Johns Hopkins University, Dept. of Biostatistics Working Papers

1-15-2003

Checking Assumptions in Latent Class Regression Models via a Markov Chain Monte Carlo Estimation Approach: An Application to Depression and Socio-Economic Status

Elizabeth Garrett
Johns Hopkins University School of Medicine, Sidney Kimmel Comprehensive Cancer Center, esg@jhu.edu

Richard Miech
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, rmiech@jhsph.edu

Pamela Owens
Center for Organization & Delivery Systems, Agency for Healthcare Research & Quality

William W. Eaton
Department of Mental Hygiene, Johns Hopkins Bloomberg School of Public Health, weaton@jhsph.edu

Scott L. Zeger
Johns Hopkins Bloomberg School of Public Health, szeger@jhsph.edu

Checking Assumptions in Latent Class Regression Models via a Markov Chain Monte Carlo Estimation Approach: An Application to Depression and Socio-Economic Status

Elizabeth S. Garrett[1] Richard Miech[2] Pamela Owens [3]

William W. Eaton[2] Scott L. Zeger[4]

January 15, 2003

1 Division of Biostatistics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205.

2 Department of Mental Hygiene, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205.

3 Center for Organization and Delivery Systems, Agency for Healthcare Research and Quality, Rockville, MD.

4 Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205.

Contact information: Elizabeth S. Garrett, esg@jhu.edu, Suite 1103, 550 N. Broadway, Baltimore, MD, 21205

Elizabeth Garrett, Ph.D. (esg@jhu.edu) is an Assistant Professor of Oncology and Biostatistics at the Sidney Kimmel Comprehensive Cancer Center and the Bloomberg School of Public Health at Johns Hopkins University, Baltimore, MD. Richard Miech, Ph.D. (rmiech@jhsph.edu) is an Assistant Professor in the Department of Mental Hygiene at the Johns Hopkins Bloomberg School of Public Health. Pamela Owens, Ph.D. (powens@ahrq.gov) is from the Center for Organization and Delivery Systems at the Agency for Healthcare Research and Quality in Rockville, MD. William W. Eaton, Ph.D.(weaton@jhsph.edu) is a Professor in the Department of Mental Hygiene at the Johns Hopkins Bloomberg School of Public Health. Scott L. Zeger, Ph.D. (sz@jhu.edu) is the Chairman and Professor in the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health.
Supported by NIH grant MH56639-03 and NIH grant MH47447.

Abstract

Latent class regression models are useful tools for assessing associations between covariates and latent variables. However, evaluation of key model assumptions cannot be performed using methods from standard regression models due to the unobserved nature of latent outcome variables. This paper presents graphical diagnostic tools to evaluate whether or not latent class regression models adhere to standard assumptions of the model: conditional independence and non-differential measurement. An integral part of these methods is the use of a Markov Chain Monte Carlo estimation procedure. Unlike standard maximum likelihood implementations for latent class regression model estimation, the MCMC approach allows us to calculate posterior distributions and point estimates of any functions of parameters. It is this convenience that allows us to provide the diagnostic methods that we introduce.

As a motivating example we present an analysis focusing on the association between depression and socioeconomic status, using data from the Epidemiologic Catchment Area study. We consider a latent class regression analysis investigating the association between depression and socioeconomic status measures, where the latent variable depression is regressed on education and income indicators, in addition to age, gender, and marital status variables. While the fitted latent class regression model yields interesting results, the model parameters are found to be invalid due to the violation of model assumptions. The violation of these assumptions is clearly identified by the presented diagnostic plots.

These methods can be applied to standard latent class and latent class regression models, and the general principle can be extended to evaluate model assumptions in other types of models.

Keywords: latent class; conditional independence; non-differential measurement; Bayesian

estimation; model diagnosis.

1. INTRODUCTION

Latent class regression models can be useful tools for measuring latent constructs and relating these constructs to covariates. However, latent class model checking is somewhat complicated because we cannot assess model fit using standard approaches which rely on comparing fitted to observed values. In latent class models, we do not observe the true class membership of individuals and so evaluation of model fit and adherence to assumptions is elusive. The goal of this paper is to provide graphical diagnostic tools which can assist in model checking. Specifically, we present methods for checking two of the key assumptions of latent class regression models: conditional independence (CI) and non-differential measurement (NDM). An integral part of these methods is the use of a Markov Chain Monte Carlo (MCMC) estimation procedure. Unlike standard maximum likelihood implementations for latent class regression model estimation, the MCMC approach allows us to calculate posterior distributions and point estimates of any functions of parameters.

Our approach has proven useful in an analysis of the association between depression and socioeconomic covariates. The relationship between depression and socioeconomic variables has been of interest to researchers as a way to investigate the social arrangements of society and their implications for individual well-being (Pearlin, 1989; Turner et al., 1995). A substantial body of evidence indicates that individuals with lower education are more likely to report depressive symptoms because of their greater exposure to "social stressors" such as unemployment, financial strain, or lack of control in the workplace (Link et al., 1993; Turner and Lloyd, 1999). Further, individuals with less education have fewer resources to successfully cope with stressful situations (McLeod and Kessler, 1990). Ongoing investigation in the relationship between depression and socioeconomic indicators, such as educational attainment, is identifying the specific factors that link individual psychological functioning

to broader social structure. We use the latent class regression model, treating depression as categorical, to investigate the hypotheses listed above. An important part of this application is to assess whether or not our model is valid. The methods that we introduce demonstrate that misleading results can be reported is violation of LCR assumptions are not identified and addressed.

This paper is organized as follows. Section 2 describes the latent class regression model and its assumptions. In section 3, the MCMC estimation procedure is outlined. Section 4 describes the Epidemiologic Catchment Area Study example. Sections 5 and 6 develops the model checking methods and applies the diagnostic tools. Finally, section 7 is a discussion.

2. LATENT CLASS REGRESSION MDOEL

We begin by discussing the standard latent class model and then expanding the model to the regression setting.

2.1 The Standard Latent Class Model

First introduced in psychiatric research by Young (1983) and sociological research by Clogg (1979) and Clogg (1980), latent class models have often been used to describe the prevalence and symptomatology of disorders, as well as to assess the reliability and accuracy of psychiatric diagnoses (Faroane and Tsuang, 1994). The general situation in which to apply a latent class model is when the following are true: (1) there exists an underlying (latent) variable that can be conceptually viewed as discrete, (2) there are a number of observed categorical variables that are thought to define the underlying variable, and (3) the observed variables are recorded for a large number of individuals (i.e. a large dataset is available). There are many ways to apply the result of a latent class analysis including classifying individuals into

(clinical) categories (i.e. classes), describing the prevalence of a disease or condition in a population, and predicting prevalence for policy and planning.

In applying an M class latent class model, we assume that each individual is a member of one of the M classes, but we do not know which class. The latent class of individual i is denoted by η_i. Symptom or item prevalences vary by class. The probability that an individual in class j will report symptom k is denoted by $p_{kj},\ j = 1, ..., M$. We define $\underset{\sim}{y_i}$ to be a vector of length K indicating individual i's binary responses to K items, and $\pi_j = P(\eta_i = j)$ to be the probability that individual i is in class j for $j = 1, .., M$. As applied to the ECA depression data, $\underset{\sim}{y_i}$ is a vector representing the presence and absence of K symptoms of depression for individual i, η_i is individual i's "true" but unknown depression class, and π_j is the proportion of individuals in the representative sample in depression class j.

