10-21-2008

Multilevel Latent Class Models with Dirichlet Mixing Distribution

Chongzhi Di
Johns Hopkins University, cdi@jhsph.edu

Karen Bandeen-Roche
Johns Hopkins University, kbandeen@jhsph.edu

Multilevel Latent Class Models with Dirichlet Mixing Distribution

Chongzhi Di and Karen Bandeen-Roche

Abstract

Latent class analysis (LCA) and latent class regression (LCR) are widely used for modeling multivariate categorical outcomes in social sciences and biomedical studies. Standard analyses assume data of different respondents to be mutually independent, excluding application of the methods to familial and other designs in which participants are clustered. In this paper, we develop multilevel latent class model, in which subpopulation mixing probabilities are treated as random effects that vary among clusters according to a common Dirichlet distribution. We apply the Expectation-Maximization (EM) algorithm for model fitting by maximum likelihood (ML). This approach works well, but is computationally intensive when either the number of classes or the cluster size is large. We propose a maximum pairwise likelihood (MPL) approach via a modified EM algorithm for this case. We also show that a simple latent class analysis, combined with robust standard errors, provides another consistent, robust, but less efficient inferential procedure. Simulation studies suggest that the three methods work well in finite samples, and that the MPL estimates often enjoy comparable precision as the ML estimates. We apply our methods to the analysis of comorbid symptoms in the Obsessive Compulsive Disorder study. Our models' random effects structure has more straightforward interpretation than those of competing methods, thus should usefully augment tools available for latent class analysis of multilevel data.

KEYWORDS: Latent class analysis (LCA), Dirichlet distribution, multilevel models, EM algorithm, pairwise likelihood

1. INTRODUCTION

Latent class analysis (LCA; Clogg, 1995) and regression (LCR; Bandeen-Roche et al., 1997) are widely used in psychosocial, educational, and health research. These models treat a population of interest as being composed of several subpopulations, 1,..., M, to which subjects belong with probabilities $\pi_1, ..., \pi_M$. They also assume that responses of different subjects are independent of each other. However, this independence assumption may not be valid for commonly employed designs: for instance, in family studies, relatives may be more likely to fall into the same subpopulation, or 'class,' than members of different families. Application of the models to studies involving clustering of participants has been limited as a result.

In standard latent class models, the mixing probabilities $\pi = (\pi_1, \pi_2, ..., \pi_M)$ are assumed to be fixed parameters. Allowing these to vary randomly among clusters provides one mechanism for introducing intra-cluster dependence among responses. A class of models doing this has been proposed in recent years (Vermunt, 2003, 2008; henceforth "ML-V"). The models assume that class mixing, or 'membership,' probabilities $\underset{\sim}{u}_i = (u_{i1}, ..., u_{iM})$ vary over clusters according to unidimensional, normally distributed random effects v_i with unit variance:

$$\log \frac{u_{im}}{u_{i1}} = \alpha_{m0} + \lambda_m v_i \ \ , m = 2, ..., M. \tag{1}$$

The unidimensionality makes the approach computationally convenient. However, the random effects have latent factor interpretation that is contingent on 'loadings' λ and may therefore be somewhat obscure. Moreover, as we shall illustrate, this random effects structure sometimes has subtle undesirable implications. This paper alternatively considers models assuming Dirichlet-distributed mixing probabilities $\underset{\sim}{u}_i$. Like any parametric model, the Dirichlet distribution has implications for analytic interpretation; however, we believe its direct linking to the probability scale and freedom from loadings make it more natural and interpretable than the alternative, if clustering reasonably may be thought to induce exchangeable association. Moreover, as we shall demonstrate, conjugacy between the Dirichlet and multinomial distributions eases computation burden.

Because the Dirichlet is conjugate to the multinomial distribution, it was used in latent class models recently. For example, Potthoff et al. (2000) considered a latent-class type model in single level settings where the mixing probabilities for each individual are considered Dirichlet distributed random effects. Varki and Chintagunta (2004) proposed an augmented latent class model which is a mixture of standard LCA and Potthoff et al. (2000)'s model. However, none of these models considered multilevel setting in which individuals are clustered. To the best of our knowledge, there is no existing work that discusses multilevel latent class models with Dirichlet mixing distribution, although it is a natural choice. Estimation and Inference for such models are more complicated due to the multilevel structure. In this paper, we consider multilevel latent class models with Dirichlet mixing distribution, discuss estimation and inference using maximum likelihood and maximum pairwise likelihood methods and investigate the consequence of ignore clustering.

We were motivated to the present research by our collaboration in the Obsessive-Compulsive Disorder (OCD) study, a family-based study aiming to understand the comorbidity of OCD with other disorders. Obsessive-Compulsive Disorder is an anxiety disorder characterized by recurrent thoughts (obsessions) or repetitive behaviors (compulsions) which attempt to neutralize the obsessions (see, e.g, Jenike et al. 1990). A total of 999 subjects in 238 families were enrolled into this study, among which 706 subjects from 238 families were OCD cases. Diagnosis was made of 8 other disorders including major depression, generalized anxiety disorder, and panic disorder. It is hypothesized that there exist subtypes of OCD, based on comorbidity (Nestadt et al., 2003). Latent class analysis is a natural tool for evaluating this hypothesis; however, the clustering within families must be taken into account if correct and efficient inference is to be made. It is also of interest to estimate the subtype heritability: in statistical terms, the intra-cluster correlation among class memberships.

This paper develops multilevel latent class analysis (MLCA) models with Dirichlet mixing distribution, proposes model fitting using both maximum likelihood and maximum pairwise likelihood methods, and discusses issues of application including missing data and model selection. We also investigate the use of simple latent class model by ignoring clustering. We

evaluate methods' performance in a simulation study and in application to the OCD study.

2. MULTILEVEL LATENT CLASS ANALYSIS: MODELS

Latent class models typically involve vector data per individual, comprising multiple categorical 'item' responses. Though these handle categorical responses in general, for simplicity of notation we primarily consider binary data. Then let Y_{ijk} denote the response of the j^{th} subject of the i^{th} cluster on the k^{th} item; $i = 1, 2, ..., n$; $j = 1, 2, ..., n_i$; $k = 1, 2, ..., K$. We denote the K-vector of a subject's responses by $\underset{\sim}{Y}_{ij}$. Let η_{ij} denote the class membership for subject j in cluster i, taking values in $\{1, 2, ..., M\}$, and $\underset{\sim}{\eta}_i = \{\eta_{i1}, \eta_{i2}, ..., \eta_{in_i}\}$. Latent class analysis (LCA) ignoring clustering decomposes the mass function of a subject's item responses as

$$\begin{aligned}
\Pr(\underset{\sim}{Y}_{ij} = \underset{\sim}{y}) &= \sum_{m=1}^{M} \Pr(\eta_{ij} = m) \Pr(\underset{\sim}{Y}_{ij} = \underset{\sim}{y} \,|\, \eta_{ij} = m) \\
&= \sum_{m=1}^{M} \Pr(\eta_{ij} = m) \cdot \prod_{k=1}^{K} \Pr(Y_{ijk} = y_k \,|\, \eta_{ij} = m) \qquad (2) \\
&= \sum_{m=1}^{M} \pi_m \cdot \prod_{k=1}^{K} p_{km}^{y_k} (1 - p_{km})^{1-y_k} \,,
\end{aligned}$$

where $\pi_m = \Pr(\eta_{ij} = m)$ is the prevalence of class m and $p_{km} = \Pr(Y_{ijk} = 1 \,|\, \eta_{ij} = m)$ is the conditional probability of positive response for item k if the subject belongs to class m. Typically, LCA imposes the "conditional independence" assumption (as revealed in the equations above) that a subject's responses on the items are independent given his class membership (Clogg, 1995). The conditional probabilities define the "measurement" part of the model. They are often parameterized in logit scale, i.e, $\beta_{km} = \text{logit}(p_{km}) = \log(p_{km}/(1 - p_{km}))$. The distribution of classes in the population defines the "mixing" part of the model, which involves parameters $(\pi_1, \pi_2, \cdots, \pi_m)$. LCA assumes that the mixing distribution is the same for every individual.

Model (2) does not account for potential correlation among response vectors of subjects within the same cluster. To rectify this, we assume that class mixing probabilities $\{\pi_1, \pi_2, ..., \pi_M\}$ vary from cluster to cluster, arising from a common Dirichlet distribution.

Thus, we consider there to be between-cluster heterogeneity in the probabilities governing underlying outcome status (class membership), and not additionally in the item response distribution given class membership. Modeling class membership probabilities as Dirichlet straightforwardly expresses clusterwise heterogeneity: in the OCD example, probabilities of having each type of comorbidity may vary from family to family. It also explicitly acknowledges classes as competing, such that membership in one class precludes membership in another. Clustering is accounted for in the sense that subjects from the same cluster are more likely to fall into same classes since they share the same cluster specific random effects.

The multilevel latent class model that we consider in this paper is formulated as:

$$\begin{cases} \Pr(\underset{\sim}{Y}_{ij} = \underset{\sim}{y}) = \sum_{m=1}^{M} \Pr(\eta_{ij} = m) \cdot \prod_{k=1}^{K} p_{km}^{y_k}(1-p_{km})^{1-y_k} \\ \Pr(\eta_{ij} = m \,|\, \underset{\sim}{u}_i) = u_{im} \\ \underset{\sim}{u}_i \sim \text{Dirichlet}(\alpha_1, \alpha_2, ..., \alpha_M)\,, \end{cases} \tag{3}$$

where $\underset{\sim}{u}_i = (u_{i1}, ..., u_{iM})$ are cluster specific class mixing probabilities. For convenience, we model the conditional probabilities in the logit scale using $\beta_{km} = \text{logit}(p_{km}) = \log(p_{km}/(1-p_{km}))$, and then the natural parameters in the model are $\theta = (\beta, \alpha)$. Implicit in (3) is that all distributions are assumed to be functionally independent of n_i. The Dirichlet distribution is natural for random effects, since the u_{im}'s are non-negative probabilities constrained to sum to 1. Model (3) implies that in the whole population, the marginal prevalences of classes are $(\alpha_1/\alpha_0, \alpha_2/\alpha_0, ..., \alpha_M/\alpha_0)$, where $\alpha_0 = \sum_{m=1}^{M} \alpha_m$. The variance of the cluster-specific prevalences $\underset{\sim}{u}_i$ varies inversely with the scale parameter, α_0, such that the correlation in same-class membership between same-cluster members is

$$\rho := \text{Corr}\{I(\eta_{ij} = m), I(\eta_{ik} = m)\} \;=\; \frac{1}{\alpha_0 + 1}.$$

Here ρ can be interpreted as an intra-cluster correlation (ICC) coefficient, i.e, heritability. The model implicitly assumes that the ICCs for same-class membership are class-invariant. Appendix A provides formulas for odds ratios for association among both same-class mem-

bership between same-cluster members and different-class membership between same-cluster members, as well as other implications of Model (3).

