approach was presented to check the nondifferential measurement assumptions in the model fitted for Project EAT data. Although this approach is straightforward to carry out, it has a weakness in that surrogates $\hat{\mathbf{f}}_i$ and $\hat{c}_i$ are used and these results may not represent a check for assumptions about $\mathbf{f}_i$ and c_i. Further work is needed to investigate the sensitivity and specificity of this diagnostic approach.

In the simulation study and the example, we only considered models where the underlying latent factor was assumed to be normally distributed. Although other distributions could be considered, it is not clear whether this modeling choice can be checked as the latent factors are not directly observable. Rather, methods that only require weak assumptions on the distribution of the underlying factors should be developed in the future. One might also consider dropping the idea of underlying latent factors and instead regressing the latent class on all the individual observed variables in $\mathbf{X}$. Certainly, this is possible to do, but in practice, when a variety of indicators are chosen to measure one latent variable, they are expected to be highly correlated. In this case, most of them would not be statistically significant in the latent class regression model on these indicators. The advantage of considering a latent variable as a predictor over the observable indicators in modeling has been investigated (Wall and Li, 2003).

Finally, identifiability of model parameters is always an important issue in latent variable models. In our proposed model, the parameters are well identified theoretically when the latent class model and the latent factor model satisfy appropriate conditions. But identifiability is not a property of just the model but also of the data combined with the model. A model may have weak data identifiability if perhaps due to small sample size, some parameters are difficult or even impossible to estimate as was seen in the simulation study when the model from the Project EAT example was fit to only 200 observations and in several

cases the estimates went to the boundary. While generally it is hard to assess this data identifiability problem, some work has been done to quantify it for the latent class model within a fully Bayesian framework by comparing the prior and posterior of parameters (Garrett and Zeger, 2000). Similar methods might be possible to consider for the model presented here.

APPENDIX A

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

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

For the latent class regression on latent factor model (2.1)–(2.5), unfortunately we do not have a closed form for $f(\mathbf{d}_i|\mathbf{Z}_i,\mathbf{W}_i;\boldsymbol{\theta}_l)$ and consequently we do not have a closed-form solution for the integral in $g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta};\mathbf{Z},\mathbf{W})$. Hence, we propose to use the Monte Carlo method to obtain an approximation to $g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta};\mathbf{Z},\mathbf{W})$. First, note that

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

where the expectation is taken with respect to the random variable $\mathbf{d}_i$. Given the current $\boldsymbol{\theta}_l$, a Monte Carlo sample $(\mathbf{d}_i^1,\ldots,\mathbf{d}_i^M)$ is generated from $f(\mathbf{d}_i|\mathbf{W}_i;\boldsymbol{\theta}_l)$, then the expectation can be approximated by an average

$$g_{\boldsymbol{\theta}_l}(\boldsymbol{\theta};\mathbf{Z},\mathbf{W})\approx\sum_{i=1}^{n}\tfrac{1}{M}\sum_{m=1}^{M}[\log f(\mathbf{Z}_i,\mathbf{d}_i^m|\mathbf{W}_i;\boldsymbol{\theta})H_i^m]\equiv g^{\mathrm{MC}}_{\boldsymbol{\theta}_l}(\boldsymbol{\theta};\mathbf{Z},\mathbf{W}),$$

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

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 $\log f(\mathbf{Z}_i,\mathbf{d}_i^m|\mathbf{W}_i;\boldsymbol{\theta})$ can be factorized into four parts corresponding to the four parts of the model (2.2)–(2.5), i.e.

$$\begin{aligned}
\log f(\mathbf{Z}_i,\mathbf{d}_i^m|\mathbf{W}_i;\boldsymbol{\theta}) = {}& \log f(\mathbf{Y}_i|c_i^m;\{\pi_{j|k}\})+\log f(\mathbf{X}_i|\mathbf{f}_i^m;\boldsymbol{\lambda}_0,\boldsymbol{\Lambda},\boldsymbol{\Psi})\\
&+\log f(c_i^m|\mathbf{f}_i^m,\mathbf{W}_i;\{\alpha_k\},\{\boldsymbol{\beta}_k\},\{\boldsymbol{\gamma}_k\})+\log f(\mathbf{f}_i^m|\mathbf{W}_i;\boldsymbol{\Gamma},\boldsymbol{\Phi}).
\end{aligned}\tag{A.1}$$

