Biometrics 000, 1–17 DOI: 000
000 0000

Web-based Supplementary Materials for “Multilevel Latent Class Models with Dirichlet Mixing Distribution”

Chong-Zhi Di* and Karen Bandeen-Roche**

**email:* cdi@fhcrc.org

***email:* kbandeen@jhsph.edu

Summary: This report contains supplementary materials for the paper “multilevel latent class models with Dirichlet mixing distribution.” Section A provides insights of various multilevel latent class (MLC) models. Section B contains some technical results related to the EM algorithm. Section C includes proofs for asymptotic results. Sections D and E show additional results from simulation studies and the OCD application.

In the the paper and supplementary materials, we use the following acronyms.

LCA:	latent class analysis
LCR:	latent class regression
MLC model:	multilevel latent class model
MLC-V1 model:	MLC model with unidimensional normal random effects (Vermunt, 2003)
MLC-V2 model:	MLC model with multidimensional normal random effects (Vermunt, 2003)
MLC-VN model:	MLC model with “nonparametric” random effects (Vermunt, 2003)
MLC-D model:	MLC model with Dirichlet random effects
LC-S model:	standard latent class model that ignores clustering
ML-D:	maximum likelihood using the MLC-D model
MPL-D:	maximum pairwise likelihood using the MLC-D model
ML-S:	maximum likelihood using the simple latent class model
ML-V1:	maximum likelihood using the MLC-V1 model

A. Insights of various MLC models

As described in the paper, various multilevel latent class (MLC) models make the same assumptions on the measurement part (β parameters), and they differ primarily in the mixing part. The cluster specific class mixing probabilities $\underset{\sim}{u}_i = (u_{i1}, ..., u_{iM})$ are treated as random effects, whose distributions are specified differently by various models. Note that the domain of $\underset{\sim}{u}_i$, $\Omega_u = \{(u_1, \cdots, u_M) : u_1 + \cdots + u_M = 1 \text{ and } 0 \leqslant u_m \leqslant 1 \text{ for } m = 1, \cdots, M\}$, is an $(M-1)$ dimensional subspace of $[0,1]^M$.

A.1 *Marginal class prevalences (MCPs) and intra-cluster correlations (ICCs)*

We introduce two quantities for MLC models that are often of scientific interest and are conveniently interpretable, marginal class prevalences and intra-cluster correlations. We define marginal prevalance of class m as $\pi_m = \mathrm{E}(\eta_{ij} = m)$, which reflects the average prevalence of class m over the whole population. Note that

$$\pi_m = \mathrm{E}(\eta_{ij} = m) = \mathrm{E}\{\, \mathrm{E}(\eta_{ij} = m | u_i)\,\} = \mathrm{E}(u_{im}). \tag{1}$$

To describe within cluster dependence, we define intra-cluster correlations as follows. Let $\rho_{mm} = \mathrm{Cor}\{\, I(\eta_{ij} = m), I(\eta_{ik} = m)\,\}$ (“same-class ICC”) denote the correlation that two subjects from the same cluster belong to the same class m, and $\rho_{mq} = \mathrm{Cor}\{\, I(\eta_{ij} = m), I(\eta_{ik} = q)\,\}$ (for $m \neq q$; “different-class ICC”) denote the correlation that two subjects from the same cluster belong to different classes m and q, respectively. The former is often of more scientific interest, and can be interpreted as “heritability” in family studies. One can show that the ICCs are fully determined by the first and second order moments of the random effects u_i's. More precisely, the following

$$\begin{aligned}\rho_{mq} &= \operatorname{cor}\{\, I(\eta_{ij}=m), I(\eta_{ik}=q)\,\} \\ &= \frac{\operatorname{cov}\{\, I(\eta_{ij}=m), I(\eta_{ik}=q)\,\}}{\sqrt{\operatorname{var}\{\, I(\eta_{ij}=m)\,\}\operatorname{var}\{\, I(\eta_{ik}=q)\,\}}} \\ &= \frac{\operatorname{cov}\{\, u_{im}, u_{iq}\,\}}{\sqrt{\operatorname{var}\{\, I(\eta_{ij}=m)\,\}\operatorname{var}\{\, I(\eta_{ik}=q)\,\}}} \\ &= \frac{\mathrm{E}(u_{im}u_{iq}) - \mathrm{E}(u_{im})\mathrm{E}(u_{iq})}{\sqrt{\mathrm{E}(u_{im})\mathrm{E}(u_{iq})\{1-\mathrm{E}(u_{im})\}\{1-\mathrm{E}(u_{iq})\}}} \qquad (2)\end{aligned}$$

holds for $m = q$ and $m \neq q$.

One might also use odds ratios to measure within cluster dependency, since $I(\eta_{ij} = m)$ and $I(\eta_{ik} = q)$ are both binary random variables. The odds ratios are also fully determined by the first and second order moments of the random effects u_i's. We will not give the full formulas here.

A.2 *Vermunt's models*

Vermunt (2003, 2008) considered a few multilevel latent class models. These models assume the existence of higher level random effects v_i's, which could be unidimensional continuous, multi-dimensional continuous or discrete random variables, and build relationships between $\underset{\sim}{u}_i$ and v_i's.

A.2.1 *MLC-V1.*

The "MLC-V1" model assumes that the higher level random effects v_i's come from a one-dimensional normal distribution and that logistic transformed $\underset{\sim}{u}_i$'s depend on v_i's through a factor-analysis type structure, i.e.

$$\begin{cases} \log \dfrac{u_{im}}{u_{i1}} = \gamma_m + \lambda_m v_i \ , \ m = 2, \cdots, M, \\ v_i \sim N(0, \sigma^2), \end{cases} \qquad (3)$$

where γ_m, λ_m and σ^2 are unknown parameters. For identifiability, let $\lambda_2 = 1$. We also denote $\gamma_1 = 0, \lambda_1 = 0$ for $m = 1$ for convenience of notations in the following. The assumption that the $(M-1)$ generalized logits of $\underset{\sim}{u}_i$'s depend only on one-dimensional normal random effects makes computation convenient, but could be restrictive in practice.

The MLC-V1 model does not yield closed-form formulas for the MCPs and ICCs. Instead, one need to obtain these quantities by numerical integrations or Monte Carlo simulations. For example, the first and second moments of $\underset{\sim}{u}_i$ are given as

$$\mathrm{E}(u_{im}) = \int_R \frac{exp(\gamma_m + \lambda_m v_i)}{\sum_m exp(\gamma_m + \lambda_m v_i)} \phi(v_i)\, dv_i,$$

$$\mathrm{E}(u_{im}u_{iq}) = \int_R \frac{exp(\gamma_m + \lambda_m v_i + \gamma_q + \lambda_q v_i)}{\sum_m exp(\gamma_m + \lambda_m v_i)} \phi(v_i)\, dv_i,$$

and one can further calculate the MCPs and ICCs using formulas (1) and (2).

In this model, the variance parameter σ^2 controls the level of heterogeneity among clusters. When $\sigma^2 = 0$, $\underset{\sim}{u}_i = (\pi_1, \pi_2, \cdots, \pi_M)$ is constant for all subjects and the model reduces to a standard LC model without clustering. When σ^2 is large, $\underset{\sim}{u}_i$'s are quite different among clusters, which means large heterogeneity and high ICCs. Figure 1 illustrates implications of varying σ^2 on the distribution of $\underset{\sim}{u}_i$'s and on ICCs.

[Figure 1 about here.]

In terms of estimation, Vermunt (2003) proposed to use EM algorithm with numerical integration over random effects. The unidimensionality of the random effects makes computation feasible. However, it also imposes strong assumptions on the distribution of $\underset{\sim}{u}_i$. With the constraint $u_{i1} + \cdots + u_{iM} = 1$, the domain of $\underset{\sim}{u}_i$ is intrinsically (M-1)-dimensional subspace of $[0, 1]^M$. However, for fixed parameters γ_m and λ_m's, model (3) only allow $\underset{\sim}{u}_i$ to take values from a one-dimensional subspace (see e.g., Figure 1). This could be restrictive.