The measurement part of Model (3) is the same as for the simple LCA model. The conditional independence assumption is retained, i.e, responses on different items are assumed to be independent given class membership. The possible clustering effect is reflected in the mixing part, that is, potential associations among the class membership indicators $\{\eta_{ij} : i = 1, 2, ..., n; j = 1, ..., n_i\}$. The Dirichlet random effects structure specifies the joint distribution of class membership vector $\underset{\sim}{\eta}_i = \{\eta_{i1}, \eta_{i2}, ..., \eta_{in_i}\}$ for cluster i as:

$$\Pr(\underset{\sim}{\eta}_i = \underset{\sim}{z}) = \int \prod_{j=1}^{n_i} \Pr(\eta_{ij} = z_j | \underline{u}_i)\, f(\underline{u}_i)\, d\underline{u}_i = \frac{\Gamma(\alpha_0 + n_i)}{\Gamma(\alpha_0)} \prod_m \frac{\Gamma(\alpha_m + q_m^{(i)})}{\Gamma(\alpha_m)} ,$$

where $q_m^{(i)} = \sum_j I(z_j = m)$, the number of subjects from cluster i that belong to class m. This has nice analytic form because multinomial and Dirichlet distributions are conjugate families, and eases implementation and interpretation. In Model (1), in contrast, $\Pr(\underset{\sim}{\eta}_i = \underset{\sim}{z})$ does not have a closed form. Exchangeable within cluster association is implied by both models, meaning the sets of associations among class memberships for any two subjects from the same cluster are the same.

3. ESTIMATION AND INFERENCE: MAXIMUM LIKELIHOOD

For the multilevel latent class model (3), the complete likelihood contributed by cluster i is

$$\begin{aligned} L_i^c(\beta, \alpha) &= \Pr(\underset{\sim}{Y}_i | \underset{\sim}{\eta}_i) \Pr(\underset{\sim}{\eta}_i | \underline{u}_i)\, f(\underline{u}_i) \\ &= \prod_j \left[\prod_k \Pr(Y_{ijk} | \eta_{ij}) \Pr(\eta_{ij} | \underline{u}_i) \right] f(\underline{u}_i). \end{aligned}$$

Since the latent variables $\underset{\sim}{\eta}_i$ and $\underline{u}_i$ are unobservable, the observed likelihood is obtained by marginalizing the complete likelihood over them, i.e,

$$\begin{aligned} L_i(\beta, \alpha) &= \iint \Pr(\underset{\sim}{Y}_i | \underset{\sim}{\eta}_i) \Pr(\underset{\sim}{\eta}_i | \underline{u}_i)\, f(\underline{u}_i)\, d\underline{u}_i\, d\underset{\sim}{\eta}_i \\ &= \sum_{\underset{\sim}{z}} \left[\prod_j \prod_k \Pr(Y_{ijk} | \eta_{ij} = z_j) \right] \cdot \Pr(\underset{\sim}{\eta}_i = \underset{\sim}{z}) , \end{aligned}$$

where the marginal distribution of $\underset{\sim}{\eta}_i$ is as specified in the previous section.

3.1 Estimation: EM algorithm

The EM (Expectation-Maximization) algorithm (Dempster et al., 1977) well suits the incompletely observed nature of mixture models. Provided a set of regularity conditions (e.g, in Dempster et al 1977) which are met in our model, it is stable and ensures that the likelihood monotonely increases over iterations. For these reasons, we propose to use the EM algorithm for estimation.

For the E step, take the parameter estimates as $\beta^{(h)}, \alpha^{(h)}$ at the h^{th} iteration. Then, we need to calculate the expected value of the log complete likelihood:

$$\begin{aligned} Q(\beta,\alpha;\beta^{(h)},\alpha^{(h)}) &= E_{(\beta^{(h)},\alpha^{(h)})}\left[\sum_i \log\, L_i^c(\beta,\alpha)\,|\,Y_i;\,\beta^{(h)},\alpha^{(h)}\right] \\ &= \sum_i\sum_j\sum_m w_{ijm}\,U_{ijm}(\beta) + (\textstyle\sum_i v_i^T)\alpha \\ &\quad + n\,[\,\log\,\Gamma(\textstyle\sum_m \alpha_m) - \sum_m \log\,\Gamma(\alpha_m)\,] + Constant \end{aligned} \tag{4}$$

where $U_{ijm} = \log \Pr(\underset{\sim}{Y}_{ij}|\,\eta_{ij} = m) = \sum_k \log \Pr(Y_{ijk}|\eta_{ij} = m)$. The weights w_{ijm} and $v_i^T = (v_{i1}, v_{i2}, ..., v_{iM})$ are

$$\begin{aligned} w_{ijm} &= \Pr(\eta_{ij} = m|\,\underset{\sim}{Y}_i;\beta^{(h)},\alpha^{(h)}) \\ v_{im} &= E[\,\log(u_{im})|\,\underset{\sim}{Y}_i;\beta^{(h)},\alpha^{(h)}\,]\,. \end{aligned}$$

These are the only places where the current parameter estimates $\beta^{(h)}$, $\alpha^{(h)}$ enter the Q function.

To obtain weights w_{ij}, we need the posterior distribution of $\underset{\sim}{\eta_i}$ given Y_i, which can be calculated by Bayes' rule,

$$\Pr(\underset{\sim}{\eta_i}|\underset{\sim}{Y}_i) = \frac{\Pr(\underset{\sim}{\eta_i})\,\prod_j\prod_k \Pr(Y_{ijk}|\eta_{ij})}{\sum_{\underset{\sim}{\eta_i}}\left[\Pr(\underset{\sim}{\eta_i})\,\prod_j\prod_k \Pr(Y_{ijk}|\eta_{ij})\right]}$$

Here the sum above is taken over all possible class membership combinations for cluster i, totaling M^{n_i} possibilities. To obtain weights v_i, we use the double expectation technique $v_{im} = E\{\,E[\,log(u_{im})|\,\underset{\sim}{Y}_i,\underset{\sim}{\eta_i};\beta^{(h)},\alpha^{(h)}\,]\,|\underset{\sim}{Y}_i;\beta^{(h)},\alpha^{(h)}\,\}$. Here,

$$E[\,\log(u_{im})|\,\underset{\sim}{Y}_i,\underset{\sim}{\eta_i};\beta^{(h)},\alpha^{(h)}\,] = D\Gamma(\alpha_m^{(h)} + q_m^{(i)}) - D\Gamma(\textstyle\sum_m \alpha_m^{(h)} + n_i)$$

where $D\Gamma(x) = \frac{d}{dx}\log\,\Gamma(x)$, and $q_m^{(i)} = \sum_j I(\eta_{ij} = m)$, the number of subjects from cluster i belonging to class m; see Appendix B. We then take expectation conditional on Y_i to obtain

the v_i's, which again involves summing over M^{n_i} possible patterns of class memberships in cluster i.

Once we obtain the Q function as in equation (4), the M step is relatively straightforward. The β parameters appear only in the first term of (4), and the α parameters appear only in the second and third terms. Maximization over β is equivalent to fitting a logistic regression model with weights w_{ij}. Thus in practice, we can conveniently call any routine that fits weighted logistic regression for this part of the M step. The first and second derivatives with respect to α are:

$$\begin{aligned} \frac{\partial Q}{\partial \alpha} &= \sum_i v_i + n\,[D\Gamma(\alpha_0)\,\mathbf{1_{M\times 1}} - D\Gamma(\alpha)\,]\,, \\ \frac{\partial^2 Q}{\partial \alpha\,\partial \alpha'} &= n\,[T\Gamma(\alpha_0)\,\mathbf{1_{M\times M}} - diag\,(T\Gamma(\alpha_1),\,\ldots,\,T\Gamma(\alpha_M))\,]\,, \end{aligned}$$

where $T\Gamma(x) := \frac{\partial^2}{\partial x^2}\log\Gamma(x)$. Thus, we can carry out a one or multi-step Newton-Raphson algorithm for this part of the M step. The cross-derivative $\frac{\partial^2 Q}{\partial\beta\,\partial\alpha'}$ is 0, so the two parts can be carried out separately.

Finally, we iterate between the E step and M step until a suitable convergence criterion is met.