The M-step is to maximize $g^{\mathrm{MC}}_{\boldsymbol{\theta}_l}(\boldsymbol{\theta};\mathbf{Z},\mathbf{W})$ with respect to $\boldsymbol{\theta}$ and then update $\boldsymbol{\theta}_l$. Based on (A.1), we see that each of the four parts has distinct parameters associated with it. Hence, we can maximize each component separately as a straightforward weighted regression (weighted by H_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, although 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 the 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 lth iteration and M_0 and T are positive constants. We monitor the convergence of the EM algorithm by plotting $\boldsymbol{\theta}_l$ vs. 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)

$$I_{\mathbf{Z}}(\boldsymbol{\theta}) = E_{\mathbf{d}}\left(-\frac{\partial^2 L_{\mathrm{c}}(\mathbf{Z}, \mathbf{d}|\mathbf{W}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) - \mathrm{Var}_{\mathbf{d}}\left(\frac{\partial L_{\mathrm{c}}(\mathbf{Z}, \mathbf{d}|\mathbf{W}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right),$$

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.

APPENDIX B

```
**** Data (dat) is generated as in simulation study from model 5.1;
**** Example of data frame;
id      x1           x2            x3            x4     y1 y2 y3 y4 y5  dummy
1 -0.41747699  -0.103693622  -0.596811257  -0.827308222  0  1  1  1  1    1
2 -1.113301965 -0.098232232   0.2814094176 -0.577613842  0  0  1  1  0    1
3 -0.395122307  0.2969895587  0.1303007521  0.5346021666 0  0  1  0  1    1
4  0.375280428  1.1090689506 -0.576731743   0.1241954368 1  1  1  0  0    1
5 -0.363147587 -0.86985922   -1.039424695  -1.147074384  1  0  1  1  1    1
                           ... ...
                           ... ...
**** Note the X variables are centered (not standardized);
proc standard data = dat mean = 0 out = dat1;
     var x1-x4;
run;

**** PROC NLMIXED for the latent class regressed on latent factor model
with K = 2, Q = 1, P = 4, J = 5;
proc nlmixed data = dat1 tech = quanew lis = 2 method = gauss
     maxiter = 1000 gconv = .00000000001 fconv = .00000000001;
**** starting values (starting values are given for the logit(pi_i|j);
parms
     alpha = 0.5  beta = 1  gamma = 1
     lam11 = 1    lam12 = 1     lam13 = 1
     psi1  = 0.5  psi2  = 0.5   psi3  = 0.5   psi4  = 0.5
     bpi11 = 1    bpi21 = 1     bpi31 = 0     bpi41 = 0     bpi51 = -.5
     bpi12 = 0.5  bpi22 = 0.5   bpi32 = -0.5  bpi42 = -0.5  bpi52 = -1;
bounds -6 <= bpi11 - bpi52 <= 6;

**** latent class part;
pi11 = 1/(1+exp(-bpi11)); pi12 = 1/(1+exp(-bpi12));
pi21 = 1/(1+exp(-bpi21)); pi22 = 1/(1+exp(-bpi22));
```

```
 pi31 = 1/(1+exp(-bpi31)); pi32 = 1/(1+exp(-bpi32));
 pi41 = 1/(1+exp(-bpi41)); pi42 = 1/(1+exp(-bpi42));
 pi51 = 1/(1+exp(-bpi51)); pi52 = 1/(1+exp(-bpi52));
 prod11 = (pi11**y1)*(1-pi11)**(1-y1); prod12 = (pi12**y1)*(1-pi12)**(1-y1);
 prod21 = (pi21**y2)*(1-pi21)**(1-y2); prod22 = (pi22**y2)*(1-pi22)**(1-y2);
 prod31 = (pi31**y3)*(1-pi31)**(1-y3); prod32 = (pi32**y3)*(1-pi32)**(1-y3);
 prod41 = (pi41**y4)*(1-pi41)**(1-y4); prod42 = (pi42**y4)*(1-pi42)**(1-y4);
 prod51 = (pi51**y5)*(1-pi51)**(1-y5); prod52 = (pi52**y5)*(1-pi52)**(1-y5);

 **** relation between latent class and latent factor (structural model);
 eta1=exp(alpha+beta*fi)/(1+exp(alpha+beta*fi));
 eta2=1/(1+exp(alpha+beta*fi));

*** Note: if a covariate W were also added, it could be included in the following
way,
*** eta1=exp(alpha+beta*fi+gamma*Wi)/(1+exp(alpha+beta*fi+gamma*Wi));
*** eta2=1/(1+exp(alpha+beta*fi+gamma*Wi));

 **** the part of the likelihood coming from latent class part;
 l_latclass=eta1*prod11*prod21*prod31*prod41*prod51
           +eta2*prod12*prod22*prod32*prod42*prod52;
 ll_latclass = log(l_latclass);

 **** factor analysis part;
 mu1 = lam11*fi; mu2 = lam12*fi;
 mu3 = lam13*fi; mu4 =     1*fi;

 **** the part of the likelihood coming from latent factor part;
 ll_factpart = -.5*log(psi1) - (1/(2*psi1)) * (x1 - mu1)**2
               -.5*log(psi2) - (1/(2*psi2)) * (x2 - mu2)**2
               -.5*log(psi3) - (1/(2*psi3)) * (x3 - mu3)**2
               -.5*log(psi4) - (1/(2*psi4)) * (x4 - mu4)**2;

 **** dummy is just a place holder so that SAS has something on;
 **** the left side of equation;
 model dummy ~ general(ll_latclass + ll_factpart);
 random fi ~ normal(0,phi) subject = id;
 run;
```