The likelihood function for the latent class model is

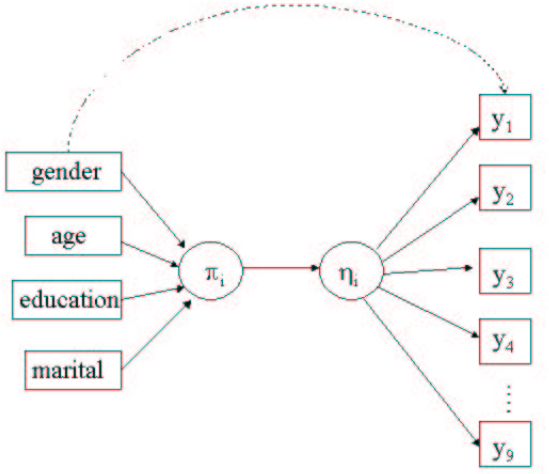
$$f(Y; \underset{\sim}{\pi}, \underset{\sim}{p}) = \prod_{i=1}^{N} \sum_{j=1}^{M} \pi_j \prod_{k=1}^{K} p_{jk}^{y_{ik}} (1 - p_{jk})^{1-y_{ik}}.$$

Clogg (1995) and McCutcheon (1987)) provide more detailed discussions of the latent class model.

2.2 The Latent Class Regression Model

Similar to previous authors (Dayton and Macready, 1988; Bandeen-Roche et al., 1997a), we extend the LCM to the regression setting. In the LCR models that we employ, the probability of class membership is related to an individual's covariates, and this relationship is described via odds ratios. To see examples of other LCM extensions, see Clogg and

Figure 1: Graphical model depiction of the LCR model. NDM is indicated by the dashed line between gender and y_1



Goodman (1984), Melton, Liang, and Pulver (1994) and Huang and Bandeen-Roche (2000). The methods we propose will add diagnostic methods for quantifying if the data is consistent with the LCR assumptions.

The difference that we see in comparing the LCM to the LCR is in the modeling of η_i. In the standard LCM, no information is known aside from symptom responses that is related to an individual's class membership. In the LCR model, we allow covariates to be associated with class membership. We are incorporating more observed information (a vector of covariates $\underset{\sim}{x}_i$) and represent the association between $\underset{\sim}{x}_i$ and η_i by the parameter vector $\underset{\sim}{\beta}$. These two quantities replace $\underset{\sim}{\pi}$ in the standard LCM representation as demonstrated in equation (**??**) below. A graphical representation of the model is shown in Figure 1, where the solid lines indicate associations in the LCR model.

To allow a vector of covariates ($\underset{\sim}{x}_i$) to be associated with individual i's class membership,

we use the parameterization suggested in Bandeen-Roche et al. (1997b):

$$f(\underset{\sim}{y}_i; \underset{\sim}{x}_i, \underset{\sim}{\pi}, \underset{\sim}{p}) = \sum_{j=1}^{M} \left\{ \pi_j(\underset{\sim}{x}_i) \prod_{k=1}^{K} p_{jk}^{y_{ik}} (1-p_{jk})^{1-y_{ik}} \right\}.$$

Specifically, we allow x_i to be associated the latent outcome using the following logistic relationship between probability of class membership (π_j) and a single covariate (x_i):

$$\pi_j(x_i) = \frac{e^{\beta_{0j}+\beta_{1j}x_i}}{\sum_{l=1}^{M} e^{\beta_{0l}+\beta_{1l}x_i}} \qquad j = 1, .., M$$

where we are constrained so that $\beta_{0M} = \beta_{1M} = 0$. In this representation, β_{1j} can be interpreted as the log odds ratio comparing individuals in class j to those in class M with respect to a one unit change in x_i. If β_{1j} is positive, the model suggests that an individual with a high value of x_i is more likely to be in class j than in class M as compared to an individual with a low value of x_i. We can rewrite the distribution of y_i as

$$f(\underset{\sim}{y}_i; \underset{\sim}{x}, \underset{\sim}{\beta}, \underset{\sim}{p}) = \sum_{j=1}^{M} \left\{ \left(\frac{e^{\beta_{0j}+\beta_{1j}x_i}}{\sum_{l=1}^{M} e^{\beta_{0l}+\beta_{1l}x_i}} \right) \prod_{k=1}^{K} p_{jk}^{y_{ik}} (1-p_{jk})^{1-y_{ik}} \right\}. \quad (1)$$

The distribution of $\underset{\sim}{y}_i$ for two or more covariates is straightforward. In the case of multiple covariates, the coefficients can be interpreted as the log odds ratios comparing probabilities of class membership adjusting for the other covariates in the model.

2.3 LCR Model Assumptions

The model in equation (1) imposes a CI assumption. This implies that, within a class, symptoms are independent. That is, conditional on class membership there is no association

between individual responses to items:

$$P(Y_{ik} = y_{ik}, Y_{ik'} = y_{ik'}|\eta_i) = P(Y_{ik} = y_{ik}|\eta_i)P(Y_{ik'} = y_{ik'}|\eta_i), \qquad k \neq k'. \tag{2}$$

An additional assumption imposed by the LCR model is a NDM assumption. For individuals within a class, there is no association between covariates and symptoms. That is, conditional on class membership, individual responses and the covariates are independent:

$$P(Y_{ij} = y_{ij}|\eta_i, X_i) = P(Y_{ij} = y_{ij}|\eta_i). \tag{3}$$

These assumptions will be more formally examined in section 4.

3. MCMC ESTIMATION APPROACH

A benefit to the MCMC estimation procedure is that in addition to posterior distributions for the p and β parameters in the LCR as described in the previous section, posterior distributions can be calculated for any functions of these parameters with relative ease. As a result, we are able to derive both point estimates and precision estimates for any function of the parameters. Using standard EM algorithm approaches for maximum likelihood (ML) estimation, we are limited to results that include the precision estimates for only the p and β parameters. Estimation of a confidence interval for, for example, the log odds ratio between symptoms 1 and 3 within individuals class 2 would not be easily obtained from a standard ML package. However, a great benefit to the ML estimation in many packages is the great speed at which results are obtained and the simplicity with which they are presented. Using the MCMC approach, the results must be post-processed to obtain meaningful results. The graphical displays presented in this paper use the posterior distributions of the LCR model

parameters, in addition to functions of those parameters.