A.2.2 *MLC-V2.* To allow more flexibility, Vermunt (2003) also introduced the “MLC-V2” model (4) with $(M-1)$-dimensional random effects, i.e.,

$$\begin{cases} \log \dfrac{u_{im}}{u_{iM}} = \gamma_m + v_{im}\ ,\ m = 2, \cdots, M-1, \\ \underset{\sim}{v}_i = (v_{i2}, \cdots, v_{i(M-1)}) \sim MVN(\mathbf{0}, \mathbf{\Sigma}), \end{cases} \tag{4}$$

where γ_m's and $\mathbf{\Sigma}$ are unknown parameters.

Under the MLC-V2 model, the MCPs and ICCs do not have closed form formulas, and numerical integration over $(M-1)$ dimensional random effects are needed to calculate them. For example, the first and second moments of $\underset{\sim}{u}_i$ are given as

$$\mathrm{E}(u_{im}) = \int_{R^{M-1}} \frac{exp(\gamma_m + v_{im})}{\sum_m exp(\gamma_m + v_{im})}\, \phi_{M-1}(v_i; 0, \Sigma)\, d\underset{\sim}{v}_i,$$

$$\mathrm{E}(u_{im}u_{iq}) = \int_{R^{M-1}} \frac{exp(\gamma_m + v_{im} + \gamma_q + v_{iq})}{\sum_m exp(\gamma_m + v_{im})}\, \phi_{M-1}(v_i; 0, \Sigma)\, d\underset{\sim}{v}_i,$$

and one can further calculate the MCPs and ICCs using formulas (1) and (2).

For fixed parameters, this model allows more flexible distribution for $\underset{\sim}{u}_i$. In fact $\underset{\sim}{u}_i$ is allowed to take any value in Ω_u. However, the computational burden of $(M-1)$ dimensional numerical integration in model estimation circumvent this approach from wide use, especially for moderate to large M. Thus, it was briefly introduced but not pursued in details by Vermunt (2003).

Similarly, the covariance matrix Σ determines the level of between cluster heterogeneity and ICCs. When $\Sigma = \mathbf{0}$, the MLC-V2 model reduces to a standard latent class model that ignores clustering.

A.2.3 *MLC-VN.*

$$\begin{cases} [\,\underset{\sim}{u}_i \,|\, v_i = s\,] = (\psi_{s1}, \cdots, \psi_{sM})\ ,\ s = 1, \cdots, S \\ v_i \sim Discrete(\tau_1, \cdots, \tau_S) \end{cases} \tag{5}$$

This model was termed "nonparametric" multilevel latent class model by Vermunt (2003, 2008). Essentially, it assumes the random effects $\underset{\sim}{u}_i$ has follows a discrete distribution with a finite number of support points $\Omega_{VN} = \{(\psi_{s1}, \cdots, \psi_{sM}) : s = 1, 2, \cdots, S\}$. Intuitively, this model assumes that there are S different types of clusters and that each cluster belongs to type s with probability τ_s.

Under the MLC-VN model, the MCPs and ICCs can be obtained by first calculating the first two moments of u_{im}'s according to

$$\mathrm{E}(u_{im}) = \sum_{s=1}^{S} \psi_{sm}\tau_s\ ,\ \ \mathrm{E}(u_{im}u_{iq}) = \sum_{s=1}^{S} \psi_{sm}\psi_{sq}\tau_s,$$

and then applying formulas (1) and (2).

The MLC-VN model was considered a more flexible "non-parametric" or "semi-parametric" approach. This argument holds if the number of support points S is large so that the discrete distribution can approximate the underlying distribution of $\underset{\sim}{u}_i$ well. Since $\underset{\sim}{u}_i$ intrinsically $(M - 1)$ dimensional, S needs to be a large number (say 5^{M-1} with 5 grid points per dimension) for a good approximation. On the other hand, a large or even moderate S would lead to a large number of parameters that might cause problems for model identification and estimation. Thus, in previous literature, S was often chosen to be a small number, say between 2 and 5. Since the domain of $\underset{\sim}{u}_i$ is an (M-1)-dimensional space, a discrete distribution with so few support points may not be flexible enough. In addition, two levels of latent class structure make model interpretation, identifiability, selection and checking really complicated. These issues were not well understood from existing literature.

Finally, if one assumes $S = 1$, the MLC-VN reduced to a standard latent class model that ignores clustering.

A.3 *MLC-D*

In this paper, we considered MLC models with Dirichlet distributed random effects, i.e.,

$$\underset{\sim}{u}_i \sim Dirichlet(\alpha_1, \alpha_2, \cdots, \alpha_M), \tag{6}$$

where α_m's are unknown parameters.

The MLC-D model assumes that the cluster specific random effects $\underset{\sim}{u}_i$'s follow a Dirichlet distribution. This is more natural compared to Vermunt's models because of the following reasons. First, its domain is exactly Ω_u, instead of a restricted subset of Ω_u, for any fixed parameter values. Second, the Dirichlet distributional assumption works directly on the probability scale and does not require any higher level random effects structure. Third, the parameters α_m's and their transformations have meaningful interpretations. For example, as revealed below, the MCPs and ICCs have simple analytic forms for convenient interpretations.

Proposition 3. The following results hold under MLC-D model for $m, q \in \{1, 2, \cdots, M\}$ and $m \neq q$,

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

$$\text{(2)}\ \ \pi_m = \Pr(\eta_{ij} = m) = \frac{\alpha_m}{\alpha_0},\ \mathrm{var\{I(\eta_{ij} = m)\}} = \frac{\alpha_\mathrm{m}(\alpha_0 - \alpha_\mathrm{m})}{\alpha_0^2};$$

$$\text{(3)}\ \ \rho_{mm} = \mathrm{cor\{I(\eta_{ij} = m), I(\eta_{ik} = m)\}} = \frac{1}{\alpha_0+1},$$

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

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

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

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

Under the MLC-D model, marginal prevalences of classes are $\underset{\sim}{\pi} = (\alpha_1/\alpha_0, \alpha_2/\alpha_0, ..., \alpha_M/\alpha_0)$, where $\alpha_0 = \sum_{m=1}^{M} \alpha_m$. The scale parameter, α_0, controls the level of heterogeneity, similar to the role of $1/\sigma^2$ for the MLC-V1 model. When α_0 is large or approaches ∞, the Dirichlet distribution approaches a discrete distribution with probability masses $\underset{\sim}{\pi}$. Under such situations, cluster specific random effects $\underset{\sim}{u}_i$ are nearly constant among all clusters and thus ICCs are close to 0, indicating little heterogeneity. On the other hand, when α_0 is small, there is large variation of $\underset{\sim}{u}_i$'s among all clusters. Thus, the same-class ICCs are high, reflecting large between cluster heterogeneity. This point is demonstrated by Figure 2.

[Figure 2 about here.]

Finally, in the limiting case with $\alpha_m/\alpha_0 \to \pi_m$ and $\alpha_0 \to \infty$ for all $m \in \{1, \cdots, M\}$, the MLC-D model reduces to a standard latent class model without clustering.

A.4 *LC-S*

$$\underset{\sim}{u}_i = (\pi_1, \pi_2, \cdots, \pi_M) \tag{7}$$

The LC-S model is the standard LC model that ignores clustering. This model assumes that the class mixing probabilities are constant, instead of varying among clusters. We note that LC-S corresponds to a degenerate special case for each of the MLC models described above.

Based on this model, the marginal class prevalences (MCPs) are exactly the parameters $(\pi_1, \pi_2, \cdots, \pi_M)$, while the same-class and different-class intra-cluster correlations (ICCs) are all implicitly assumed to 0. As described in the paper, when the true model has clustering, maximum likelihood estimates for β parameters based on this LC-S are still consistent, but their standard errors need to be corrected by the robust estimators.