3.2 Missing Data

By using the EM algorithm, we can conveniently deal with data that are missing at random (MAR) in the sense of Little and Rubin (2002). Let M_{ijk} be the missing indicator for Y_{ijk}, i.e, $M_{ijk} = 1$ if Y_{ijk} is missing (hence we denote Y_{ijk}^{miss}) and $M_{ijk} = 0$ otherwise (hence we denote Y_{ijk}^{obs}). If Y_{ijk} is observed, its contribution to the complete log likelihood and the Q function are

$$\begin{aligned} &\sum_{m=1}^{M} I(\eta_{ij} = m)\log\Pr(Y_{ijk}^{obs}\,|\,\eta_{ij} = m) \\ &\sum_{m=1}^{M} w_{ijm}\left[\,Y_{ijk}^{obs}\log p_{km} + (1 - Y_{ijk}^{obs})\log(1 - p_{km})\,\right] \end{aligned} \tag{5}$$

respectively, where $p_{km} = \Pr(Y_{ijk} = 1|\eta_{ij} = m) = exp(\beta_{km})/\{1 + exp(\beta_{km})\}$. If Y_{ijk} is missing, its contribution to the complete log likelihood is

$$\sum_{m=1}^{M} I(\eta_{ij} = m) \log \Pr(Y_{ijk}^{miss} \,|\, \eta_{ij} = m)\,,$$

and its contribution to the Q function is

$$\begin{aligned} & E\left[\sum_{m=1}^{M} I(\eta_{ij} = m) \log \Pr(Y_{ijk}^{miss} \,|\, \eta_{ij} = m)\,|Y_i\,; \beta^{(h)}, \alpha^{(h)}\right] \\ = & \sum_{m=1}^{M} w_{ijm}\left[p_{km}^{(h)} \log p_{km} + (1 - p_{km}^{(h)}) \log(1 - p_{km})\right]\,, \end{aligned} \tag{6}$$

where $p_{km}^{(h)} = \Pr(Y_{ijk} = 1|\eta_{ij} = m; \beta^{(h)}) = exp(\beta_{km}^{(h)})/\{1 + exp(\beta_{km}^{(h)})\}$ is the probability of a positive response in the current iteration.

We can see that if the response Y_{ijk} is missing, the EM algorithm "imputes" it based on current knowledge, i.e, $Y_{ijk}^{miss} = 1$ with probability $p_{km}^{(h)}$ and $Y_{ijk}^{miss} = 0$ with probability $1 - p_{km}^{(h)}$ for a member of the m^{th} class. Only the first term of the Q function changes when the data are missing.

3.3 Inference and Prediction

We use the observed Fisher information matrix to estimate the standard errors of the estimated parameters. The EM algorithm does not directly provide the Hessian matrix of log likelihood; rather, methods are available to estimate it from EM outputs, e.g. Louis (1982). For our problem the application of such methods is computationally complex. Instead, we numerically calculate the observed Fisher information matrix following Oakes (1999):

$$\frac{\partial \log\, L(\theta)}{\partial \theta} = \left[\frac{\partial Q(\psi;\theta)}{\partial \psi}\right]|_{\psi=\theta} \tag{7}$$

$$\frac{\partial^2 \log\, L(\theta)}{\partial \theta\, \partial \theta'} = \left[\frac{\partial^2 Q(\theta;\psi)}{\partial \theta\, \partial \theta'} + \frac{\partial^2 Q(\theta;\psi)}{\partial \theta\, \partial \psi'}\right]|_{\psi=\theta} \tag{8}$$

where $\theta = (\beta, \alpha)$ are the parameters and $\psi = (\beta^{(h)}, \alpha^{(h)})$ are the current estimates. The technical details may be found in Appendix C.

As a by-product of the EM algorithm, we can easily obtain best predictions of the random effects given the data, which includes both latent class memberships η_{ij} and cluster-specific class prevalence u_i. The posterior probabilities of class membership for each subject, $\Pr(\eta_{ij} = m|\, \underset{\sim}{Y}_i; \beta, \alpha), m = 1, ..., M$, are calculated as weights in the E-step. As for the cluster-specific

random effect $\underset{\sim}{u}_i = (u_{i1}, ..., u_{iM})$, the best prediction is the posterior expectation, with the m^{th} component provided by

$$E[\, u_{im} |\, \underset{\sim}{Y}_i; \beta, \alpha\,] = E\{\, E[\, u_{im} |\, \underset{\sim}{Y}_i, \underset{\sim}{\eta}_i; \beta, \alpha\,]\, |\underset{\sim}{Y}_i; \beta, \alpha\,\}.$$

Appendix B shows that the inner expectation is $(\alpha_m + q_m^{(i)})/(\sum_m \alpha_m + n_i)$, since $[\, \underset{\sim}{u}_i \,|\, \underset{\sim}{\eta}_i, \tilde{Y}_i\,]$ is Dirichlet-distributed with parameter $(\alpha_1 + q_1^{(i)}, ..., \alpha_M + q_M^{(i)})$. We then marginalize over all possible patterns of $\underset{\sim}{\eta}_i$ to obtain the outer expectation.

3.4 Selecting the number of classes

To select among models with different numbers of classes is widely considered as a challenging problem. Even in latent class models without clustering, the likelihood ratio test comparing an M-class model and an (M+1)-class model does not follow the typical χ^2 distribution, because under the null hypothesis, some parameters lie on the boundary of the parameter space, or may be not identifiable. Instead, the AIC (Akaike Information Criterion, Akaike, 1974) and BIC (Bayesian Information Criterion, Schwarz, 1978) and similar statistics have been widely used for selecting among models. In the multilevel latent class model, appropriate specification of AIC and BIC is challenged by the random effects structure. Thus, we recommend an alternative method for model selection, based on marginalizing model (3). That marginalization yields

$$\begin{aligned}
\Pr(\underset{\sim}{Y}_{ij} = \underset{\sim}{y}) &= \sum_{m=1}^{M} \Pr(\eta_{ij} = m) \cdot \Pr(\underset{\sim}{Y}_{ij} = \underset{\sim}{y} \,|\, \eta_{ij} = m) \\
&= \sum_{m=1}^{M} \int \Pr(\eta_{ij} = m \,|\, \underset{\sim}{u}_i) f(\underset{\sim}{u}_i)\, d\, \underset{\sim}{u}_i \cdot \prod_{k=1}^{K} \Pr(Y_{ijk} = y_k \,|\, \eta_{ij} = m) \\
&= \sum_{m=1}^{M} \frac{\alpha_m}{\alpha_0} \cdot \prod_{k=1}^{K} p_{km}^{y_k} (1 - p_{km})^{1-y_k} \\
&= \sum_{m=1}^{M} \pi_m^* \cdot \prod_{k=1}^{K} p_{km}^{y_k} (1 - p_{km})^{1-y_k}, \qquad (9)
\end{aligned}$$

so that the marginal distribution of a single subject's response vector is a simple latent class model with the same number of classes as the MLCA model. This relationship suggests a simple method for selecting the number of classes: randomly choose one subject per cluster,

and then apply latent class analysis on the resulting subsample. Standard methods, such as BIC, could then be used to choose the number of classes. Finally, one would fix the number of classes, M, in a subsequent multilevel latent class model.

The method just outlined may lose precision since only a subset of the data is used. Instead, we propose to randomly draw multiple mutually independent subsamples, resulting in the following algorithm:

1. Draw a subsample $S = \{(i, j_i^*) : i = 1, ..., n\}$, where j_i^* is a subject randomly chosen from subjects $\{1, ..., n_i\}$ in cluster i.

2. Fit a latent class model using sample S and obtain the BIC (or other model selection criterion) statistics for all candidate models with $\{1, ..., M^*\}$ classes.

3. Repeat steps 1-2 to get L such random subsamples. Record $\{\text{BIC}_l^{(m)} : l = 1, ..., L, m = 1, ..., M^*\}$, where $\text{BIC}_l^{(m)}$ is the BIC for m-class model using the l^{th} subsample.

4. Choose the model with the smallest average BIC statistic, i.e, M=arg min $\{\bar{BIC}^{(m)} : m = 1, ..., M^*\}$, where $\bar{BIC}^{(m)} = \sum_l \text{BIC}_l^{(m)}/L$.

Step 4 is justified under weak law regularity conditions so long as the model selection statistic has additive form: then, the average estimates the same limiting quantity as the original statistic.

For standard LCA models, it is known that the BIC would consistently choose the right model in large samples (Haughton, 1988). The proposed BIC method has the same asymptotic property for multilevel data as the usual BIC method. However, the subsampling creates a different finite sample tradeoff, especially when the sample size is small to medium. It is known that BIC may underestimates the number of classes in small samples (Yang, 2006). In such cases, one could use the modified BIC with sample size adjustment (Sclove, 1987), which is shown to perform better according to Yang (2006). We suggest that the method described here be used with caution. Model selection on the number of classes for multilevel latent class models is a complex problem. A comprehensive study on this issue would be possible future research directions and is out of scope of this paper.

4. ESTIMATION AND INFERENCE: MAXIMUM PAIRWISE LIKELIHOOD

In computing weights for the EM Q function (4), the computational burden increases exponentially ($O(n \cdot M^J)$) with the number of classes M and cluster size $J := max\{n_i, i = 1, ..., n\}$. Thus, we recommend using EM fitting when both M and n_i are relatively small, and otherwise using the maximum pairwise likelihood approach we now propose.

4.1 Pairwise likelihood

The idea of using pairwise likelihood for clustered data is not new. Particularly when clusters present complex (e.g. spatial) correlation structure, the joint likelihood may be difficult to specify. Even if we can specify the joint likelihood, maximizing it may be computationally complicated (e.g. with large clusters), and inferences may be sensitive to the model assumptions. The pairwise likelihood approach nicely overcomes these difficulties. Pairwise likelihood falls within the general concept of "composite likelihood" (Lindsay, 1988), which has been used for a variety of correlated data problems (Nott and Ryden, 1999; Kuk and Nott, 2000; Cox and Reid, 2004; Renard et al., 2004; Varin et al., 2005).