We have used the MCMC approach to fitting latent class model in several published examples (Garrett and Zeger, 2000; Garrett et al., 2002). Our MCMC estimation procedure is as follows. We define $\underset{\sim}{x_i}$ to be the vector of covariates for individual i. We use a linear logistic reparameterization for symptom prevalences ($g_{jk} = log(\frac{p_{jk}}{1-p_{jk}})$) as described in Formann (1996) so that the full-likelihood for a dataset with N individuals is defined to be

$$L(\underset{\sim}{\beta}, \underset{\sim}{g}; \underset{\sim}{x}, \underset{\sim}{y}) = \prod_{i=1}^{N} \sum_{j=1}^{M} \left\{ \left(\frac{e^{\underset{\sim}{\beta_j} \underset{\sim}{x_i}}}{\sum_{l=1}^{M} e^{\underset{\sim}{\beta_l} \underset{\sim}{x_i}}} \right) \prod_{k=1}^{K} \frac{e^{g_{jk} y_{ij}}}{1 + e^{g_{jk}}} \right\}.$$

The MCMC algorithm for an M class model uses the following full-conditional distributions

$$\begin{aligned}
p(g_{jk} | \underset{\sim}{\eta}, \underset{\sim}{y_k}) &\propto P(g_{jk}) \times \prod_{i=1}^{N} \left(\frac{e^{y_{ik} g_{jk}}}{1 + e^{g_{jk}}} \right)^{I(\eta_i = j)} \\
p(\beta_{vj} | \underset{\sim}{\eta}, \underset{\sim}{x}) &\propto P(\beta_{vj}) \times \prod_{i=1}^{N} \left(\frac{e^{\underset{\sim}{\beta_{\eta_i}} \underset{\sim}{x_i}}}{\sum_{l=1}^{M} e^{\underset{\sim}{\beta_l} \underset{\sim}{x_i}}} \right) \\
p(\eta_i | \underset{\sim}{\beta}, \underset{\sim}{g}, \underset{\sim}{y_i}, \underset{\sim}{x_i}) &\propto \frac{e^{\underset{\sim}{\beta_{\eta_i}} \underset{\sim}{x_i}}}{\sum_{l=1}^{M} e^{\underset{\sim}{\beta_l} \underset{\sim}{x_i}}} \times \prod_{k=1}^{K} \frac{e^{y_{ik} g_{\eta_i k}}}{1 + e^{g_{\eta_i k}}} \\
\text{for } k &= 1, \ldots, K; \; j = 1, \ldots, M; \; v = 0, 1, 2; \; i = 1, \ldots, N
\end{aligned}$$

where $P(g_{jk})$ and $P(\beta_{vj})$ are the priors for g_{jk} and β_{vj}. All of the priors in the models that we consider in later sections are specified so that $g_{jk} \sim N(0, 2.25)$, and $\beta_{vj} \sim N(0, 2.25)$ unless otherwise noted. These densities translate to proper, yet relatively flat, prior distributions on the model parameters. See Garrett and Zeger (2000) for more information about the choice of priors.

Models were fit and traceplots of model parameters were produced using WinBugs version 1.3 (Imperial College of Science and Medicine, 2000). See Chib and Greenberg (1995),

Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and Hastings (1970) for details of the estimation procedure. In the WinBugs implementation, a Gibbs sampler is used which includes Metropolis-Hastings steps. Convergence diagnostic methods described in Kass, Carlin, Gelman, and Neal (1997) were implemented via Splus 2000 (MathSoft, 1999), and include running multiple chains using different starting values and checking comparability of results. WinBugs program files for estimation of the LCR model and the Splus programs used to create the figures are available at `http:/astor/ esg/software.html`. Using WinBugs on a Dell Inspiron 7500 with 750mHz Pentium III processor, allowing for 500 burn-in iterations and an additional 5000 iterations, total running time was between 30 and 60 minutes (depending on the number of parameters in model). Output from the MCMC chain were saved in text files and graphical displays and diagnostics were performed using Splus2000 (MathSoft, Inc., 1999). The simulated parameters $\underset{\sim}{g}$ are back-transformed to obtain the values of $\underset{\sim}{p}$ to which we refer below.

All models were also fit using a maximum likelihood (ML) approach to validate the results found in the MCMC approach. Mplus (Muthen and Muthen, 2001), which employs an EM algorithm, was used for ML model estimation. Running time for each model was under 20 seconds. For details of the EM algorithm, see Mooijaart and van der Heuden (1992).

4. THE EPIDEMIOLOGIC CATCHMENT AREA STUDY

The National Institute of Mental Health Epidemiologic Catchment Area (ECA) Program consists of coordinated sample surveys at five sites of research. In the Baltimore site, 3481 individuals in the community were interviewed in 1981, with a response rate of 78%. The National Institute of Mental Health Diagnostic Interview Schedule (DIS) was used in the

interviews to collect information on mental health. The DIS is a structured interview similar to a typical psychiatric interview and designed to produce similar diagnoses as would a psychiatrist. The design of the study is described in detail in Robins and Regier (1991) and validity and reproducibility of the DIS are shown in Robins, Helzer, and Orvaschel (1985). The Baltimore ECA sample was followed and interviewed again between 1993 and 1996 and DIS interviews were obtained from 1920 of the original 1981 sample of 3481, which amounted to 73% of the survivors of the baseline sample. More detailed description of the rationale for follow-up and the sample attrition can be found in Eaton et al. (1997) and Badawi, Eaton, Myllyluoma, Weimer, and Gallo (1999), respectively.

We investigated the six-month prevalence of depressive episodes (i.e. symptoms that were reported in the previous six months) in the Baltimore wave 3 data. The DSM-IIIR criteria for diagnosis of major depression requires evidence of symptoms in five of a total of nine possible symptom groups where one of the five groups must be group 1 (depressed mood) or group 2 (loss of interest or pleasure). These symptoms and their prevalences in the Baltimore ECA data at round 3 are listed in Table 1.

In our analyses, we considered only individuals who had full information on the 22 questions in the DIS pertaining to depression and relevant covariates. In addition, we restricted our attention to adults who were younger than age 66 in light of evidence that depression may manifest itself differently in the elderly (Gallo et al., 1994). Due to missingness primarily in income (where many respondents reported not knowing their household income), we are left with a sample size of 1126. To have comparability across models, we used this sample for all LCMs and LCRs that were estimated. The symptom prevalences of the sample are shown in the first second column of Table 1 and the demographic characteristics in Table 2. Although a series of LCR models were estimated, we only present one model in this paper

Table 1: Overall Symptom Prevalences and LCR model results for Symptom Prevalences in the ECA Wave 3 data (N = 1126). Class sizes are 82%, 14%, and 4% in classes 1, 2, and 3, respectively.

		prevalence	class 1 non-depressed	class 2 mild depression	class 3 severe depressi
Symptom Group	Symptoms				
1	depressed mood	0.11	0.02	0.41	0.82
2	loss of interest	0.11	0.02	0.42	0.86
3	loss of appetite weight loss increased appetite weight gain	0.10	0.04	0.31	0.67
4	trouble falling asleep waking too early sleeping too much	0.13	0.04	0.44	0.73
5	fast movement slow movement	0.05	0.01	0.07	0.74
6	fatigue	0.06	0.01	0.21	0.64
7	feel worthless/sinful feel inferior low self-confidence	0.05	0.01	0.15	0.67
8	trouble concentrating slow thoughts unable to decide	0.06	< 0.01	0.14	0.91
9	thoughts of death want to die thoughts of suicide suicide attempt	0.09	0.03	0.29	0.65

to demonstrate our proposed methods.

Our regression analyses include four variables. *Gender* is a dichotomous variable coded 1 for women and 0 for men. To assess marital status, currently married is our reference category, and we include indicator variables for *single* (1 = never married, 0 = otherwise), and *sep/wid/div* (1 = separated, widowed, or divorced; 0 = otherwise). The age effect was assesed using the natural log of *age*. We considered other functional forms of age (e.g. quadratic and spline models of age), but the log transformation described the data well and uses only one degree of freedom (analyses not shown). High school *diploma* is an indicator variable coded as 1 for individuals who received a diploma or GED and as 0 otherwise. Finally, *poverty* (1 = below the poverty line, 0 = above the poverty line) was derived from the poverty index used by the federal government (United States Census Bureau, 1993) and was based on age, household size, and family income. Other forms of income and education were considered (e.g. number of years of education, income in 1000's of dollars). The dichotomous variables, however, seemed to describe the associations with depression as well or better than continuous variables and are easy to interpret in the LCR setting.