REFERENCES

ABRAMOWITZ, M. AND STEGUN, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover Publications, Inc.

BANDEEN-ROCHE, K., MIGLIORETTI, D., ZEGER, S. AND RATHOUZ, P. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* **92**, 1375–1386.

BARTHOLOMEW, D. J. AND KNOTT, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd edition. Kendall's Library of Statistics, New York: Oxford University Press.

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.

BOWLING, A. (1997). *Measuring Health: A Review of Quality of Life Measurement Scales*, 2nd edition. Philadelphia: Open University Press.

CARROLL, R. J., RUPPERT, D. AND STEFANSKI, L. A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman & Hall.

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

CLOGG, C. C. (1995). Latent class models. In Arminger, G., Clogg, C. C. and Sobel M. E. (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum Press, pp. 311–359.

COLLINS, L. M., FIDLER, P. L., WUGALTER, S. E. AND LONG, J. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research* **28**, 375–389.

CROUDACE, T. J., JARVELIN, M. R., WADSWORTH, M. AND JONES, P. B. (2003). Developmental typology of trajectories to nighttime bladder control: epidemiologic applications of longitudinal latent class analysis. *American Journal of Epidemiology* **157**, 834–842.

DAYTON, M. AND MACREADY, G. (1988). Concomitant-variable latent class models. *Journal of the American Statistical Association* **83**, 173–178.

FLAHERTY, B. P. (2002). Assessing reliability of categorical substance use measures with latent class analysis. *Drug and Alcohol Dependency* **68**, 7–20.

FULLER, W. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

GARRETT, E. S. AND ZEGER, S. L. (2000). Latent class model diagnosis. *Biometrics* **56**, 971–1295.

GOLUB, G. H. AND WELSCH, J. H. (1969). Calculation of Gaussian quadrature rules. *Mathematical Computing* **23**, 221–230.

HABERMANN, S. J. (1977). Product models for frequency tables derived by indirect observation. *Annals of Statistics* **5**, 1124–1147.

HAGENAARS, J. A. AND MCCUTCHEON, A. L. (eds) (2002). *Applied Latent Class Analysis Models*. Cambridge: Cambridge University Press.

JÖRESKOG, K. G. AND SORBORM, D. (1996). *LISREL8: User's Reference Guide*. Chicago, IL: Scientific Software International.

LAWLEY, D. N. AND MAXWELL, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd edition. London: Butterworths.

LAZARSFELD, P. F. AND HENRY, N. W. (1968). *Latent Structure Analysis*. Boston, MA: Houghton Mifflin.

LEE, N. K., OEI, T. P. S., GREELEY, J. D. AND BAGLIONI, A. J. (2003). Psychometric properties of the drinking expectancy questionnaire: a review of the factor structure and a proposed new scoring method. *Journal of Studies on Alcohol* **64**, 432–436.

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

MCHUGH, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* **43**, 551–560.

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.

MUTHEN, B. AND SHEDDEN, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics* **55**, 463–469.

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? *American Journal of Public Health* **92**, 844–851.

NEUMARK-SZTAINER, D., WALL, M. M., PERRY, C. AND STORY, M. (2003a). Correlates of fruit and vegetable intake among adolescents: findings from Project EAT. *Preventive Medicine* **37**, 198–208.

NEUMARK-SZTAINER, D., WALL, M. M., STORY, M. AND PERRY, C. (2003b). Correlates of unhealthy weight control behaviors among adolescent girls and boys: implications for the primary prevention of disordered eating. *Health Psychology* **22**, 88–98.

UEBERSAX, J. S. AND GROVE, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine* **9**, 559–572.

WALL, M. M. AND LI, R. (2003). A comparison of multiple regression to two latent variable techniques for estimation and prediction. *Statistics in Medicine* **22**, 3671–3685.

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. *Journal of the American Statistical Association* **85**, 699–704.

[*Received April 6, 2005; revised July 15, 2005; accepted for publication July 25, 2005*]