To summarize, various MLC models make different assumptions on the distribution of random effects $\underset{\sim}{u}_i$'s, and thus have different implications on the domain of $\underset{\sim}{u}_i$, model interpretation, computational burden, etc. Due to reasons described above, the MLC-D model seems more natural and have advantages in interpretability.

A.5 *Grade of membership models vs. latent class models*

The Dirichlet distribution has been used in mixture type models. For example, Potthoff et al. (2000) considered a type of grade of membership (GoM) model, which assumes that each subject can be partial members of all classes and that partial membership weights are Dirichlet distributed random effects. Erosheva (2003) discussed Bayesian inference for GoM models, employing the conjugacy between the Dirichlet and multinomial distributions. Varki and Chintagunta (2004)'s model is a mixture of LCA and Dirichlet based GoM. In the following, we briefly explain differences between these models and our MLC-D model.

Potthoff et al. (2000) and Varki and Chintagunta (2004) are applicable to a sample of *independent* subjects, assume that each subject *belongs partially to all M classes*, and use Dirichlet distributed random effects to reflect *heterogeneity of partial membership weights.* Potthoff et al. (2000) considered a grade of membership (GoM) model, which is applicable to a sample of independent subjects and assumes that each subject can be partial members of all classes and that such mixing probabilities are Dirichlet distributed random effects. Varki and Chintagunta (2004)'s model is a mixture of LCA and Dirichlet based GoM, and is also used to a sample of independent subjects only.

More precisely, we give the mathematical formulas for three models here. First, we compare the GoM model and LCA in the single level case. The mathematical form for Dirichlet based GoM model is

$$\begin{cases} \Pr(\underset{\sim}{Y}_i = \underset{\sim}{y} | \underset{\sim}{u}_i) = \prod_{k=1}^{K} \Pr(Y_{ik} = y_k \mid \underset{\sim}{u}_i) = \prod_{k=1}^{K} \left\{ \sum_m u_{im} p_{km} \right\}^{y_k} \left\{ 1 - \sum_m u_{im} p_{km} \right\}^{1-y_k} \\ \\ \Pr(\underset{\sim}{Y}_i = \underset{\sim}{y}) = \int_{\underset{\sim}{u}_i} \prod_{k=1}^{K} \left\{ \sum_m u_{im} p_{km} \right\}^{y_k} \left\{ 1 - \sum_m u_{im} p_{km} \right\}^{1-y_k} \, dF(\underset{\sim}{u}_i) \\ \underset{\sim}{u}_i = (u_{i1}, \cdots, u_{iM}) \sim \mathit{Dirichlet}(\alpha_1, \cdots, \alpha_M), \end{cases} \tag{8}$$

where each subject i is assumed to be partial members of M classes with random weights $\underset{\sim}{u}_i = (u_{i1}, \cdots, u_{iM})$. The parameter p_{km} is the probability of reporting positive on item k if a subject belongs fully to class m. Conditional on $\underset{\sim}{u}_i$, because subject i are partial members of all classes, its probability of reporting positive on item k is a weighted average of p_{km}'s with weights given by partial memberships, i.e., $\Pr(Y_{ik} = 1 \mid \underset{\sim}{u}_i) = \sum_m u_{im} p_{km}$. In contrast, standard latent class models have formulation

$$\begin{cases} \Pr(\underset{\sim}{Y}_i = \underset{\sim}{y} | \eta_i = m) = \prod_{k=1}^{K} \Pr(Y_{ik} = y_k \mid \eta_i = m) = \prod_{k=1}^{K} p_{km}^{y_k} (1 - p_{km})^{1-y_k} \\ \\ \Pr(\underset{\sim}{Y}_i = \underset{\sim}{y}) = \sum_{m=1}^{M} \pi_m \prod_{k=1}^{K} p_{km}^{y_k} (1 - p_{km})^{1-y_k} \\ \underset{\sim}{\pi} = (\pi_1, \cdots, \pi_M) \text{ fixed,} \end{cases} \tag{9}$$

reflecting the assumption that each subject belongs to one and only one class and that responses of a subject are independent conditional on its class membership. Varki and Chintagunta (2004)'s model is a mixture of the previous two models, by assuming that a proportion of the population follows a latent class model and that the remaining follows a Dirichlet based GoM model, i.e.,

$$\begin{aligned} \Pr(\underset{\sim}{Y}_i = \underset{\sim}{y}) \;&=\; \delta \sum_{m=1}^{M} \pi_m \cdot \prod_{k=1}^{K} p_{km}^{y_k} (1 - p_{km})^{1-y_k} + \\ & (1 - \delta) \int_{\underset{\sim}{u}_i} \prod_{k=1}^{K} \left\{ \sum_m u_{im} p_{km} \right\}^{y_k} \left\{ 1 - \sum_m u_{im} p_{km} \right\}^{1-y_k} dF(\underset{\sim}{u}_i)\,, \end{aligned} \tag{10}$$

where δ and $1 - \delta$ are mixture proportions of the two components, respectively.

In contrast, our multilevel latent class model is applicable to a sample of *clustered* subjects, assume that each subject is a *full member of one and only one class*, and uses Dirichlet distributed random effects to induce *within cluster dependence*. Thus, our model is fundamentally different from Potthoff et al. (2000) and Varki and Chintagunta (2004).

B. On the EM algorithm

In this section, we provide some technical details for the EM algorithm that are briefly mentioned in the paper.

B.1 *Some Details of the EM algorithm*

Proposition 4: Let $\underset{\sim}{z} = (z_1, ..., z_M) \sim \textit{Dirichlet}(\gamma_1, ..., \gamma_M)$ 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 5: The following results hold for the EM algorithm defined in Section 3.1:

(1) $[\,\underset{\sim}{u}_i\,|\,\underset{\sim}{\eta}_i\ ;\ \beta^{(h)}, \alpha^{(h)}\,]\ \sim\ \textit{Dirichlet}(\alpha_1^{(h)} + q_1^{(i)}, ..., \alpha_M^{(h)} + q_M^{(i)})$;
(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) $E[\,u_{im}|\,\underset{\sim}{Y}_i, \underset{\sim}{\eta}_i; \beta^{(h)}, \alpha^{(h)}\,] = \dfrac{\alpha_m^{(h)} + q_m^{(i)}}{\sum_m \alpha_m^{(h)} + n_i}$

Proposition 4 can be proved by direct calculation using properties of the Dirichlet distribution. Result 1 in Proposition 5 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 Proposition 4.

B.2 *Details on estimating the observed Fisher information*

We numerically calculate the observed Fisher information matrix following Oakes (1999):

$$\frac{\partial \log\, L(\theta)}{\partial \theta} = \left[\frac{\partial Q(\psi;\theta)}{\partial \psi}\right]|_{\psi=\theta} \tag{11}$$

$$\frac{\partial^2 \log\, L(\theta)}{\partial \theta\, \partial \theta'} = \left[\frac{\partial^2 Q(\theta;\psi)}{\partial \theta\, \partial \theta'} + \frac{\partial^2 Q(\theta;\psi)}{\partial \theta\, \partial \psi'}\right]|_{\psi=\theta} \tag{12}$$

where $\theta = (\beta, \alpha)$ are the parameters and $\psi = (\beta^{(h)}, \alpha^{(h)})$ are the current estimates. Specifically, we plug in the parameter estimates in the final EM iteration $\hat{\theta} = (\hat{\beta}, \hat{\alpha})$, i.e,

$$\frac{\partial^2 \log\, L(\theta)}{\partial \theta\, \partial \theta'}|_{\theta=\hat{\theta}} = \left[\frac{\partial^2 Q(\theta;\psi)}{\partial \theta\, \partial \theta'} + \frac{\partial^2 Q(\theta;\psi)}{\partial \theta\, \partial \psi'}\right]|_{\theta=\hat{\theta},\psi=\hat{\theta}}\ . \tag{13}$$

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,

$$\frac{\partial^2 Q(\theta;\psi)}{\partial\theta\,\partial\psi'}|_{\theta=\hat\theta,\psi=\hat\theta} \approx \frac{\left[\frac{\partial Q(\theta;\psi)}{\partial\theta} - \frac{\partial Q(\theta;\psi+\Delta\psi)}{\partial\theta}\right]}{\Delta\psi}|_{\theta=\hat\theta,\psi=\hat\theta}\ . \tag{14}$$

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 (13);

(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 (14) to estimate the second term in equation (13);

(4) Obtain the observed Fisher information by equation (13).