The model that we present in section 6 includes age, gender, marital status, and education effects. As mentioned above, although a series of models have been estimated and evaluated, we have chosen just one to present for brevity.

5. Methods for Assessing Model Assumptions

The model assumes that there is a relationship between covariates (e.g. gender, age) and risk of class membership (π_i) and that there is a relationship between class membership (η_i) and symptom responses (p_{ik}). These associations are expressed in Figure 1 by the solid black arrows linking covariates to class prevalences and linking class membership to

Table 2: Demographic Characteristics of the ECA Wave 3 data (N = 1126)

	Mean	Std. Dev.	Range
Age	44.2	9.01	30, 65
Income (thousands of dollars)	33.3	24.3	0, 150[1]
Years of Education	12.3	2.6	0, 17[2]
	%		
Female	61		
Marital Status			
married	54		
single	29		
divorced/separated/widowed	17		
Poverty	21		
High School Diploma	72		

[1] 150 indicates 150 thousand or greater.
[2] 17 years of education indicates graduate school.

symptoms. Note that conditional independence requires that there are no arrows between the $y's$: the $y's$ are only associated with each other through their relationship with class. Similarly, NDM requires that there be no arrows from the covariates to the symptoms. In the example shown in figure 2, a dotted arrow is drawn between gender and y_1 to indicate how we would indicate NDM.

5.1 Assessing the Condtional Independence Assumption

The CI assumption in equation (2) implies that the odds ratio between items k and k' within class j should be equal to 1 for all $k, k' = 1, ..., K,\ k \neq k'$:

$$OR_{kk'j} = \frac{P(y_k = 1, y_{k'} = 1|\eta_i = j)P(y_k = 0, y_{k'} = 0|\eta_i = j)}{P(y_k = 0, y_{k'} = 1|\eta_i = j)P(y_k = 1, y_{k'} = 0|\eta_i = j)}. \quad (4)$$

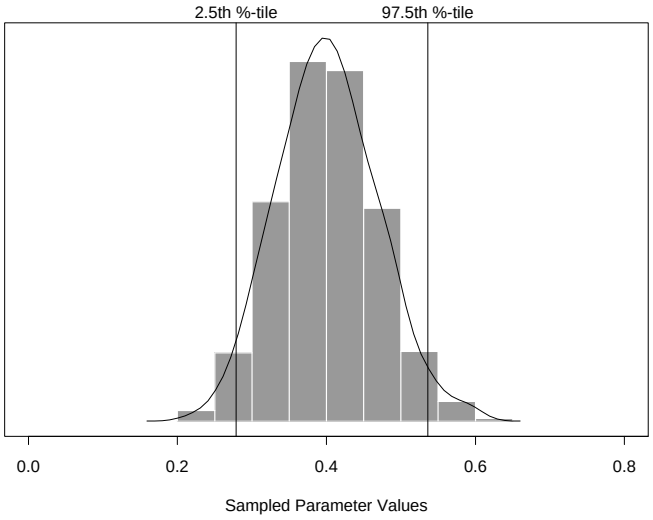
Equivalently, $\log(OR_{kk'j})$ should be equal to 0 in the case of CI. This may appear to be a simple quantity to estimate, but recall that we do not know η_i for $i = 1, ..., N$. Hence, we need to a strategy for assigning individuals to classes and, most importantly, for accounting

for possible misclassifications.

In the MCMC approach, at each of the T iterations of the chain, η_i is sampled for each individual. Unlike other approaches, there is no need to "manually" assign individuals to classes in a *post hoc* procedure because individuals are already assigned to classes at every iteration of the chain. As as result, for each iteration we can then calculate the odds ratio in equation (4) between each pair of items for each class. This boils down to simple pairwise comparisons of symptoms within classes and relies on no model parameters aside from the sampled η_i values. This procedure results in $J \times K$ odds ratios at each of T iterations. We can then use the same standard inferential method for MCMC estimation procedures, which is to estimate the posterior density using the parameter estimates of $\log(OR_{kk'j})$ at the T iterations. Essentially, we calculate the "empirical" density of the values across iterations. This is demonstrated in Figure 2, where a histogram of sampled values is plotted and a density fitted to the distribution. This density represents the posterior distribution of the parameter of interest The mean and standard deviation of the posterior distribution provide us with the point estimate and standard error of the parameter.

Of interest in this paper is whether or not the assumption that $logOR_{kk'j}$ is equal to 0 is reasonable for all of $k, k' = 1, ..., K$ and $k \neq k'$, . We can evaluate this by looking to see where the posterior distribution of $logOR_{kk'm}$ overlaps 0. If the interval between the $2.5th$ and the $97.5th$ quantiles of the estimated posterior distribution contains 0, then we can conclude that there is not evidence against CI. If 0 is outside this 95% posterior interval, then we conclude that the assumption is reasonable. Other quantiles can be used, but we have chosen the $2.5th$ and $97.5th$ to be consistent with the idea of 95% confidence intervals.

Figure 2: Example of histogram of sampled values of a parameter and the associated posterior interval generated from empirical distribution. 95% posterior interval is marked by vertical lines.



5.2 Assessing the Non-differential Measurement Assumption

Similar to the CI assumption, if the NDM assumption in equation (2) holds, then the odds ratio between item k and covariate x_r within class j should be equal to 1 for all $k = 1, ..., K$ items, and $r = 1, ..., R$ covariates. We show this for a binary covariate in equation (**??**), however it is true for both continuous and categorical covariates.

$$OR_{krm} = \frac{P(y_k = 1, x_r = 1|\eta_i = j)P(y_k = 0, x_r = 0|\eta_i = j)}{P(y_k = 0, x_r = 1|\eta_i = j)P(y_k = 1, x_r = 0|\eta_i = j)}.$$

Equivalently, $\log(OR_{krm})$ should be equal to 0 if the NDM assumption holds. The methods for creating the posterior interval for checking the NDM assumption are equivalent to those for CI, described in the previous section.

Table 3: Log odds ratios from LCR model regression: posterior mean (posterior standard deviation). Log odds ratios with Z-values larger than 2 or smaller than -2 are indicated by *.

	Class 3 vs. 1 (severe vs. none)	Class 2 vs. 1 (mild vs. none)	Class 3 vs. 2 (severe vs. mild)
Log(age)	-1.23 (0.77)	-1.45 (0.54)*	0.23 (0.89)
Gender	0.85* (0.37)	0.76* (0.25)	0.09 (0.47)
Single	0.44 (0.44)	0.38 (0.30)	-0.05 (0.53)
Sep/Wid/Div	0.86* (0.36)	0.83* (0.24)	-0.01 (0.42)
Diploma	-0.01 (0.36)	-0.56* (0.22)	0.51 (0.42)

6 . THE LCR ON THE ECA: CHECKING FOR VIOLATION OF ASSUMPTIONS

6.1 LCR Model Results

Based on previously described methods (Garrett and Zeger, 2000), we chose to fit a LCR model assuming 3 classes. There are five covariates in the model that we present: log(age), gender, single (versus married), separated/widowed/divorced (versus married), and diploma. The estimated symptom prevalences (p) and class sizes (π) are in Table 1 and the regression coefficients relating covariates to classes are in Table 3.