In the multilevel latent class setting, rather than specifying the joint distribution for each cluster, we specify only pairwise distributions and then take the product over all possible pairs:

$$L^p(\beta, \alpha) = \prod_{i=1}^{n} L_i^p(\beta, \alpha) = \prod_{i:\, n_i>1} \left[\prod_{j_1<j_2} \Pr(Y_{ij_1}, Y_{ij_2}; \beta, \alpha) \right] \cdot \prod_{i:\, n_i=1} \Pr(Y_i;\, \beta, \alpha) \qquad (10)$$

where

$$\Pr(Y_{ij_1}, Y_{ij_2}; \beta, \alpha) = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \Pr(Y_{ij_1} \,|\, \eta_{ij_1} = m_1 \,; \beta) \Pr(Y_{ij_2} \,|\, \eta_{ij_2} = m_2 \,; \beta) \cdot \Pr(\eta_{ij_1} = m_1, \eta_{ij_2} = m_2; \alpha)$$

and

$$\Pr(Y_i; \beta) = \sum_{m=1}^{M} \Pr(Y_i \,|\, \eta_i = m;\, \beta) \Pr(\eta_i = m; \alpha) \text{ , for i in } \{i : n_i = 1\}$$

Under typical regularity conditions, the pairwise likelihood estimate, which maximizes L^p, is the solution to the pairwise score equation

$$\frac{\partial \log L^p}{\partial \theta} = \frac{\partial \log L_i^p}{\partial \theta} = \sum_{i:n_i>1} \sum_{j_1<j_2} \frac{\partial \log L_{ij_1j_2}}{\partial \theta} + \sum_{i:n_i=1} \frac{\partial \log L_i}{\partial \theta} = 0.$$

As pointed out by Lindsay (1988), each component in (10) is a likelihood function, and the corresponding score function is unbiased provided correct pairwise specification. Thus, the first derivative of the pairwise likelihood is an unbiased estimating function. Hence assuming that the number of clusters (with at least two subjects) goes to infinity, the pairwise likelihood estimators for both β and α must be consistent and asymptotically normal with variance

$$\left\{ E_\theta \left[\frac{\partial^2 \log L^p}{\partial\theta\partial\theta'} \right] \right\}^{-1} \cdot E_\theta \left\{ \left[\frac{\partial \log L^p}{\partial\theta} \right]^T \left[\frac{\partial \log L^p}{\partial\theta} \right] \right\} \cdot \left\{ E_\theta \left[\frac{\partial^2 \log L^p}{\partial\theta\partial\theta'} \right] \right\}^{-1}.$$

The asymptotic variance can be consistently estimated by the "sandwich" variance estimator (Royall, 1986) that replaces the expectations in the above formula with empirical estimates. After we obtain parameter estimates, prediction for both cluster-specific and subject-specific random effects follows by similar methods as described in Section 3.3.

4.2 Estimation: Pairwise EM algorithm

We can view the pairwise likelihood in another way. If we think of "pseudo-data" comprised of the pairs, and assume the pairs' responses are mutually independent, then the pairwise likelihood is exactly the joint likelihood of the "pseudo-data". This connection enables us to modify the EM algorithm in Section 3 to maximize the pairwise likelihood. We call this algorithm Pairwise EM (PEM). Under typical regularity conditions and suitable conditions on the missing data mechanism stated below, PEM shares similar properties with EM, for example, the ascent property and linear rate of convergence. The essential reason is that each pairwise likelihood component satisfies the information inequality,

$$E_{\theta_0} \left[\log \frac{f(\underset{\sim}{Y}_{ij}^{full}, \underset{\sim}{Y}_{ik}^{full}, \eta_{ij}, \eta_{ik}; \theta)}{f(\underset{\sim}{Y}_{ij}^{obs}, \underset{\sim}{Y}_{ik}^{obs}; \theta)} \, | \underset{\sim}{Y}_{ij}^{obs}, \underset{\sim}{Y}_{ik}^{obs} \right] \leq E_{\theta_0} \left[\log \frac{f(\underset{\sim}{Y}_{ij}^{full}, \underset{\sim}{Y}_{ik}^{full}, \eta_{ij}, \eta_{ik}; \theta_0))}{f(\underset{\sim}{Y}_{ij}^{obs}, \underset{\sim}{Y}_{ik}^{obs}; \theta_0))} \, | \underset{\sim}{Y}_{ij}^{obs}, \underset{\sim}{Y}_{ik}^{obs} \right],$$

thus does the whole pairwise likelihood by additivity of expectation. The ascent property for PEM follows by an analogous argument to that which proves the ascent property for EM (Dempster et al., 1977).

The PEM can handle missing data conveniently, similarly as the EM. However, it requires stricter assumptions on the missing data mechanism to ensure consistency. One set

of conditions sufficient to ensure the information inequality is that the data be missing at random (MAR) and that the missing distribution have no more than second-order pairwise dependence. Equivalently, the needed assumption is that, conditional on one's own observed data and that of each single family member, missingness is independent of all other family members' observed data as well as data not observed. Under such conditions, the information inequality holds, and thus the validity of PEM is justified.

4.3 Comparison: MPL vs. ML

The pairwise likelihood (MPL) approach has both advantages and disadvantages compared to the ML approach. MPL relies only on bivariate distributional assumptions rather than those for the full distribution, thus is more robust than ML. On the other hand, the asymptotic efficiency of MPL can be no better than for ML, and may be worse if the true joint distribution is correctly specified.

In terms of computational burden, MPL has a clear advantage over ML, since each pseudo-data cluster contains at most two subjects. The computational complexity is $O(n \cdot M^2 J(J-1)/2)$ for MPL, as opposed to $O(n \cdot M^J)$ for ML. Table 1 displays the ratio of computational complexity comparing ML to MPL. When the number of classes or the cluster size is less than 5, the difference in computational burden may still be acceptable. However, the improvement of MPL is huge if the cluster size is greater than 5 and the number of classes is greater than 3. For instance, ML requires 146 times computations as MPL does to fit a four class model with cluster size 8. For the OCD example, the cluster sizes range from 1 to 10. It takes approximately 3 hours to fit a three class model using ML, compared to 30 minutes using MPL.

[Table 1 about here.]

Finally, though MPL is proposed to reduce computational burden for the Dirichlet model (3), it is not restricted to this model. In fact, we can assume any bivariate distributions for pairs (η_{ij}, η_{ik}), and still use the MPL method to obtain consistent estimation.

5. COMPARISON WITH SIMPLE LATENT CLASS ANALYSIS

It is of interest how simple latent class analysis performs when it is incorrectly applied to multilevel data, as is the performance of multilevel latent class analysis for independent subjects. To investigate the former, we consider a more general class of models than model (3), i.e. the semiparametric model

$$\begin{cases} \Pr(\underset{\sim}{Y}_{ij} = \underset{\sim}{y}) = \sum_{m=1}^{M} \Pr(\eta_{ij} = m) \cdot \prod_{k=1}^{K} p_{km}^{y_k}(1-p_{km})^{1-y_k} \\ \underset{\sim}{\eta}_i \sim f(\underset{\sim}{\eta}_i; \underset{\sim}{\pi}^*, \alpha^*) \\ \Pr(\eta_{ij} = m\,) = \pi_m^* \end{cases} \tag{11}$$

where the joint distribution $f(\underset{\sim}{\eta}_i; \alpha^*)$ is unspecified but subject to the constraint that each subject belongs to class m with probability π_m^* marginally. Both model (3) and the ML-V model are parametric submodels of the general model (11). The following result implies that the application of ML to the simple latent class model consistently estimates π_m^* and β parameters.

Proposition 1. Assume that $\{\underset{\sim}{Y}_{ij} : i = 1, ..., n; j = 1, ..., n_i\}$ are generated from the semiparametric model (11). Let $(\tilde{\beta}, \tilde{\pi})$ denote the maximum likelihood estimators from the simple latent class model (2), and let $l^S(\theta_1) := l^S(\beta, \pi) := \sum_i \sum_j \log f(\underset{\sim}{Y}_{ij})$ denote the log likelihood function from it. Then under suitable regularity conditions

1. $\mathrm{E}\{\frac{\partial\, l^S}{\partial \beta}; \beta, \alpha^*\} = 0\,, \mathrm{E}\{\frac{\partial\, l^S}{\partial \pi}; \beta, \alpha^*\} = 0\,;$
2. As $n \to \infty$, $\tilde{\beta} \xrightarrow{P} \beta, \tilde{\pi} \xrightarrow{P} \pi^*$;
3. $\sqrt{n} \begin{pmatrix} \tilde{\beta} - \beta \\ \tilde{\pi} - \pi^* \end{pmatrix} \xrightarrow{D} N(0, \Sigma)$, where $\Sigma := \left\{ E_\theta \left[\frac{\partial^2 l^S}{\partial \theta_1 \partial \theta_1'} \right] \right\}^{-1} \cdot E_\theta \left\{ \left[\frac{\partial\, l^S}{\partial \theta_1} \right]^2 \right\} \cdot \left\{ E_\theta \left[\frac{\partial^2 l^S}{\partial \theta_1 \partial \theta_1'} \right] \right\}^{-1}$.

The proof is given in Appendix D. Since our MLCA model (3) is a parametric submodel of the general semiparametric model (11), the results of Proposition 1 apply to it as a corollary.

Remark 1a. Proposition 1 suggests an alternative correct inference procedure if the goal is to understand the measurement model and the average proportions of the subclasses:

one can simply fit the simple latent class model and fix the standard errors by the sandwich estimator. This method is simple and fast to implement, compared to the two methods developed above. However, it may suffer some loss of efficiency when the multilevel model is true. Moreover it does not provide a measure of within cluster association.

Remark 1b. There is an important connection with marginal modeling for longitudinal or clustered data. If we ignore the measurement part of the model, the latent class η_{ij}'s are clustered data, correlated within clusters. The simple latent class model corresponds to a marginal model for η_{ij}'s with working independence correlation, while our multilevel latent class model corresponds to a marginal model with working exchangeable correlation. Similarly as with generalized estimating equations (GEE, Liang and Zeger 1986), even if the working correlation is misspecified as independence, the estimators of marginal parameters π_m^*'s are consistent, and their standard errors can be consistently estimated using the robust variance estimator. Moreover, Proposition 1 indicates that the measurement model parameters (β's) can also be consistently estimated under such model misspecification.