B.3 *Dealing with Missing Data in EM algorithm*

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

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

$$\sum_{m=1}^{M} w_{ijm}\left[Y_{ijk}^{obs}\log p_{km} + (1-Y_{ijk}^{obs})\log(1-p_{km})\right] \tag{15}$$

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

$$E\left[\sum_{m=1}^{M} I(\eta_{ij}=m)\log\Pr(Y_{ijk}^{miss}\,|\,\eta_{ij}=m)\,|Y_i\,;\beta^{(h)},\alpha^{(h)}\right]$$

$$= \sum_{m=1}^{M} w_{ijm}\left[p_{km}^{(h)}\log p_{km} + (1-p_{km}^{(h)})\log(1-p_{km})\right]\ , \tag{16}$$

where $p^{(h)}_{km} = \Pr(Y_{ijk} = 1|\eta_{ij} = m; \beta^{(h)}) = exp(\beta^{(h)}_{km})/\{1 + exp(\beta^{(h)}_{km})\}$ 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, $Y^{miss}_{ijk} = 1$ with probability $p^{(h)}_{km}$ and $Y^{miss}_{ijk} = 0$ with probability $1 - p^{(h)}_{km}$ for a member of the m^{th} class. Only the first term of the Q function changes when the data are missing.

B.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 MLC models, appropriate specification of AIC and BIC is challenged by the multilevel structure. Thus, we recommend an alternative method for model selection that avoids such difficulty. We described a subsampling procedure to select the number of classes for MLC-D model. Based on marginalization,

$$\begin{aligned} \Pr(\underset{\sim}{Y}_{ij} = \underset{\sim}{y}) &= \sum_{m=1}^{M} \Pr(\eta_{ij} = m) \cdot \prod_{k=1}^{K} p_{km}^{y_k}(1 - p_{km})^{1-y_k} \\ &= \sum_{m=1}^{M} \frac{\alpha_m}{\alpha_0} \cdot \prod_{k=1}^{K} p_{km}^{y_k}(1 - p_{km})^{1-y_k} \ , \end{aligned} \tag{17}$$

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 MLC 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 independent subsample. Standard methods, such as BIC, could then be used to choose the number of classes, say M, using the subsample. 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 $SS = \{(i, j_i) : i = 1, ..., n\}$, where j_i is a subject randomly chosen from all subjects $\{1, ..., n_i\}$ in cluster i;

(2) Fit a latent class model using sample SS 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 $\{\mathrm{BIC}_l^{(m)} : l = 1, ..., L, m = 1, ..., M^*\}$, where $\mathrm{BIC}_l^{(m)}$ 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 $\{B\bar{I}C^{(m)} : m = 1, ..., M^*\}$, where $B\bar{I}C^{(m)} = \sum_l \mathrm{BIC}_l^{(m)}/L$.

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.

Note that our approach is different from using the number of clusters as the sample size in the BIC formula. The BIC with the number of clusters as the sample size is

$$BIC_{nested} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log Pr(\underset{\sim}{Y}_{ij}; \hat{\beta}, \hat{\alpha}) + df \cdot \log n \,,$$

while our approach with one subsample induces formula

$$BIC_{subsample} = \sum_{i=1}^{n} \log Pr(\underset{\sim}{Y}_{ij_i}; \tilde{\beta}, \tilde{\alpha}) + df \cdot \log n \,,$$

where j_i is a randomly selected subject from cluster i, $j_i \in \{1, 2, \cdots, n_i\}$.

We then draw multiple copies of such subsamples and take averages of $BIC_{subsample}$'s to improve accuracy. Thus, essentially, we did not really propose new procedures for nested data. Rather, we created subsamples that contain independent subjects, and utilize BIC for standard LC models.

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 underestimate the number of classes in small samples (Yang, 2006). Thus the method above should be used with caution. In general, model selection on the number of classes for MLC models is a complex problem. A more comprehensive study on this issue would be possible future research directions and is out of scope of this paper.

B.5 *Estimation: Pairwise EM algorithm*

In this subsection, we briefly discuss estimation via maximum pairwise likelihood (Section 4 of the paper). 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,

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

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.

C. Sketch of the proof for Proposition 1 and 2

Both results in Proposition 1 and 2 can be obtained by using the general composite likelihood theory (Lindsay, 1988).

Sketch of proof for Proposition 1: (1). As pointed out by Lindsay (1988), each component of $l^p(\beta,\alpha)$ is a true log 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.

(2) and (3). Since the score functions of the simple likelihood are unbiased, one can 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.

Sketch of proof for Proposition 2: (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 marginal likelihood 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) can be proved using general estimating function theory.

In fact, the likelihood function derived from the standard LC model can also be viewed as a special case of the composite likelihood, since each component $\log f(\underset{\sim}{Y}_{ij};\beta,\pi)$ is a true marginal likelihood. Thus, consistency and asymptotic normality follows.

D. Simulation studies under *Settings II* and *III*

In this section, we show simulation results under *Settings II* and *III.* The aim of these simulation studies include: 1) evaluate finite sample performance of various MLC models/methods under more complex settings (3 classes and 8 items, compared to *Setting I* with 2 classes and 5 items); 2) assess the robustness of various MLC models/methods, e.g., the performance of MLC-D when the true model is MLC-V1 and vice versa.

The simulation settings will mimic the OCD application described in the main paper. The following two subsections report simulation results when the true models are MLC-D and MLC-V1, respectively. The true parameter values were taken to be the MLEs of corresponding models using the OCD data for each setting.

In the following simulation studies, the methods ML-D, MPL-D and ML-S were implemented using R codes written by the authors, while simulations via ML-V1 were implemented with the MPlus 5.21 software with 40 quadrature points for numerical integration and default stopping criteria.

D.1 *Setting II: the true model is MLC-D*

In this setting, data were generated from the following true settings: $n = 200$ clusters, $J = 4$ subjects per cluster, $K = 8$ items, $M = 3$ classes. The true model was the multilevel latent class model (6) with true parameters values being the MLEs from the OCD application (see Table 3 of the paper). We conducted 500 simulation runs, and in each run four methods were used to fit the multilevel latent class model, maximum likelihood for the MLC-D model (ML-D), maximum pairwise likelihood for the MLC-D model (MPL-D), maximum likelihood for simple latent class model with robust standard errors (ML-S) and maximum likelihood for the MLC-V1 model (ML-V1).

First we look at findings for estimation of the measurement models. Figures 3 and 4 display boxplots of estimated conditional probabilities, i.e., $p_{km} = \exp(\beta_{km})/\{1 + \exp(\beta_{km})\}$. The solid gray lines in each figure represent true parameter values. For all four methods, estimator distributions centered closely around true values, exhibited relatively small dispersion, and included few outliers. The dispersion of MPL-D was similar to that of ML-D, suggesting high relative efficiency of the MPL-D estimates. The dispersion of ML-S, however, was larger than that for ML-D or MPL-D, implying loss of efficiency by ignoring the within cluster correlation. The ML-V1, which fitted the MLC-V1 model when the data were generated from the MLC-D model, seemed to provide unbiased estimates for p_{km}'s with dispersion generally smaller than ML-S but larger than ML-D.

[Figure 3 about here.]

[Figure 4 about here.]