Without checking the model and simply interpreting the model assuming the CI and NDM requirements are met, we would have some interesting findings to report. Women are more likely to be in the severe and mild classes than men, and previously married individuals are more likely to be in the severe and mild classes than those who are currently married. High school diploma is not associated with risk of being in the severe versus no depression classes, but it does appear to be associated with risk of mild versus no depression. In other words, we might conclude that individuals who do not have a high school are at increased

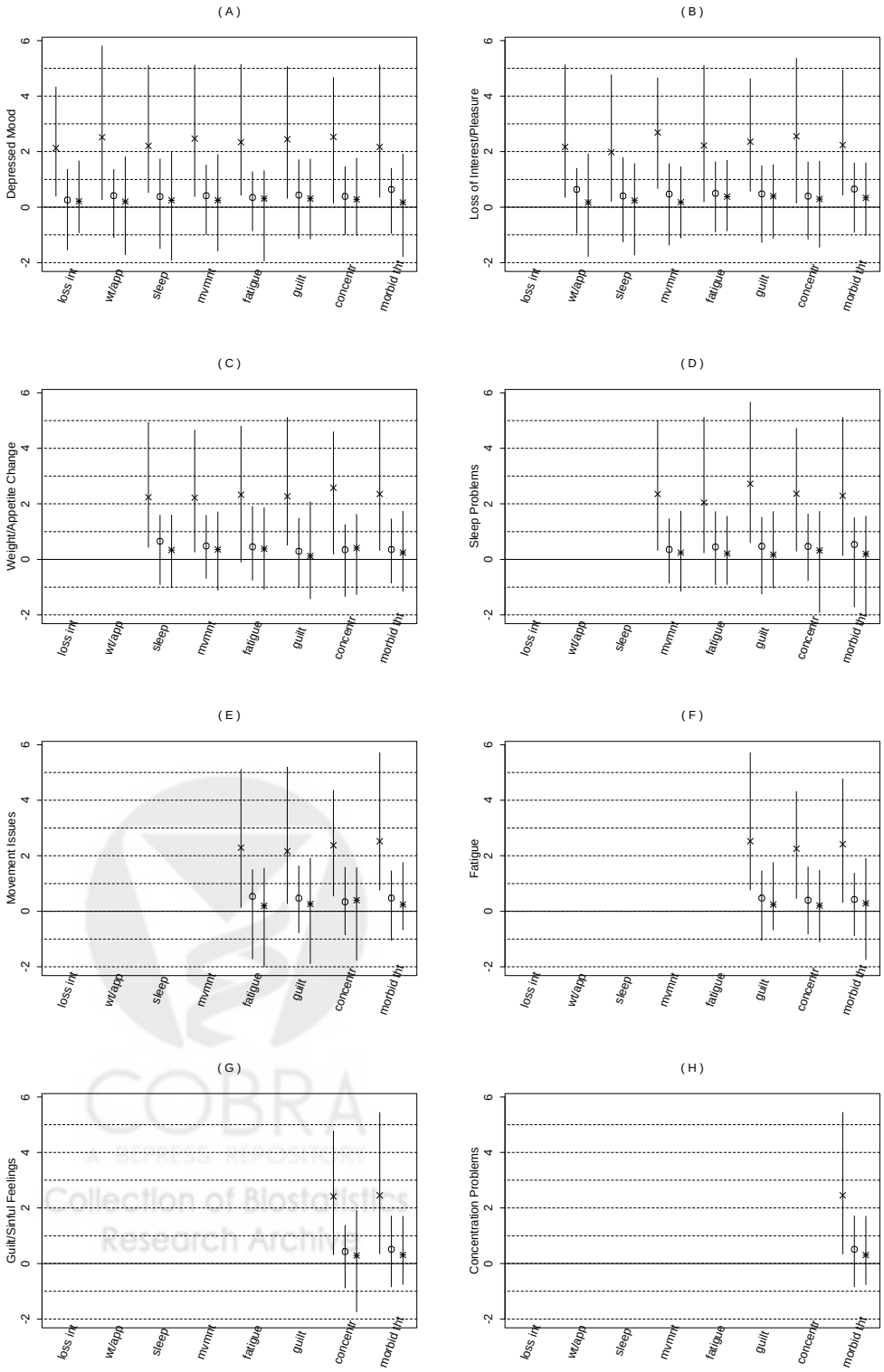
risk of mild versus no depression compared to those who do. We also see that individuals with a diploma are more likely to fall into the severe than the mild class (although not significant due to the relatively small sizes of classes 2 and 3).

We may be able to come up with a rationale for this unusual association that we observe. First, attainment of a high school diploma may be associated with a particular interpretation and response to questions about depression symptoms. For instance, individuals with a high school diploma may be more cautious or reluctant to indicate that they have symptoms of negative affect, while individuals without a high school diploma may be less concerned about the negative stigma associated with a positive response to such symptoms. Second, individuals with a high school diploma may be less likely to experience stressful circumstances which could be related to subthreshold symptoms of depression or they may have additional resources that protect against negative affective symptoms that result from stressful circumstances. Finally, individuals without a high school diploma may be more susceptible to stressful life circumstances or lack coping resources to deal with the stress, which thereby increases the likelihood of subthreshold symptoms of depression. However, none of these explanations mean anything unless the model is valid.

6.2 Assessment of Assumptions

In Figure 3, plots of the 95% posterior intervals as described in section 5 are shown. For K items, there are $K(K-1)/2$ log odds ratios for each of the M classes to examine for conditional independence. In the plots in figure 3, a horizontal line is drawn at 0 and the vertical lines on the plot indicate the 95% posterior intervals, with a symbol plotted at the posterior median (X = class 1, O = class 2, and * = class 3).

Figure 3: Graphical diagnostic plot for assessing CI. Values plotted are log odds ratios. Vertical lines range from the $2.5th$ percentile to the $97.5th$ percentile of the posterior distribution. Posterior median estimates are plotted for class 1 (X), class 2 (O), and class 3 (*). Vertical lines which do not overlap 0 indicate evidence of violation of conditional independence assumption.