Remark 1c. If the within cluster association is of interest, or higher efficiency is needed, the simple latent class model would not be appropriate. A parametric submodel, such as the ML-V model and our Dirichlet model, provides one solution when the parametric assumptions are reasonable, but robustness no longer holds generally. Alternatively, one might make the second moment assumptions for $\underset{\sim}{\eta}_i$ in addition to the semiparametric model (11), that is, specify a model for bivariate distributions $g(\eta_{ij}, \eta_{ik}; \pi^*, \alpha^*)$ for every possible pair of latent class indicators (η_{ij}, η_{ik}). Two methods might be utilized for estimation of this general semiparametric model. The first method is the MPL approach. As we pointed out in Section 4, MPL provides consistent estimates as long as the first two moments for $\underset{\sim}{\eta}_i$ are correctly specified. The second possible method is an estimating equation approach, for instance, mimicking that of Reboussin et al. (1999).

6. SIMULATION

We evaluated the finite sample performance of our procedure in simulation studies. Data were generated from the following true settings: $n = 200$ or 500 clusters, $J = 4$ subjects per cluster, $K = 5$ items, $M = 2$ classes. The true model was the multilevel latent class model (3) with true parameters values chosen randomly. The true α parameters were $(1.5, 2.3)$. The log odds of reporting "1" for class 1 members ($\beta_{\cdot 1}$) were $(-1.21, 0.28, 1.08, -2.35, 0.43)$ for five items, respectively, and the log odds for class 2 members ($\beta_{\cdot 2}$) were $(0.51, -0.57, -0.55, -0.56, -0.89)$. We conducted 1000 simulation runs, and in each run three methods were used to fit the multilevel latent class model, maximum likelihood for Dirichlet model (ML), maximum pairwise likelihood for Dirichlet model (MPL), and maximum likelihood for simple latent class model with robust standard errors (ML-S).

First we consider findings for estimation of the measurement models. Figures 1 display boxplots of estimated β parameters in 1000 runs using 200 clusters. The solid lines in each figure represent true parameter values. For each method and parameter, estimator distributions centered closely around true values, exhibited relatively small dispersion, and included few outliers. The dispersion of MPL was similar to that of ML, suggesting high relative efficiency of the MPL estimates. The dispersion of ML-S, however, was larger than that for ML or MPL, implying loss of efficiency by ignoring the within cluster correlation. As shown in Table 2, the loss in efficiency was about 10-20% on average, and up to 40% for some parameters. Simulation results using 500 clusters displayed similar patterns, but with narrower confidence intervals. In summary, the β parameters were well estimated by both ML and MPL methods based on the Dirichlet model, and the simple latent class model estimators were consistent, but generally less efficient.

[Figure 1 about here.]

[Table 2 about here.]

Turning to findings relating to the mixing distribution, the distributions of the α parameter estimates were widely dispersed and exhibited heavy tails (Figure 2). Researchers

typically will be more interested in conveniently interpreted transformations of the α parameters, including the population-average class prevalences $(\alpha_1/\alpha_0, ..., \alpha_M/\alpha_0)$, the random effects scale parameter $\alpha_0 = \alpha_1 + ... + \alpha_M$, and the intra-cluster correlation parameter ρ. Figure 2 shows that the population-average class prevalences and the intra-cluster correlation were well estimated, with distributions centering around the true values and having narrow spreads. Estimates of the scale parameter α_0 exhibited substantial variability, as is often the case for variance components. Finally, MPL estimates for α parameters enjoyed high finite-sample efficiency compared to ML estimates. In fact, for α parameters, finite sample performance of the MPL estimates even seemed slightly superior to the ML estimates. The simple latent class model (ML-S) did not provide information on the scale parameter α_0 or intra-cluster correlation ρ. It did estimate the population-average class prevalences π^* consistently.

[Figure 2 about here.]

Table 2 displays standard errors and coverage probabilities of model-based 95% confidence intervals for the three methods. The simulated standard errors are the sample standard deviations of estimates across runs, and thus reflect the underlying uncertainty. The estimated standard errors are the average of model-based standard errors across simulations, thus indicate the uncertainty estimated by the model. The two sets of standard errors were generally close to each other for both methods. Coverage probabilities primarily were close to the 95% nominal value. Standard error agreement and coverage probabilities were worse for the α parameters than for the β parameters. Finally, Table 2 confirmed high efficiency of the MPL estimators.

To summarize, our simulation study suggests that both ML and MPL well accomplish estimation and inference for multilevel latent class models in finite samples.

7. APPLICATION: ANALYSIS OF OBSESSIVE COMPULSIVE DISORDER DATA

We apply the multilevel latent class model to the OCD data described in the Introduction. Our colleagues identified 8 disorders that often co-occur with OCD: generalized anxiety

disorder (GAD), separation anxiety disorder (SAD), panic disorder (PD), tics disorder, major depressive disorder (MDD), mania disorder, grooming disorders (GrD; trichotillomania, pathological skin picking), and body dysmorphic disorder (BDD). The analytic aim is to identify subtypes of OCD based on comorbidity with the 8 disorders. Data for the 706 OCD cases from 238 families were used for the analysis. The family sizes range from 1 to 10, and most families contain two to five members.

[Table 3 about here.]

We began by selecting among models with two, three and four classes. Using marginalization techniques presented in Section 3.4, the two-class model was modestly preferred over a three class model. However, each random subsample contains only 238 subjects, and it is known that BIC may underestimates the number of classes in such small samples (Yang, 2006). Given that the choice was equivocal, we present the more illustrative three class model. For the three class model, both ML and MPL methods converged successfully, and they gave similar results, hence we only report the model fitted by ML (Table 3). Subjects in the first class were characterized by low prevalence of each comorbid disorder except depression, which was estimated to occur in rougly a quarter of class members. In the second class there were moderate prevalences of GAD, SAD, tics, MDD and GrD, in conjunction with low prevalences of panic disorder and mania. Subjects in the third class were at moderate to high risk for nearly all disorders. The population average prevalence of the three classes were estimated as 38%, 32% and 30%, respectively.

The intra-cluster correlation, ρ, was estimated as 0.44 (95% CI: 0.30, 0.59). The odds ratios of same-class membership between same-cluster members were estimated as 7.0, 7.5, 7.7, respectively for classes 1, 2, and 3, and the estimated odds ratios of different-class membership for same-cluster members were approximately 0.32. This indicates a moderate level of heritability for OCD subtypes, such that members of the same family are considerably more likely to have similar types of OCD comorbidity than subjects from different families.

We compared results from our Dirichlet model (ML) with those from the ML-V model. The two models gave similar latent class structure for the measurement model (β estimates), but the estimated standard errors differed by methods. On average, ML-V based standard errors were 5-10% larger than those from ML. As to the mixing parts of models (α parameters), the two models had different implications. Figure 3 showed the density estimates of family specific mixing probabilities u_{i1}, u_{i2}, and u_{i3}, from two models. Both models implied the density shape that have peaks near the boundary (0 or 1) and is flat in the middle. However, one curious feature of the ML-V is that the mixing probability for class 2, u_{i2}, does not take any values above 0.702. This phenomenon appears to be due to the inflexibility of unidimensional factor analysis type structure for modelling dependence in three classes. More precisely, the two-dimensional mixing probabilities (u_{i1}, u_{i2}), as functions of the unidimensional random effect v_i, are restricted to take values only in a one-dimensional subspace of their domain (the space $[0,1] \times [0,1]$ subject to the constraint $u_{i1} + u_{i2} \leq 1$). In contrast, our Dirichlet model allows u_{im}'s to take values from 0 to 1. The mixing probabilities (u_{i1}, u_{i2}) are allowed to take any value in the domain.

[Figure 3 about here.]

8. DISCUSSION

Latent class models have proven useful for modeling multiple categorical outcomes in social sciences and biomedical studies. In such studies multilevel or hierarchical designs are increasingly common. This paper considered an alternate model to the one proposed by Vermunt (2003, 2008), employing a Dirichlet mixing distribution. Two methods for model fitting and inference, ML and MPL, were developed and compared. We also investigated the consequences of ignoring clustering with a simple latent class model. Our models' random effects structure has more straightforward interpretation than those of competing methods, thus should usefully augment tools available for latent class analysis of clustered data.

Our model has limitations due to the Dirichlet distributional assumption. Subjects within clusters were treated as exchangeable. The Dirichlet assumption restricts the density of u_{km}'s

to be bell-shaped, flat or "U"-shaped, excluding others such as a bimodal shape with two modes in the middle. Moreover our model assumes ICCs to be the same within each class. Such assumption may sometimes be questionable, for example, in genetic studies where different types of relatives may have different heritability. If we are concerned about the validity of those assumptions, the semiparametric model with MPL estimation (Remark 1c in Section 5) would serve as a robust alternative. In contrast, the ML-V allows the ICCs to differ, but its normality and unidimensionality assumptions impose restrictions that may sometimes be undesirable (Section 7). As mentioned in Vermunt (2003), one could generalize the ML-V model to allow multivariate normal random effects, and this has been implemented in the MPlus software (Muthén and Muthén, 2007). However, fitting such models might be computationally intensive because it involves high dimensional integration ($M-1$ dimensional integration for a M-class model). For the OCD data, we fitted a generalized ML-V model with three classes and two dimensional random effects, and the findings regarding association structure were very similar to those from the ML-V model with one dimensional random effects.

There remain issues that would benefit from further research. First, model selection is complicated by the multilevel structure. Though marginalization provides a workable solution, simpler criteria would be useful. Second, diagnostics and model checking techniques are needed. Third, the MLCA model makes the conditional independence assumption. The clustering is assumed to affect only the mixing model, not the measurement model. Models allowing dependence in family members' tendency to report specific items, and not only their class memberships, are needed to address this. Finally, it would be of interest to develop multilevel latent class regression models that incorporate covariates in subpopulation mixing distribution.