Turning to findings relating to the mixing distribution, the distributions of the α parameter estimates were centered around the true values using ML-D and MPL-D, according to the upper left panel of Figure 5. Note that the methods ML-S and ML-V1 do not contain such parameters and cannot be compared with the former two methods in terms of α parameters.

Researchers typically will be most interested in conveniently interpreted quantities, including the marginal class prevalences $(\pi_1, \cdots, \pi_M)$, and the intra-cluster correlation parameter ρ's. Figure 5 shows that the marginal class prevalences and the intra-cluster correlation were well estimated by the ML-D and MPL-D, with distributions centering around the true values and having narrow spreads. The MPL-D estimates enjoyed high finite-sample efficiency compared to ML-D estimates. The ML-S that ignores clustering could still estimate the MCPs $\underset{\sim}{\pi}$ consistently, although with less efficiency compared to the ML-D. However, the ML-S assumes a working independence correlation and could not provide unbiased estimates of within cluster ICCs. The ML-V1 method seemed to be able to estimate the MCPs well, even though the distribution of random effects u_i's are mis-specified. In terms of ICCs, however, such misspecification led to biased estimates and very large dispersions.

[Figure 5 about here.]

To summarize, simulation studies suggests that both ML-D and MPL-D well accomplish estimation and inference in finite samples under correct model specification. In our simulation settings that mimic the true OCD application, MPL-D enjoyed similar finite sample efficiency compared to ML-D. When the model is mis-specified as either LC-S (ignores clustering) or MLS-V1, the measurement parameters and marginal class prevalences could still be consistently estimated with less efficiency, while estimates the ICCs would be subject to bias and large dispersions.

D.2 *Setting III: the true model is MLC-V1*

The simulation setting here is similar to the previous one except that the true model was MLC-V1 instead of MLC-D. The true parameters were also taken to be the MLEs for the

OCD example, with $\lambda_2 = 1$, $\lambda_3 = 1.704$, $\gamma_2 = 1.542$, $\gamma = 0.011$ and $\sigma^2 = 24.154$. We conducted 500 simulation runs, and in each run four methods were used for estimation, ML-D, MPL-D, ML-S and ML-V1.

First we consider findings for estimation of the measurement models. Figures 6 and 7 display boxplots of estimated conditional probabilities, i.e., $p_{km} = \exp(\beta_{km})/\{1+\exp(\beta_{km})\}$. The solid gray lines in each figure represent true parameter values. For all four methods, estimator distributions centered closely around true values, exhibited relatively small dispersion, and included few outliers. The dispersion of ML-S, however, was generally larger, implying loss of efficiency by ignoring the within cluster correlation.

[Figure 6 about here.]

[Figure 7 about here.]

In terms of the mixing parts, only the ML-V1 method provided estimates for parameters γ_m's, λ_m's and σ^2 directly. The upper left panel of Figure 8 displays boxplots for estimates of σ^2 and λ_3 in logarithm scale. Although in theory the ML-V1 provide consistent estimates in large samples under correct model specification, its finite sample performance is not very good, at least in our simulation setting. More precisely, the estimates did not center around the true value and there were large dispersion. This is probably due to weak empirical identifiability of variance parameters, as is often the case in random effects models.

We also evaluate estimation of marginal class prevalences $(\pi_1, \cdots, \pi_M)$ and the intra-cluster correlation parameter ρ's and show results in Figure 8. The ML-V1, which correctly specified the underlying random effects distribution, estimated the MCPs and ICCs well, i.e., centering around the true values and having narrow spreads. The ML-S that ignores clustering could estimate the MCPs $\underset{\sim}{\pi}$ consistently, although with less efficiency compared to the ML-V1. However, the ML-S assumes a working independence correlation and could not provide unbiased estimates of within cluster ICCs. The ML-D and MPL-D methods seemed to estimate the MCPs well, even though the distribution of random effects u_i's are mis-specified. In terms of ICCs, however, such mis-specification led to biased estimates.

[Figure 8 about here.]

To summarize, simulation studies suggests that ML-V1 well estimate measurement parameters and important features of the mixing part (MCPs and ICCs) in finite samples when the true model is MLC-V1. However, estimation of variance component parameters (say σ^2) might not be accurate, perhaps due to weak empirical identifiability. That is, a wide range of σ^2 values induce similar within cluster dependence in finite samples.

When the model is mis-specified as either LC-S (ignores clustering) or MLC-D, the measurement parameters and marginal class prevalences could still be consistently estimated with less efficiency, while estimates the ICCs would be subject to bias.

E. On OCD application

Section 7 of the paper showed main results for the OCD application. In this section, we provided additional results, especially on the comparison between MLC-D and MLC-V1.

As mentioned in the paper, MLC-D and MLC-V1 provided similar estimates for the measurement part (β parameters). We now look at the mixing part, i.e., the distribution of random effects $\underset{\sim}{u}_i$'s . Figure 9 showed estimated pairwise and marginal distributions of u_{i1}, u_{i2}, and u_{i3}, from MLC-D and MLC-V1 models, respectively. First, from the marginal density functions, both models implied "U" shaped density with peaks near the boundary (0 or 1). However, one curious feature of the MLC-V1 is that the mixing probability for class 2, u_{i2}, is not allowed take any values above 0.702. This is a subtle consequence of the unidimensional factor analysis type structure for modelling dependence in MLC-V1. Second, from the pairwise distributions, it is clear that MLC-V1 restricts (u_{i1}, u_{i2}, u_{i3}) to take values only in a one-dimensional subspace of its domain $\Omega_u = \{(u_1, u_2, u_3) \in [0, 1]^3 : u_1 + u_2 + u_3 \leqslant 1\}$. In contrast, the Dirichlet model allows $\underset{\sim}{u}_i$'s to take values freely in Ω_u. Thus, we think that the MLC-D is more natural. However, we do not claim it is superior than other models. As the reviewers pointed out, the fact that MLC-V1 does not allow clusters with high class 2 prevalence is not relevant as long as model fitting is not compromised.

[Figure 9 about here.]

References

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

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

Erosheva, E. (2003). Bayesian estimation of the grade of membership model. *Bayesian Statistics* **7,** 501–510.

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

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

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

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

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

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

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

Varki, S. and Chintagunta, P. (2004). The augmented latent class model: Incorporating additional heterogeneity in the latent class model for panel data. *Journal of Marketing Research* **41,** 226–233.

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

Vermunt, J. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research* **17,** 33.