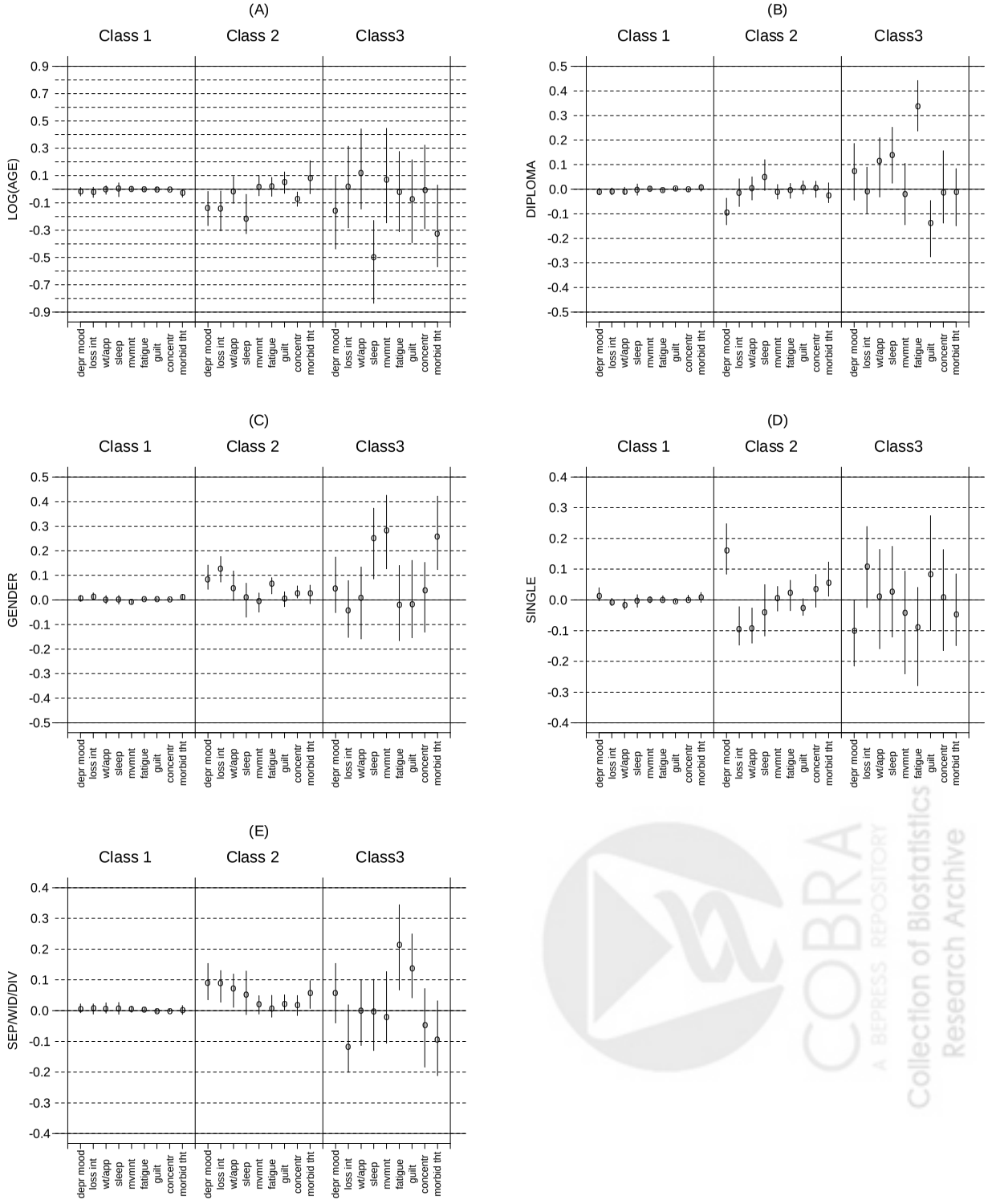
We can clearly see that CI appears to be violated in class 1 across all of the items: the posterior medians of the log odds ratios all range from 2 to 3 (i.e. which correspond to odds ratios from 7.4 to 20.1) and none of the posterior intervals includes 0. However, it does not appear that any two symptoms are more related than any other pair of symptoms. Also, note the large width of the posterior intervals for class 1, which seems counter-intuitive as class 1 is the largest class ($\pi_1 = 0.82$). The reason for this consistent pattern in the class 1 log odds ratios is the relatively low prevalence of all items in class 1. In short, class 1 individuals almost always report no symptoms. However, there are instances in which they report symptoms, but usually very few. In this case, if a class 1 individual reports a particular symptom, then s/he has an increased chance at reporting two. This explains the relatively large magnitude of the estimated log odds ratios. The large width of intervals is due, again, to the small number of individuals reporting symptoms. The log odds ratio is based on the 2×2 tabulation of symptoms within a class. When there are cells that have small counts, the estimate of the log odds ratio becomes unstable and its precision is poor. In the case of class 1 individuals, the 2×2 tables of symptom responses have almost all entries in the $[0, 0]$ cell, and few in the other cells (i.e $[0, 1]$, $[1, 0]$, and $[1, 1]$) making the variance of the estimated log odds ratio large.

In classes 2 and 3, it appears that all log odds ratios show consistency with the CI assumption. We can see this by looking at the magnitude of the estimates (all are close to 0) and observing that 0 is in all of the posterior intervals. However, there is a consistent pattern seen: the estimated log odds ratios in classes 2 and 3 are all positive. This can also be attributed to the same reason as mentioned above: there is a slight positive association between reporting symptoms, but again no symptom pairs appear to be more related than any other symptom pairs.

Figure 4 shows the plots for assessing NDM. We see the opposite result in these plots that we saw in those assessing CI: the non-depressed class seems to obey NDM, but the mild and severe classes do not for at least one of the depression symptoms for each of the covariates of interest. Note, however, that the magnitude of the estimated log odds ratios tend to be small: for the binary covariates, all log odds ratios are between -0.15 and 0.35. A log odds ratio of 0.35 corresponds to an odds ratio of 1.42, which is relatively small. (Similarly, log odds ratios of 0.1, 0.2, and 0.3 correspond to odds ratios of 1.11, 1.22, and 1.35).

In examining the diagnostic plot for age (figure 4a), the item that shows the most notable departure is sleep. It appears that the log odds ratios between sleep and log(age) within class 2 and within class 3 are less than

Figure 4: Graphical diagnostic plot for assessing NDM. Values plotted are log odds ratios. Vertical lines range from the $2.5th$ percentile to the $97.5th$ percentile of the posterior distribution. Posterior median estimates are plotted with "o". Vertical lines which do not overlap 0 indicate evidence of violation of differential measurement assumption.



zero, implying that there is still an association between sleep and age above that which is accounted for in the LCR model. In other words, the LCR model assumes that age is only associated with sleep problems through the latent class. However, what we find is that there is still a residual association between sleep and age. As a result, we find that older individuals within the depressed classes are less likely to report sleep problems than young people and we must conclude that we have violated the NDM assumption.

In Figure 4b, several items violate the NDM assumption via high school education (i.e. diploma). Diploma has a positive association with fatigue, meaning that within the severe depression class, individuals who have a high school diploma are more likely to report fatigue than those without a diploma. Additionally, severely depressed individuals with a diploma are less likely to report guilty and sinful feelings than those with a diploma.

Women in the depressed classes tend to be at higher risk of reporting several symptoms than men based on the results shown in figure 4c. In the severely depressed class, women are at higher risk of reporting sleep problems, movement problems, and morbid thoughts than men. In the mildly depressed class, women are more likely to report depressed mood and loss of interest or pleasure in normal activities.

In comparing previously married (i.e. separated, widowed, or divorced)

to married individuals, those previously married and in the mildly depressed class are more likely to report depressed mood than mildy depressed married individuals and slightly less likely to report loss of interest and weight and appetite changes. Those who are previously married and mildy depressed are slightly more likely to report depressed mood, loss of interest, and weight and appetite changes than mildly depressed married individuals. Among the severely depressed individuals, those who have previously been married are more likely to report fatigue and guilt.