APPENDIX A. MORE INSIGHTS INTO MULTILEVEL LATENT CLASS MODEL WITH DIRICHLET MIXING DISTRIBUTION

Proposition 2. The following results hold under MLCA model (3):

1. $E(u_{im}) = \frac{\alpha_m}{\alpha_0}$, $\mathrm{var}(\mathrm{u_{im}}) = \frac{\alpha_{\mathrm{m}}(\alpha_0 - \alpha_{\mathrm{m}})}{\alpha_0^2(\alpha_0+1)}$, $\mathrm{cov}(\mathrm{u_{im}}, \mathrm{u_{iq}}) = -\frac{\alpha_{\mathrm{m}}\alpha_{\mathrm{q}}}{\alpha_0^2(\alpha_0+1)}$;

2. $\Pr(\eta_{ij} = m \,|\, \underset{\sim}{u}_i) = u_{im}$, $\mathrm{var}\{\mathrm{I}(\eta_{\mathrm{ij}} = \mathrm{m}) \,|\, \underset{\sim}{u}_{\mathrm{i}}\} = \mathrm{u_{im}}(1 - \mathrm{u_{im}})$;
   $\Pr(\eta_{ij} = m) = \frac{\alpha_m}{\alpha_0}$, $\mathrm{var}\{\mathrm{I}(\eta_{\mathrm{ij}} = \mathrm{m})\} = \frac{\alpha_{\mathrm{m}}(\alpha_0 - \alpha_{\mathrm{m}})}{\alpha_0^2}$;

3. $\mathrm{cor}\{\mathrm{I}(\eta_{\mathrm{ij}} = \mathrm{m}), \mathrm{I}(\eta_{\mathrm{ik}} = \mathrm{m})\} = \frac{1}{\alpha_0+1}$,
   $\mathrm{cor}\{\mathrm{I}(\eta_{\mathrm{ij}} = \mathrm{m}), \mathrm{I}(\eta_{\mathrm{ik}} = \mathrm{q})\} = -\frac{1}{\alpha_0+1} \cdot \sqrt{\frac{\alpha_{\mathrm{m}}\alpha_{\mathrm{q}}}{(\alpha_0 - \alpha_{\mathrm{m}})(\alpha_0 - \alpha_{\mathrm{q}})}}$;

4. $\mathrm{OR}\{\mathrm{I}(\eta_{\mathrm{ij}} = \mathrm{m}), \mathrm{I}(\eta_{\mathrm{ik}} = \mathrm{m})\} = \frac{(\alpha_{\mathrm{m}}+1)(\alpha_0 - \alpha_{\mathrm{m}} + 1)}{\alpha_{\mathrm{m}}(\alpha_0 - \alpha_{\mathrm{m}})}$,
   $\mathrm{OR}\{\mathrm{I}(\eta_{\mathrm{ij}} = \mathrm{m}), \mathrm{I}(\eta_{\mathrm{ik}} = \mathrm{q})\} = 1 - \frac{1+\alpha_0}{(\alpha_0 - \alpha_{\mathrm{m}} + 1)(\alpha_0 - \alpha_{\mathrm{q}} + 1)}$.

5. $\Pr(\eta_{ij} = m, \eta_{ik} = m) = \frac{\alpha_m(\alpha_m+1)}{\alpha_0(\alpha_0+1)}$, $\Pr(\eta_{ij} = m, \eta_{ik} = q) = \frac{\alpha_m\alpha_q}{\alpha_0(\alpha_0+1)}$

Result 2 implies that in the population, the average (marginal) prevalence of classes is $(\alpha_1/\alpha_0, \alpha_2/\alpha_0, ..., \alpha_M/\alpha_0)$. In contrast, $(u_{i1}, u_{i2}, ..., u_{iM})$ is cluster specific (conditional) class prevalence. Result 5 gives the joint distribution of latent class membership for any two subjects in the same cluster, which is useful for the pairwise likelihood approach.

APPENDIX B. SOME DETAILS OF THE EM ALGORITHM

Lemma 1: Let $\underset{\sim}{z} = (z_1, ..., z_M) \sim Dirichlet(\gamma_1, ..., \gamma_M)$ and define $\gamma_0 = \sum_m \gamma_m$. Then (i) $E(z_m) = \frac{\gamma_m}{\gamma_0}$; (ii) $E[\log(z_m)] = D\Gamma(\gamma_m) - D\Gamma(\gamma_0)$, where $D\Gamma(x) := \frac{d}{dx}\log\{\Gamma(x)\}$.

Proposition 3: The following results hold for the EM algorithm defined in Section 3.1:

1. $[\underset{\sim}{u}_i \,|\, \underset{\sim}{\eta}_i \,;\, \beta^{(h)}, \alpha^{(h)}] \sim Dirichlet(\alpha_1^{(h)} + q_1^{(i)}, ..., \alpha_M^{(h)} + q_M^{(i)})$;

2. $\underset{\sim}{u}_i \perp \underset{\sim}{Y}_i \,|\, \underset{\sim}{\eta}_i$;

3. $E[\log(u_{im}) | \underset{\sim}{Y}_i, \underset{\sim}{\eta}_i; \beta^{(h)}, \alpha^{(h)}] = D\Gamma(\alpha_m^{(h)} + q_m^{(i)}) - D\Gamma(\sum_m \alpha_m^{(h)} + n_i)$;

4. $E[u_{im} | \underset{\sim}{Y}_i, \underset{\sim}{\eta}_i; \beta^{(h)}, \alpha^{(h)}] = \frac{\alpha_m^{(h)} + q_m^{(i)}}{\sum_m \alpha_m^{(h)} + n_i}$

Lemma 1 can be proved by direct calculation using properties of the Dirichlet distribution. Result 1 in Proposition 3 can be derived by Bayes' rule and the conjugacy of the Dirichlet distribution to the multinomial distribution. Result 2 follows from the formulation of multilevel latent class model. Results 3 and 4 follows immediately from Lemma 1.

APPENDIX C. DETAILS ON ESTIMATING THE OBSERVED FISHER INFORMATION

As stated in Section 3.3, we use formulas in Oakes (1999) to obtain the observed information matrix. Specifically, we plug in the parameter estimates in the final EM iteration $\hat{\theta} = (\hat{\beta}, \hat{\alpha})$, i.e,

$$\frac{\partial^2 \log\ L(\theta)}{\partial\theta\,\partial\theta'}|_{\theta=\hat{\theta}} = \left[\ \frac{\partial^2 Q(\theta;\psi)}{\partial\theta\,\partial\theta'} + \frac{\partial^2 Q(\theta;\psi)}{\partial\theta\,\partial\psi'}\ \right]|_{\theta=\hat{\theta},\psi=\hat{\theta}}\ . \tag{A.1}$$

The first term on the right hand side of the equation above is relatively easy to obtain. After the EM algorithm converges, we can carry out one more E-step and obtain the second derivatives of the Q function evaluated at the final iteration. It is generally hard to obtain an analytic form for the second term. Instead, we calculate it by numerical derivatives, i.e, using the formula,

$$\frac{\partial^2 Q(\theta;\psi)}{\partial\theta\,\partial\psi'}|_{\theta=\hat{\theta},\psi=\hat{\theta}} \approx \frac{\left[\ \frac{\partial Q(\theta;\psi)}{\partial\theta} - \frac{\partial Q(\theta;\psi+\Delta\psi)}{\partial\theta}\ \right]}{\Delta\psi}|_{\theta=\hat{\theta},\psi=\hat{\theta}}\ . \tag{A.2}$$

In practice, we can choose $\Delta\psi$ to be a small number, such as 10^{-5}. One can also use iterative algorithm, i.e, choose a $\Delta\psi$ at first, then decrease until the estimated derivatives stabilize.

To summarize, the algorithm to estimate the observed Fisher information is as follows.

1. Use the EM algorithm until it converges. Denote the parameter estimates in the last iteration $\hat{\theta}^{final}$;

2. Perform one more EM step and obtain $\frac{\partial Q(\theta;\hat{\theta}^{final})}{\partial\theta}|_{\theta=\hat{\theta}^{final}}$ and $\frac{\partial^2 Q(\theta;\hat{\theta}^{final})}{\partial\theta\,\partial\theta'}|_{\theta=\hat{\theta}^{final}}$ using formulas in Section 3.1. The latter is the first term in equation (A.1);

3. Choose a small number $\Delta\psi$, and carry out EM-steps to obtain the first order derivatives $\frac{\partial Q(\theta;\hat{\theta}^{final}+\Delta\psi)}{\partial\theta}|_{\theta=\hat{\theta}^{final}}$. Use (A.2) to estimate the second term in equation (A.1);

4. Obtain the observed Fisher information by equation (A.1).

APPENDIX D. PROOF OF PROPOSITION 1

Sketch of Proof: (1). Since the simple latent class model is the marginalization of the semiparametric model, $f(\underset{\sim}{Y}_{ij}) = \sum_m \pi_m \Pr(Y_{ijk} = y_k|\eta_{ij} = m)$ is the true likelihood (with

true parameter values β and π^*) contributed by the j^{th} subject of the i^{th} cluster. Under typical regularity conditions, its derivatives with respective to β and π are unbiased. By additivity of expectations, this would lead to the unbiasedness of the score functions of the likelihood from the simple latent class model.

(2) and (3). Since the score functions of the simple likelihood are unbiased, one can immediately obtain the consistency and asymptotic normality based on estimating functions theory (e.g, in van der Vaart, 2000), when the number of clusters goes to infinity and the cluster size is fixed.

REFERENCES

Akaike, H. "A new look at the statistical identification model." *IEEE Transactions on Automatic Control*, 19(3):716–723 (1974).

Bandeen-Roche, K., Miglioretti, D., Zeger, S., and Rathouz, P. "Latent Variable Regression for Multiple Discrete Outcomes." *Journal of the American Statistical Association*, 92(440) (1997).

Clogg, C. "Latent class models." *Handbook of statistical modeling for the social and behavioral sciences*, 311–359 (1995).

Cox, D. and Reid, N. "A note on pseudolikelihood constructed from marginal densities." *Biometrika*, 91(3):729 (2004).

Dempster, A., Laird, N., and Rubin, D. "Maximum likelihood from incomplete observations." *Journal of the Royal Statistical Society, Series B*, 39:1–38 (1977).