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

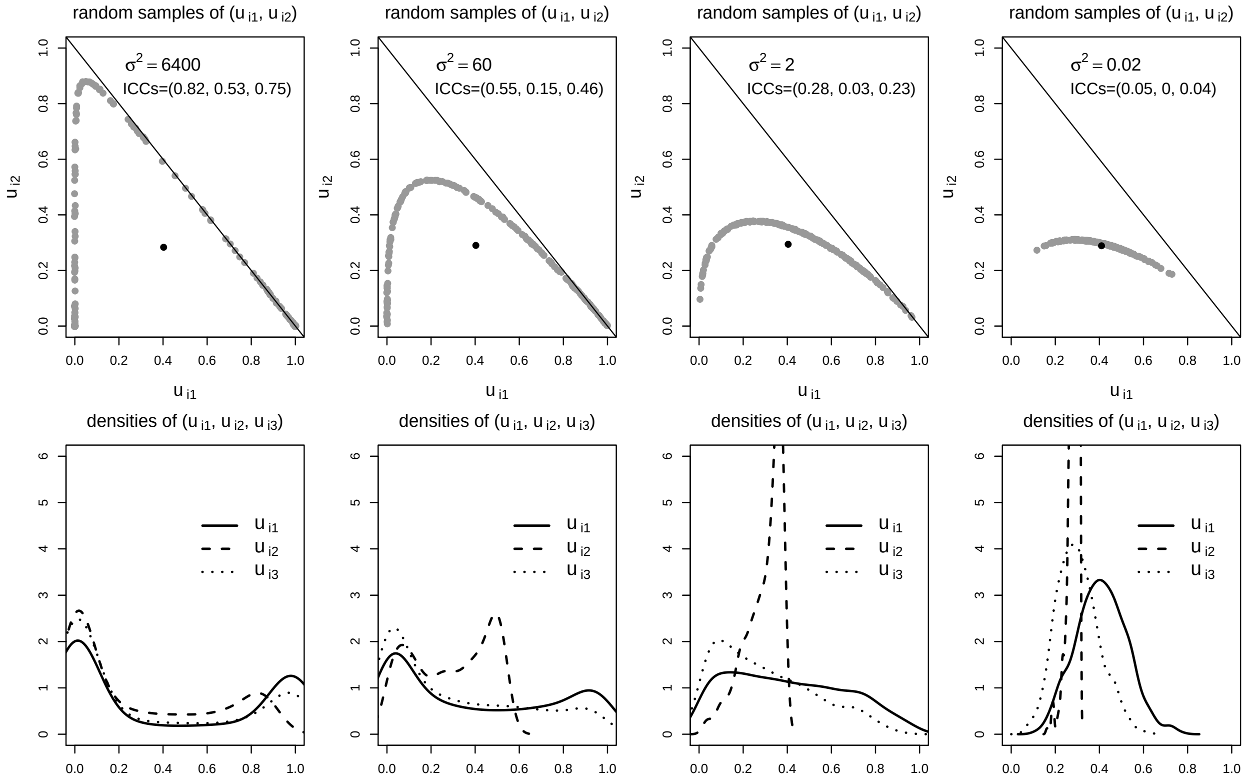

Figure 1. Distribution for class mixing random effects $u_i = (u_{i1}, u_{i2}, u_{i3})$ when the MLC-V1 model is true. In the four scenarios, the marginal mean of u_i is constant $(0.4, 0.27, 0.33)$, while ICCs vary from high (0.82) to low (0.04). For each scenario, the top panel displays 200 random samples of (u_{i1}, u_{i2}) and the bottom panel shows marginal density functions for u_i's.

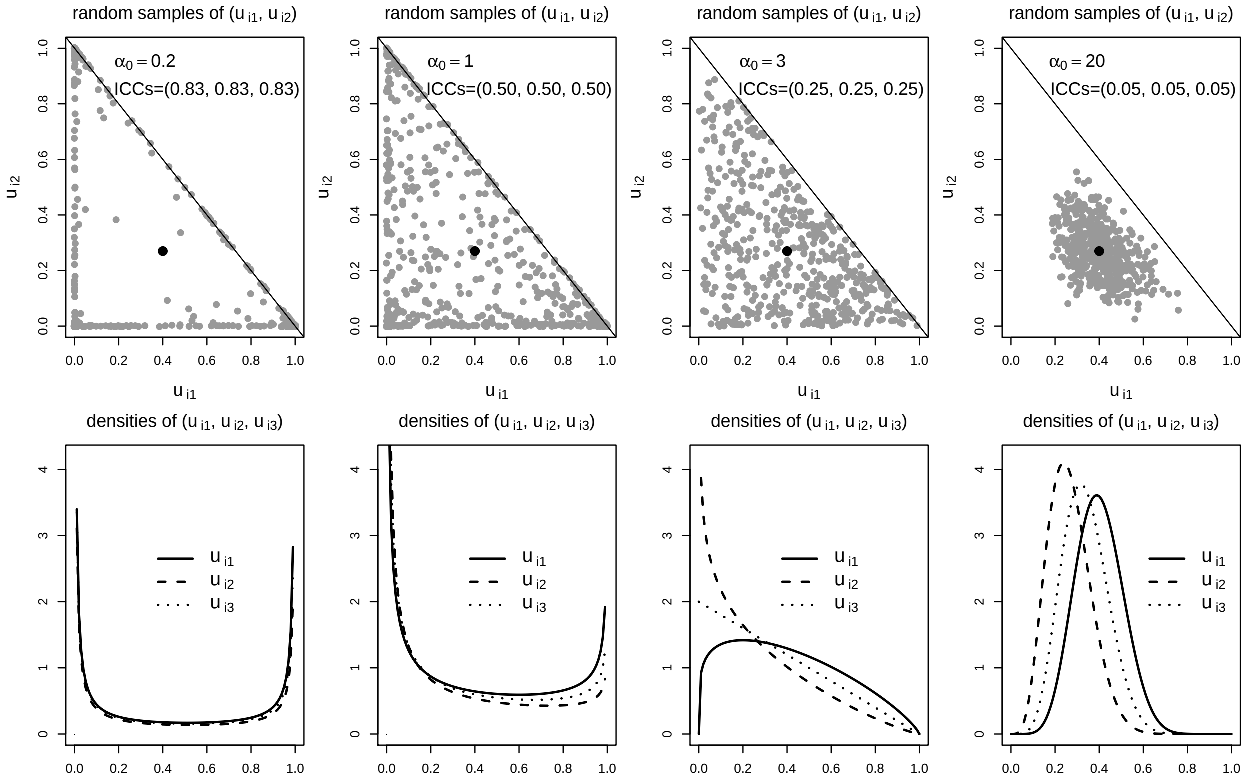

Figure 2. Distribution for class mixing random effects $u_i = (u_{i1}, u_{i2}, u_{i3})$ when the MLC-D model is true. In the four scenarios, the marginal mean of u_i is constant $(0.4, 0.27, 0.33)$, while ICCs vary from high (0.83) to low (0.05). For each scenario, the top panel displays 200 random samples of (u_{i1}, u_{i2}) and the bottom panel shows marginal density functions for u_i's.