7. DISCUSSION

LCR models can help to summarize the relationship between risk factors and latent variables in a succinct way, but it is important to check that the application of the model is valid. In the above example, a latent class regression model was fit to the ECA data for depression with covariates age, gender, marital status, and an indicator of high school diploma. If we had not performed model diagnosis, we might have interpreted the model, making false claims about associations between, for example, gender and depression class. The NDM diagnostic plots allowed us to see that, conditional on the symptom prevalences in the classes, women and men in the same depression

classes tend to report symptoms differently. For example, severely depressed women are more likely to report sleep problems than men. We have further explored the differential measurement by men and women by fitting separate latent class models for women and for men, shown in table 4. If there were no differential measurement, we would expect the symptom prevalences to be spproximately the same in the two models, although the class sizes might be different (this equivalence of symptom prevalences can be formally tested, but we have not done so in this example). This is another approach to assessing NDM, but not a feasible one: each categorical covariate needs to be analyzed individually, and continuous covariates need to be discretized to be analyzed in this way. Even so, fitting the separate models for men and women in this case allows us to see how depression either manifests itself differently in men and women, or how men and women simply respond differently to symptom questions.

Notice in table 4 that fewer men tend to be in the depressed classes. In comparing the symptom prevalences across classes, we see that classes two and three tend to be different most notably for movement problems, sleep problems, depressed mood, morbid thoughts, and concentration problems. Most of these (all except for concentration problems) were identified as problematic via the diagnostic plots in Figure 4. Instead of going to the trouble

Table 4: Comparison of fitted latent class models for males and females.

	Females ($N = 685$)			Males($N = 441$)		
	class 1	class 2	class 3	class 1	class 2	class 3
Class Size	0.79	0.16	0.05	0.90	0.07	0.03
Symptom Group						
depressed mood	0.03	0.44	0.86	0.02	0.43	0.69
loss of interest	0.04	0.48	0.82	0.01	0.44	0.79
weight/appetite	0.04	0.39	0.62	0.05	0.25	0.65
sleeping problems	0.05	0.45	0.82	0.05	0.49	0.47
movement too slow/fast	0.01	0.10	0.87	0.02	0.10	0.47
fatigue	0.01	0.24	0.64	0.01	0.18	0.56
sinful/worthless	0.01	0.16	0.68	<0.01	0.25	0.58
concentration problems	0.01	0.18	0.91	<0.01	0.20	0.72
morbid thoughts	0.04	0.27	0.74	0.02	0.42	0.37

of performing stratified analyses such as this, by using the methods proposed, we can more easily assess the NDM assumption.

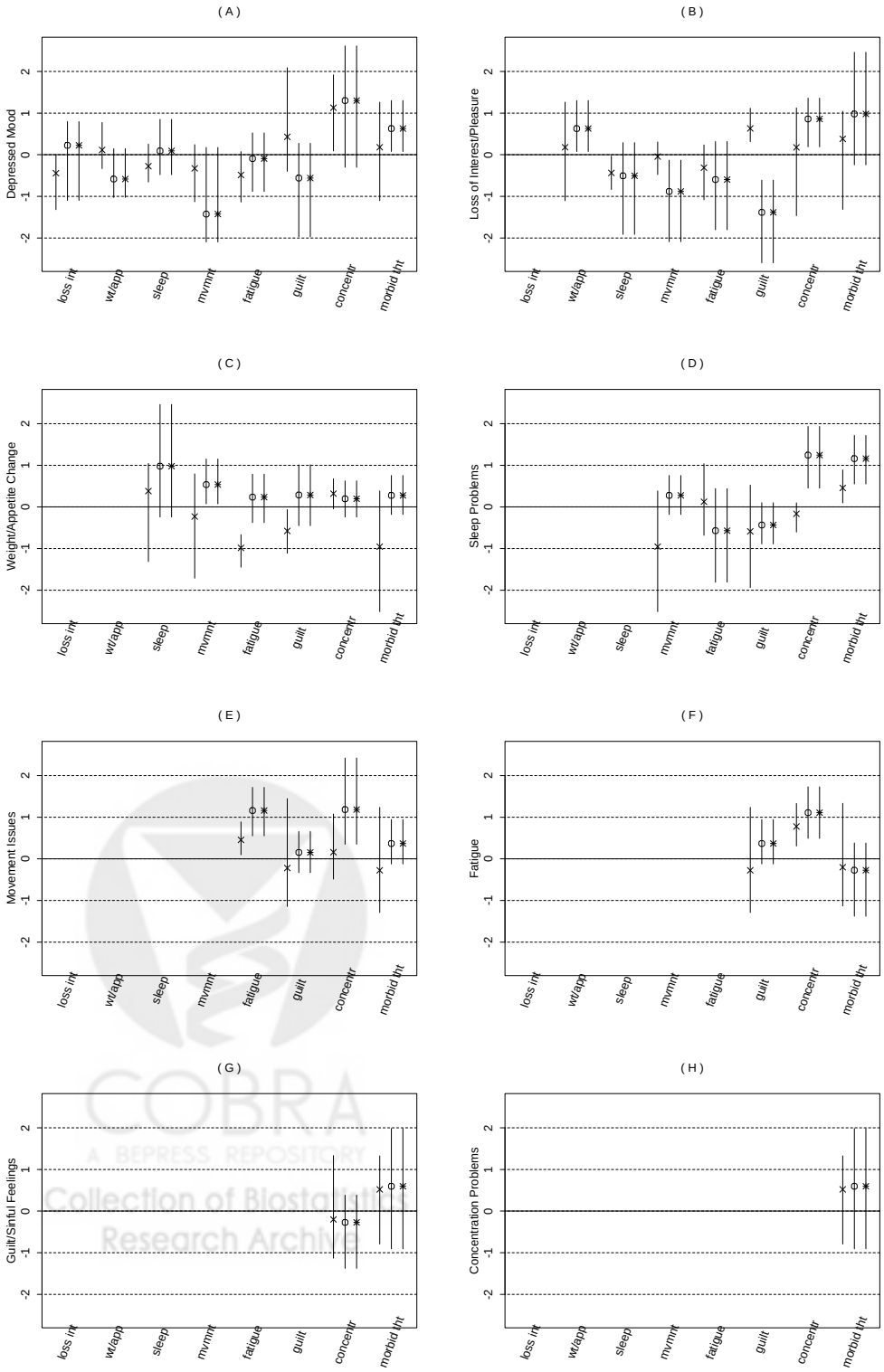
We saw that CI assumption was also violated in class 1, but appeared to be viable in classes 2 and 3. To investigate our hypothesis that the large number of individuals with no reported symptoms we removed individuals from the dataset who reported no symptoms and were left with a total of 337 individuals who reported at least one symptom. The latent class model was fit and the symptom prevalences were compared to the model fit in section 5. By eliminating those who reported no symptoms, we have essentially changed our "reference" population. Although we see improvements in the conditional independence plots shown in figure 5, by changing the dataset by

removing asymptomatic people, we have changed the resulting definition of depression (as can be seen in the symptom prevalences, not shown here for issues of space).

The obvious question to pose is "now that we know we have violated one (or more) of these assumptions, how can we fix things so that the LCR model can be applied?" In looking at the plot in figure 4, we can see what needs to occur to deal with the NDM assumption: we need to account for the association between the covariates and the symptoms. To do so, a model more complicated than the standard LCR model should be derived. As is shown in the graphical representation in Figure 1, we need a model that allows for arrows between symptoms and covariates, as is indicated by the dotted lines between gender and y_1. For examples of extensions of the LCR that allow additional associations, see Melton et al. (1994) and Huang and Bandeen-Roche (2000).