Haughton, D. "On the choice of a model to fit data from an exponential family." *Annals of Statistics*, 16(1):342–355 (1988).

Jenike, M., Baer, L., and Minichiello, W. *Obsessive Compulsive Disorders: Theory and Management*. Chicago: Year Book Medical Publishers (1990).

Kuk, A. and Nott, D. "A pairwise likelihood approach to analyzing correlated binary data." *Statistics and Probability Letters*, 47(4):329–335 (2000).

Liang, K. and Zeger, S. "Longitudinal data analysis using generalized linear models." *Biometrika*, 73(1):13 (1986).

Lindsay, B. "Composite likelihood methods." *Contemporary Mathematics*, 80:221–239 (1988).

Little, R. and Rubin, D. "Statistical analysis with missing data . Hoboken." (2002).

Louis, T. "Finding the Observed Information Matrix when Using the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233 (1982).

Muthén, L. and Muthén, B. "Mplus Users Guide (Fifth Edition)." *Los Angeles, Muthén & Muthén* (2007).

Nestadt, G., Addington, A., Samuels, J., Liang, K., Bienvenu, O., Riddle, M., Grados, M., Hoehn-Saric, R., and Cullen, B. "The identification of OCD-related subgroups based on comorbidity." *Biological Psychiatry*, 53(10):914–920 (2003).

Nott, D. and Ryden, T. "Pairwise likelihood methods for inference in image models." *Biometrika*, 86(3):661 (1999).

Oakes, D. "Direct Calculation of the Information Matrix via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(2):479–482 (1999).

Potthoff, R., Manton, K., and Woodbury, M. "Dirichlet generalizations of latent-class models." *Journal of classification*, 17(2):315–353 (2000).

Reboussin, B., Liang, K., and Reboussin, D. "Estimating Equations for a Latent Transition Model with Multiple Discrete Indicators." *Biometrics*, 55(3):839–845 (1999).

Renard, D., Molenberghs, G., and Geys, H. "A pairwise likelihood approach to estimation in multilevel probit models." *Computational Statistics and Data Analysis*, 44(4):649–667 (2004).

Royall, R. "Model Robust Confidence Intervals Using Maximum Likelihood Estimators." *International Statistical Review/Revue Internationale de Statistique*, 54(2):221–226 (1986).

Schwarz, G. "Estimating the Dimension of a Model." *The Annals of Statistics*, 6(2):461–464 (1978).

Sclove, S. "Application of model-selection criteria to some problems in multivariate analysis." *Psychometrika*, 52(3):333–343 (1987).

van der Vaart, A. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press (2000).

Varin, C., Høst, G., and Skare, Ø. "Pairwise likelihood inference in spatial generalized linear mixed models." *Computational Statistics and Data Analysis*, 49(4):1173–1191 (2005).

Varki, S. and Chintagunta, P. "The Augmented Latent Class Model: Incorporating Additional Heterogeneity in the Latent Class Model for Panel Data." *Journal of Marketing Research*, 41(2):226–233 (2004).

Vermunt, J. "Multilevel Latent Class Models." *Sociological Methodology*, 33(1):213–239 (2003).

—. "Latent class and finite mixture models for multilevel data sets." *Statistical Methods in Medical Research*, 17(1):33 (2008).

Yang, C. "Evaluating latent class analysis models in qualitative phenotype identification." *Computational Statistics and Data Analysis*, 50(4):1090–1104 (2006).

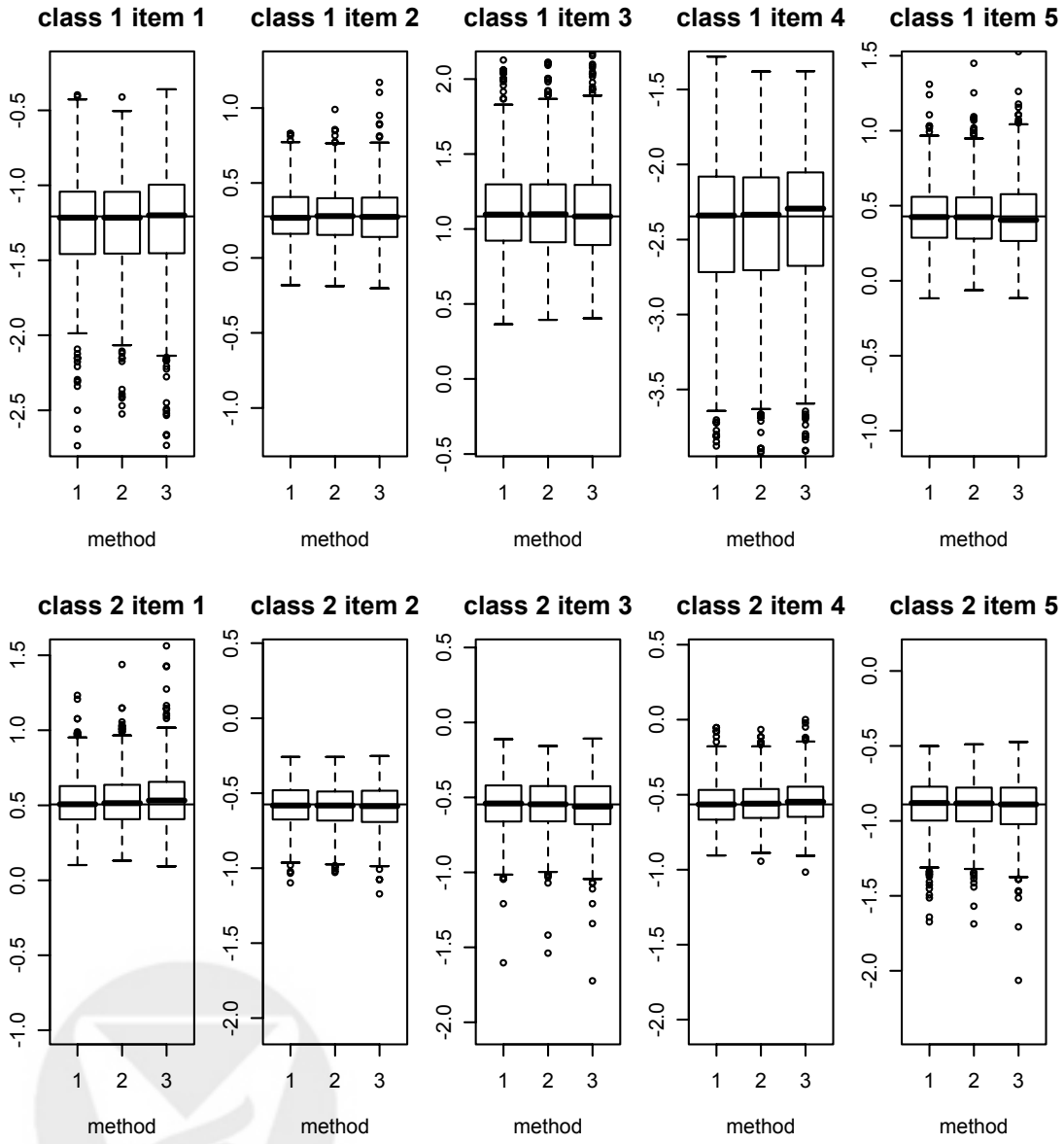

Figure 1: Boxplot of measurement model estimates using 200 clusters: β parameters. The first row displays estimates for class 1 parameters, and the second row shows class 2 parameters. The five columns correspond to 5 items. Methods 1, 2 and 3 correspond to "ML", "MPL", and "ML-S", respectively. The solid lines are true parameter values.