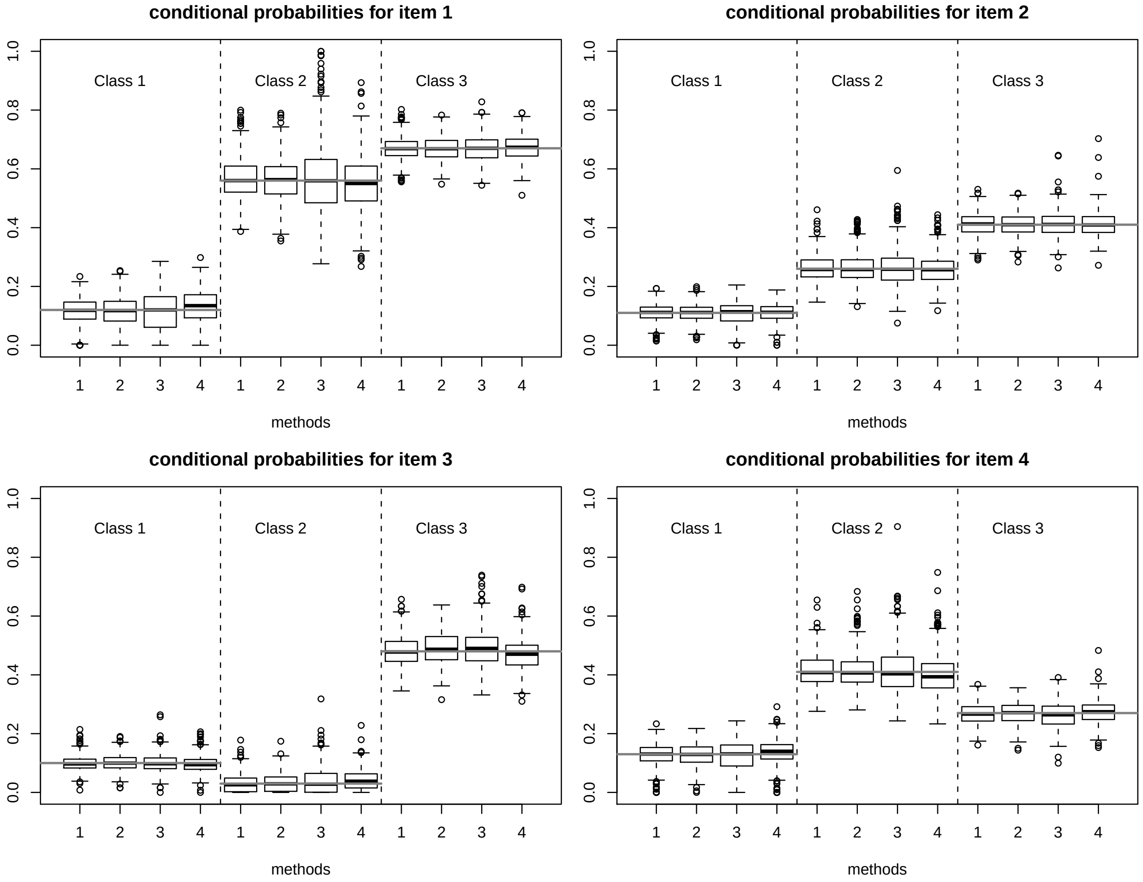

Figure 3. Estimates of conditional probabilities when the MCL-D model is true: items 1–4. Rows and Columns correspond to classes and items respectively. Methods 1, 2, 3 and 4 correspond to “ML-D”, “MPL-D”, ”ML-S” and “ML-V1”, respectively. The solid gray lines represent true values.

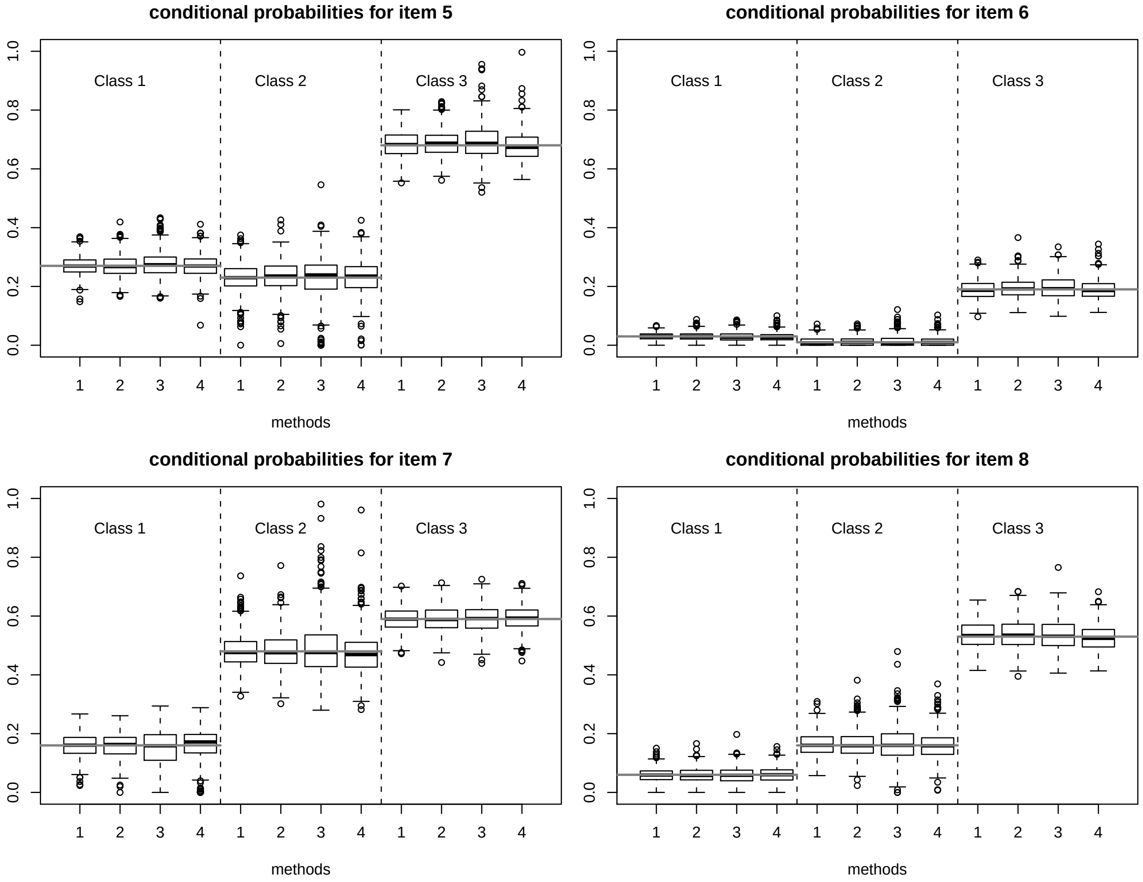

Figure 4. Estimates of conditional probabilities when the MLC-D model is true: items 5–8. Rows and Columns correspond to classes and items respectively. Methods 1, 2, 3 and 4 correspond to "ML-D", "MPL-D", "ML-S" and "ML-V1", respectively. The solid gray lines represent true values.

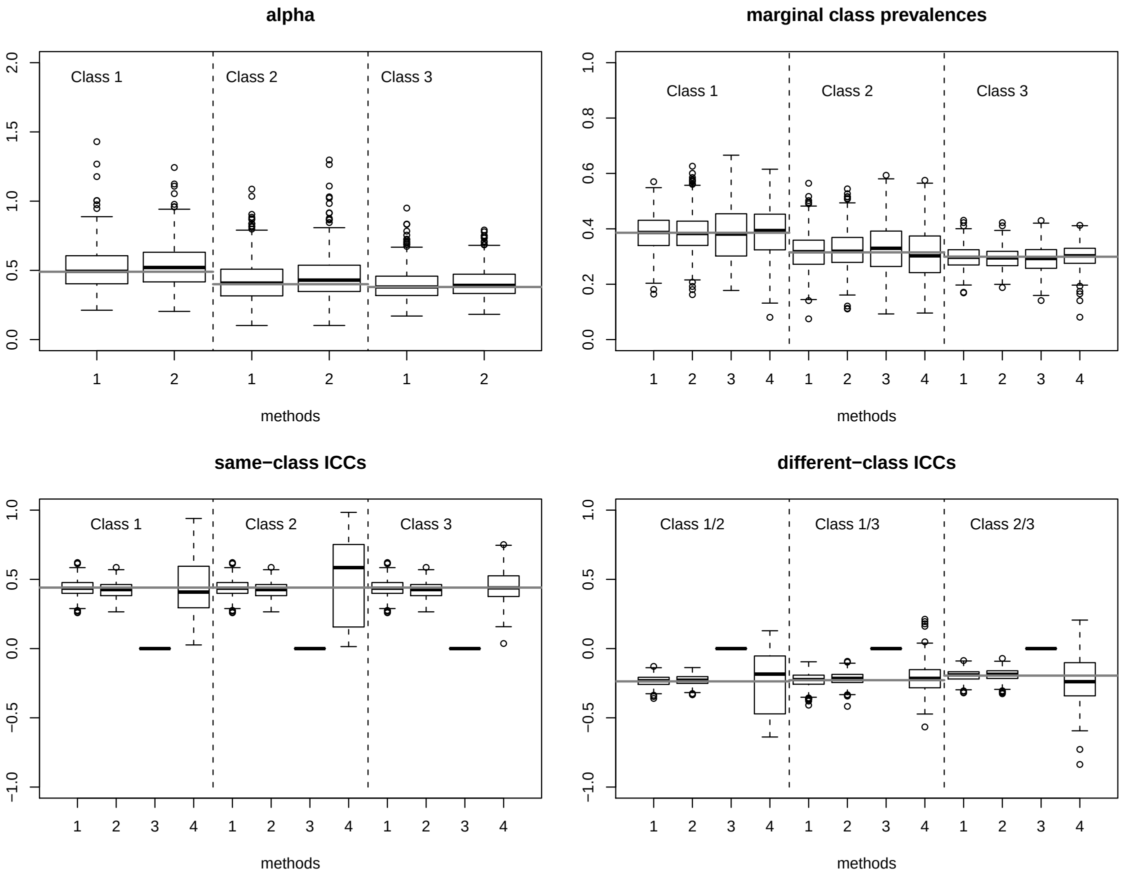

Figure 5. Mixing distribution model estimates when the MLC-D model is true. Methods 1, 2, 3 and 4 correspond to "ML-D", "MPL-D", "ML-S" and "ML-V1", respectively. The solid gray lines represent true values. In the first row, the left and right panels display boxplots of α estimates (only applicable to "ML-D" and "MPL-D") and marginal class prevalences for three classes, respectively. The second row contain boxplots for ICC estimates, including three same-class ICCs and three different-class ICCs.