To correct for CI, the same approach as above can be taken: remove individuals who could be inducing conditional independence based on their patterns of reporting. Another sensible approach is to combine items that show strong associations. For example, notice in table 1 that symptoms are collapsed into symptom groups (e.g. loss of appetite and weight loss are both in symptom group 3). These symptoms are in the same group due

Figure 5: Assessment of CI after removing individuals who reported no symptoms. Vertical lines range from the $2.5th$ percentile to the $97.5th$ percentile of the posterior distribution. Posterior median estimates are plotted for class 1 (X), class 2 (O), and class 3 (*). Vertical lines which do not overlap 0 indicate evidence of violation of conditional independence assumption.



to their strong association with each other. In the presence of conditional independence where there appear to be pairs of symptoms that are strongly related, they can be collapsed in this way. However, in our ECA example, we do not have pairs that are associated more than other pairwise combinations: the violation is due to a different source (namely, the floor effect caused by the large group of individuals reporting no symptoms).

We have used the MCMC estimation approach because it naturally provides us with ways of assigning individuals to classes without additional post hoc computation. There are other approaches, however, that can be used in the maximum likelihood setting. Bandeen-Roche et al (1997b), and Bandeen-Roche, Huang, Munoz, and Rubin (1999) use the estimated posterior probabilities of class memberships, which can be calculated for each individual based on his response pattern and the model parameters. The approach they take is to simulate class assignments for each individual using the posterior probability of membership and to repeat multiple times. Using this method, plots similar to those presented here can be made. However, these plots will not have the same interpretation: even for a large number of repeated simulations of class assignments, the empirical distribution of the estimated log odds ratios will not be the posterior distribution as we get using our MCMC results. The reason for this is that the pseudo-class ap-

proach assumes that the model parameters are fixed or known. As a result, the empirical distribution of log odds ratios resulting from the pseudo-class approach will be narrower than the true posterior distribution.

Lastly, it is important to realize that this method has been used in the case of checking LCR models, but it is easily generalizable to other situations for checking model assumptions where there are certain parameters that are assumed to be constant. For example, in Cox proportional hazards models, it is assumed that the the hazard ratio comparing two groups is independent of time. This is often not the case, but formal testing of this assumption is often not implemented. Using an approach similar to that which we propose, graphical displays could be created to test the proportionality of hazards over time.

References

Badawi, M. A., Eaton, W. W., Myllyluoma, J., Weimer, L., and Gallo, J. (1999). Psychopathology and attrition in the baltimore eca follow-up 1981-1996. *Social Psychiatry and Psychiatric Epidemiology*, 92:3491–3498.

Bandeen-Roche, K., Huang, G. H., Munoz, B., and Rubin, G. S. (1999). Determination of risk factor associations with questionnaire outcomes: a methods case study. *American Journal of Epidemiology*, 150:1165–1178.

Bandeen-Roche, K., Miglioretti, D., Zeger, S. Z., and Rathouz, P. (1997a). Latent variable regression for multiple discrete outcomes. *The Journal of the American Statistical Association*, 92:1375–1386.

Bandeen-Roche, K., Miglioretti, D., Zeger, S. Z., and Rathouz, P. (1997b). Latent variable regression for multiple discrete outcomes. *The Journal of the American Statistical Association*, 92:1375–1386.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49:327–335.

Clogg, C. C. (1979). Some latent structure models for the analysis of likert-type data. *Social Science Research*, 8:287–301.

Clogg, C. C. (1980). Characterizing the class organization of labor market opportunity: A modified latent structure approach. *Sociological Methods and Research*, 8:243–272.

Clogg, C. C. (1995). *Latent Class Models*, chapter 6. Plenum Press, New York.

Clogg, C. C. and Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *The Journal of the American Statistical Association*, 79:762–771.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *The Journal of the American Statistical Association*, 83:173–178.

Eaton, W. W., Anthony, J. C., Gallo, J., Cai, G., Tien, A., Romanoski, A., Lyketsos, C., and Chen, L. S. (1997). Natural history of dis/dsm major depression: The baltimore eca follow-up. *Archives of General Psychiatry*, 54:993–999.

Faroane, S. V. and Tsuang, M. T. (1994). Measuring diagnostic accuracy in the absence of a "gold-standard". *American Journal of Psychiatry*, 151:650–657.

Formann, A. (1996). Latent class analysis in medical research. *Statistical Methods in Medical Research*, 5:179–211.

Gallo, J. J., Anthony, J. C., and Muthen, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology*, 49:251–264.

Garrett, E. S., Eaton, W. W., and Zeger, S. (2002). Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: A latent class model approach. *Statistics in Medicine*, 21:1289–1307.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics*, 56:1055–1067.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their application. *Biometrika*, 57:97–109.

Huang, G. H. and Bandeen-Roche, K. (2000). Latent variable regression

with covariate effects on underlying and measured variables: an approach of analyzing multiple polytomous surrogates. Submitted to *Psychometrika*.

Imperial College of Science, T. and Medicine (2000). *WinBugs version 1.3*. Cambridge, UK.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1997). Markov chain monte carlo in practice: A roundtable discussion. available via anonymous ftp at muskie.biostat.umn.edu in file /pub/1997/rr97-001.ps.Z.

Link, B. G., Lennon, M. C., and Dohrenwend, B. P. (1993). Socioeconomic status and depression: The role of occupations involving direction, control, and planning. *American Journal of Sociology*, 98:1351–1387.

MathSoft (1999). *Splus2000, Professional Release*. Cambridge, MA.

McCutcheon, A. L. (1987). *Latent Class Analysis*, volume 64 of *Quantitative Applications in the Social Sciences*. Sage Publications, Inc., London.

McLeod, J. D. and Kessler, R. C. (1990). Socioeconomic status differences in vulnerability to undesirable life events. *Journal of Health and Social Behavior*, 31:162–172.

Melton, B., Liang, K. Y., and Pulver, A. E. (1994). Extended latent class approach to the study of familial/sporadic forms of disease: Its application to the study of heterogeneity of schizophrenia. *Genetic Epidemiology*, 11:311–327.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. H. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 30:1087–1092.

Mooijaart, A. and Van der Heuden, P. G. M. (1992). The em algorithm for latent class analysis with equality constraints. *Psychometrika*, 57:261–269.

Muthen, L. K. and Muthen, B. O. (2001). *Mplus, version 2.01: The Comprehensive Modeling Program for Applied Researchers*. Los Angeles.

Pearlin, L. (1989). The sociological study of stress. *Journal of Health and Social Behavior*, 19:2–21.

Robins, L. N., Helzer, J. E., and Orvaschel, H. (1985). *The Diagnostic Interview Schedule*. Academic Press, New York.

Robins, L. N. and Regier, D. A., editors (1991). *Psychiatric Disorders in America - The Epidemiologic Catchment Area Study*. Free Press, New York.

Turner, R. J. and Lloyd, D. A. (1999). The stress process and the social distribution of depression. *Journal of Health and Social Behavior*, 40:374–404.

Turner, R. J., Wheaton, B., and Lloyd, D. A. (1995). Epidemiology of social stress. *American Sociological Review*, 60:104–125.

United States Census Bureau (1993). Current population survey 1993.

www.census.gov/hhes/poverty/threshld/thresh93.html.

Young, M. (1982-1983). Evaluating diagnostic criteria: A latent class paradigm. *Journal of Psychiatric Research*, 17:285–296.