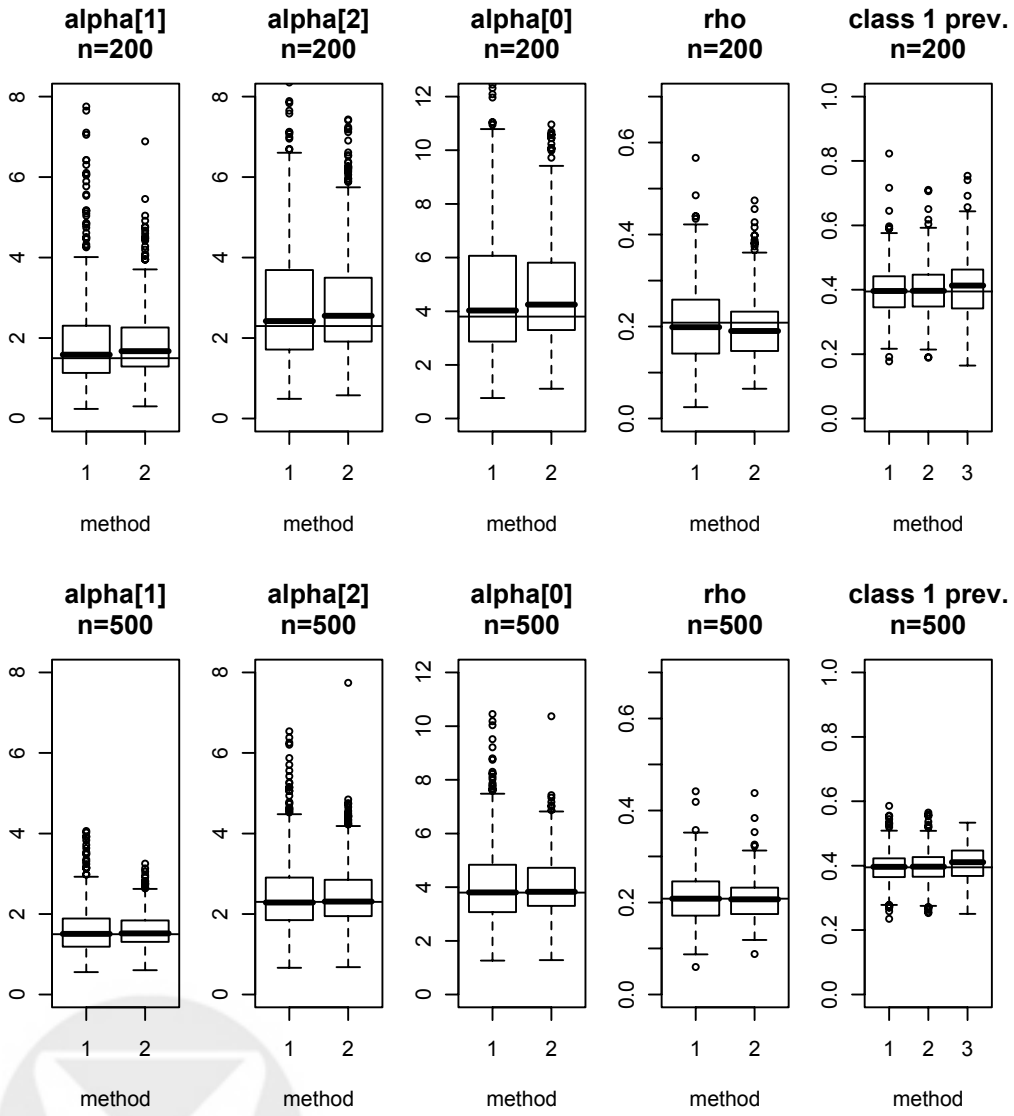

Figure 2: Mixing distribution model estimates: α parameters. The first row displays results using 200 clusters, and the second row shows results using 500 clusters. The four columns correspond to α_1,α_2, α_0 (scale parameter), ρ (intra-cluster correlation) and α_1/α_0 (average class 1 prevalence), respectively. Methods 1 and 2 correspond to "ML", and "MPL" for the Dirichlet model, while method 3 corresponds to the "ML-S" for the simple latent class model.

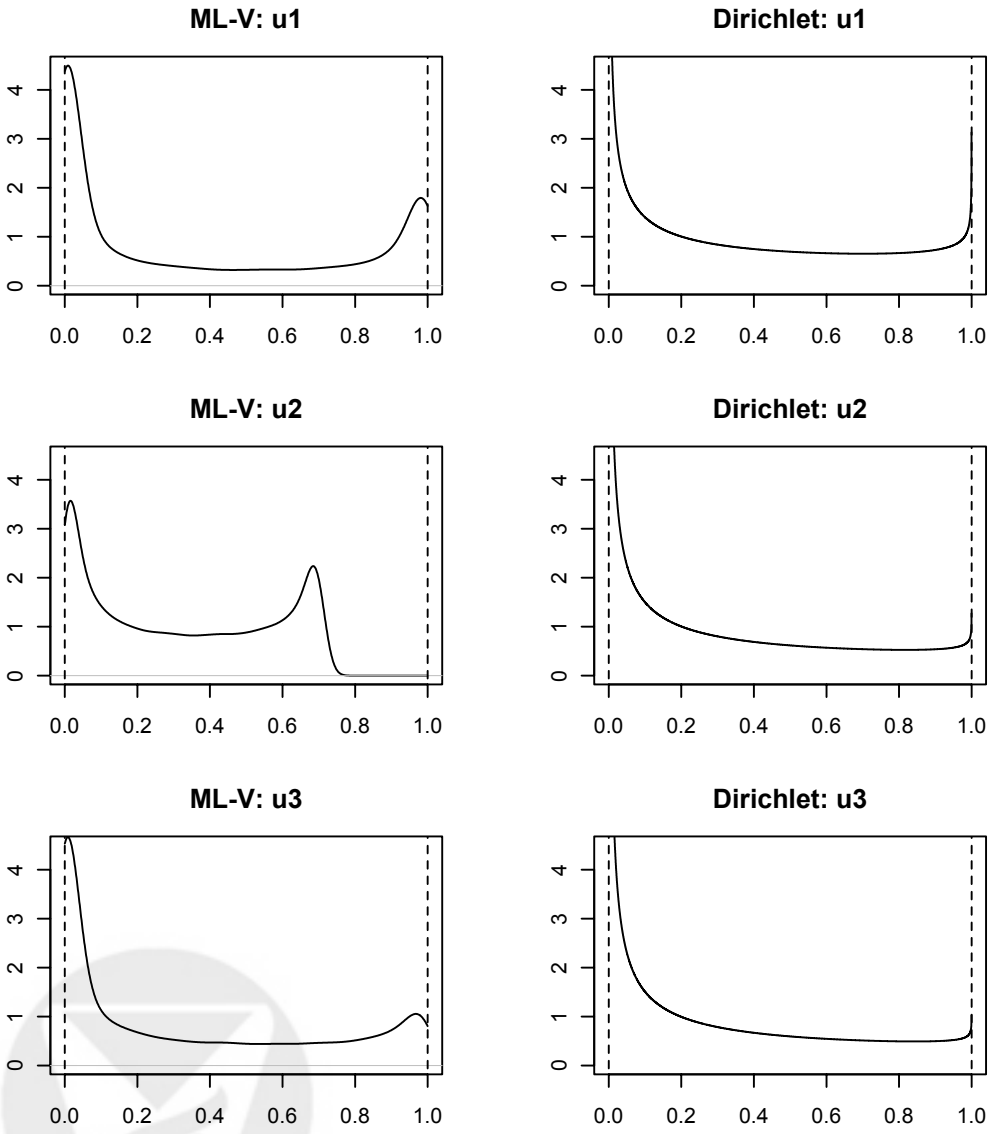

Figure 3: Density estimates for class mixing random effects $u_i = (u_{i1}, u_{i2}, u_{i3})$ from both the ML-V model and our Dirichlet MLCA model for OCD data. The three rows correspond to estimated density for random mixing probabilities of class 1, 2, and 3, respectively.

Table 1: Ratio of computational complexity: comparing ML to MPL.

cluster size (J)	number of classes (M)				
	2	3	4	5	6
2	1.00	1.00	1.00	1.00	1.00
3	0.70	1.00	1.30	1.70	2.00
4	0.70	1.50	2.70	4.20	6.00
5	0.80	2.70	6.40	12.50	21.60
6	1.10	5.40	17.10	41.70	86.40
7	1.50	11.60	48.80	148.80	370.30
8	2.30	26.00	146.30	558.00	1666.30
9	3.60	60.80	455.10	2170.10	7776.00
10	5.70	145.80	1456.40	8680.60	37324.80

Table 2: Standard errors and coverage probabilities for methods "ML", "MPL" and "ML-S" using 200 clusters. "ML" and "MPL" are ML and MPL estimates from the Dirichlet model, while "ML-S" stands for ML estimates from the simple latent class model. "simu. SE" means empirical standard errors across simulations, "est. SE" means model-based standard errors, and "cov. prob" means coverage probabilities (%) for model based 95% nominal confidence intervals.

Method	Class	SE & cov. prob.	α	$\frac{\alpha}{\alpha_0}$	β item 1	β item 2	β item 3	β item 4	β item 5
ML	Class 1	simu. SE	1.90	0.08	0.37	0.19	0.34	0.77	0.22
		est. SE	1.96	0.08	0.36	0.19	0.32	0.69	0.22
		cov. prob.	92.0	92.2	95.2	96.6	95.6	93.4	95.0
	Class 2	simu. SE	2.85	0.08	0.19	0.14	0.26	0.15	0.18
		est. SE	3.17	0.08	0.19	0.14	0.23	0.16	0.18
		cov. prob.	90.8	92.2	95.6	95.6	93.6	96.0	94.0
MPL	Class 1	simu. SE	0.87	0.08	0.37	0.19	0.33	0.86	0.23
		est. SE	1.08	0.09	0.40	0.20	0.36	0.83	0.24
		cov. prob.	95.4	95.8	95.4	95.8	95.6	94.0	96.2
	Class 2	simu. SE	1.40	0.08	0.18	0.14	0.19	0.15	0.18
		est. SE	1.69	0.09	0.20	0.15	0.20	0.17	0.19
		cov. prob.	95.6	95.8	96.4	96.6	95.4	97.0	95.0
ML-S	Class 1	simu. SE	-	0.09	0.52	0.20	0.37	1.03	0.24
		est. SE	-	0.10	0.43	0.20	0.38	0.90	0.26
		cov. prob.	-	93.8	95.6	95.0	96.2	93.0	95.8
	Class 2	simu. SE	-	0.09	0.20	0.15	0.20	0.16	0.19
		est. SE	-	0.10	0.22	0.16	0.21	0.17	0.20
		cov. prob.	-	93.8	95.0	96.0	94.8	96.4	95.4

Table 3: Model fitting for OCD data using ML. (MPL results are similar and thus are omitted.) "est." means the point estimates for conditional probabilities of reporting a certain disorder given the subject belongs to a certain class.

	Class 1			Class 2			Class 3		
	est.	95%	CI	est.	95%	CI	est.	95%	CI
GAD	0.12	(0.05,	0.25)	0.56	(0.42,	0.69)	0.67	(0.57,	0.76)
SAD	0.11	(0.05,	0.20)	0.26	(0.17,	0.37)	0.41	(0.33,	0.50)
Panic	0.10	(0.06,	0.17)	0.03	(0.00,	0.21)	0.48	(0.38,	0.59)
Tics	0.13	(0.07,	0.23)	0.41	(0.30,	0.52)	0.27	(0.20,	0.35)
MDD	0.27	(0.20,	0.36)	0.23	(0.15,	0.35)	0.68	(0.56,	0.77)
Man	0.03	(0.01,	0.09)	0.00	(0.00,	0.00)	0.19	(0.13,	0.27)
GrD	0.16	(0.08,	0.28)	0.48	(0.37,	0.59)	0.59	(0.50,	0.68)
BDD	0.06	(0.02,	0.13)	0.16	(0.09,	0.26)	0.53	(0.43,	0.63)
α	0.49	(0.25,	0.96)	0.40	(0.19,	0.86)	0.38	(0.20,	0.71)
average prev.	0.38	(0.27,	0.52)	0.32	(0.20,	0.46)	0.30	(0.22,	0.39)