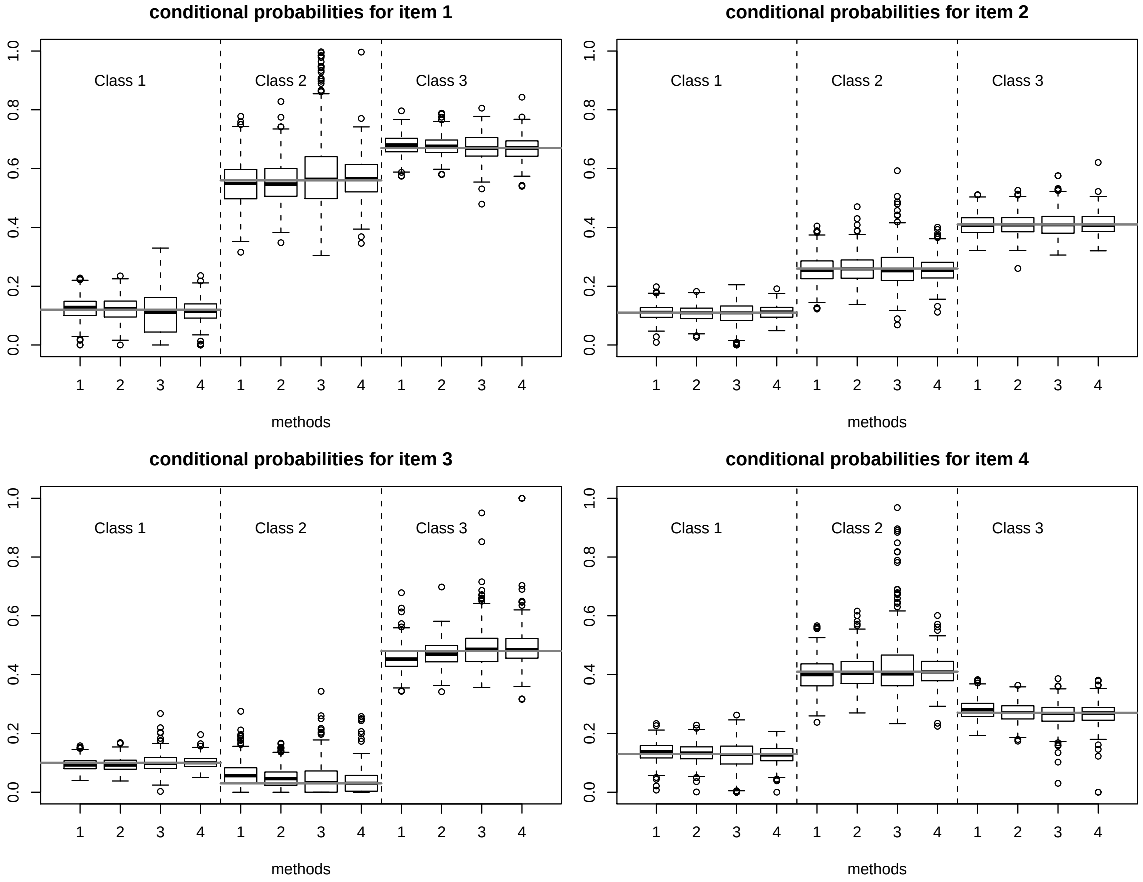

Figure 6. Estimates of conditional probabilities when the MLC-V1 model is true: items 1–4. Rows and Columns correspond to classes and items respectively. Methods 1, 2, 3 and 4 correspond to "ML-D", "MPL-D", "ML-S" and "ML-V1", respectively. The solid gray lines represent true values.

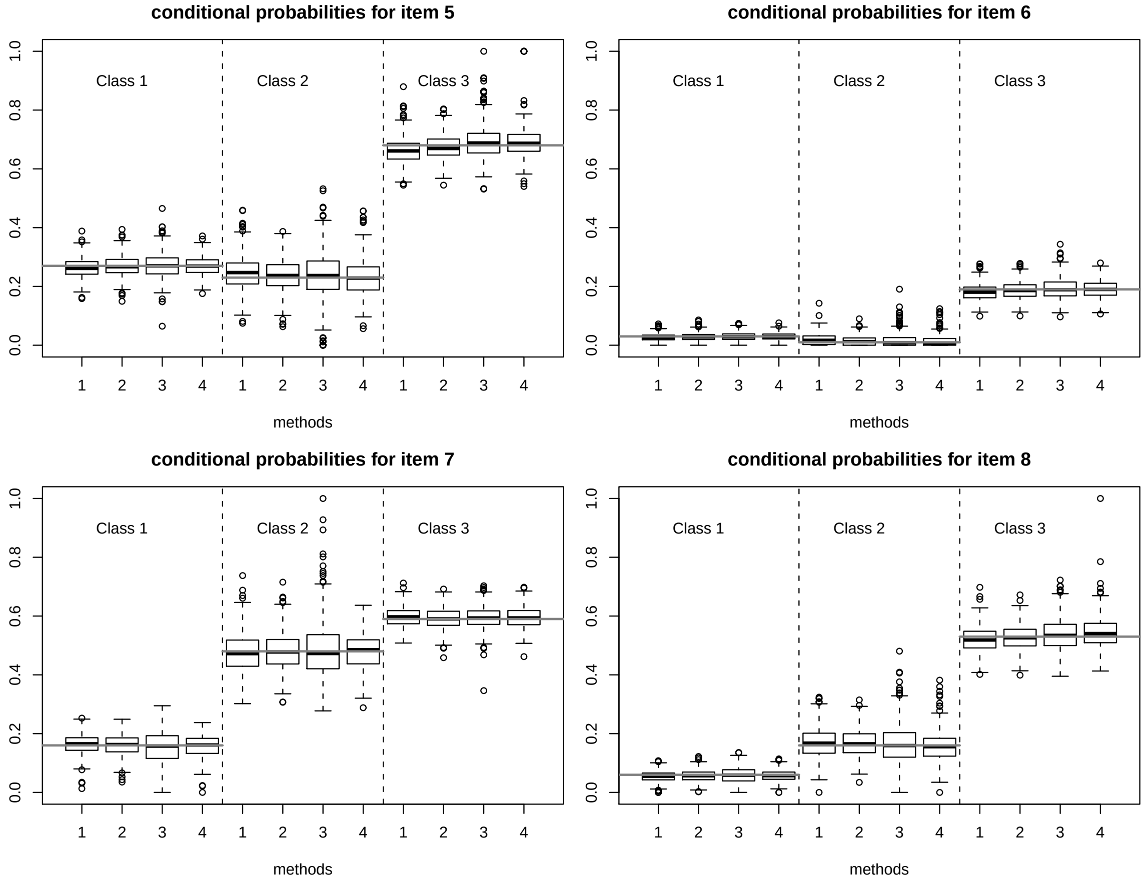

Figure 7. Estimates of conditional probabilities when the MLC-V1 model is true: items 5–8. Rows and Columns correspond to classes and items respectively. Methods 1, 2, 3 and 4 correspond to “ML-D”, “MPL-D”, ”ML-S” and “ML-V1”, respectively. The solid gray lines represent true values.

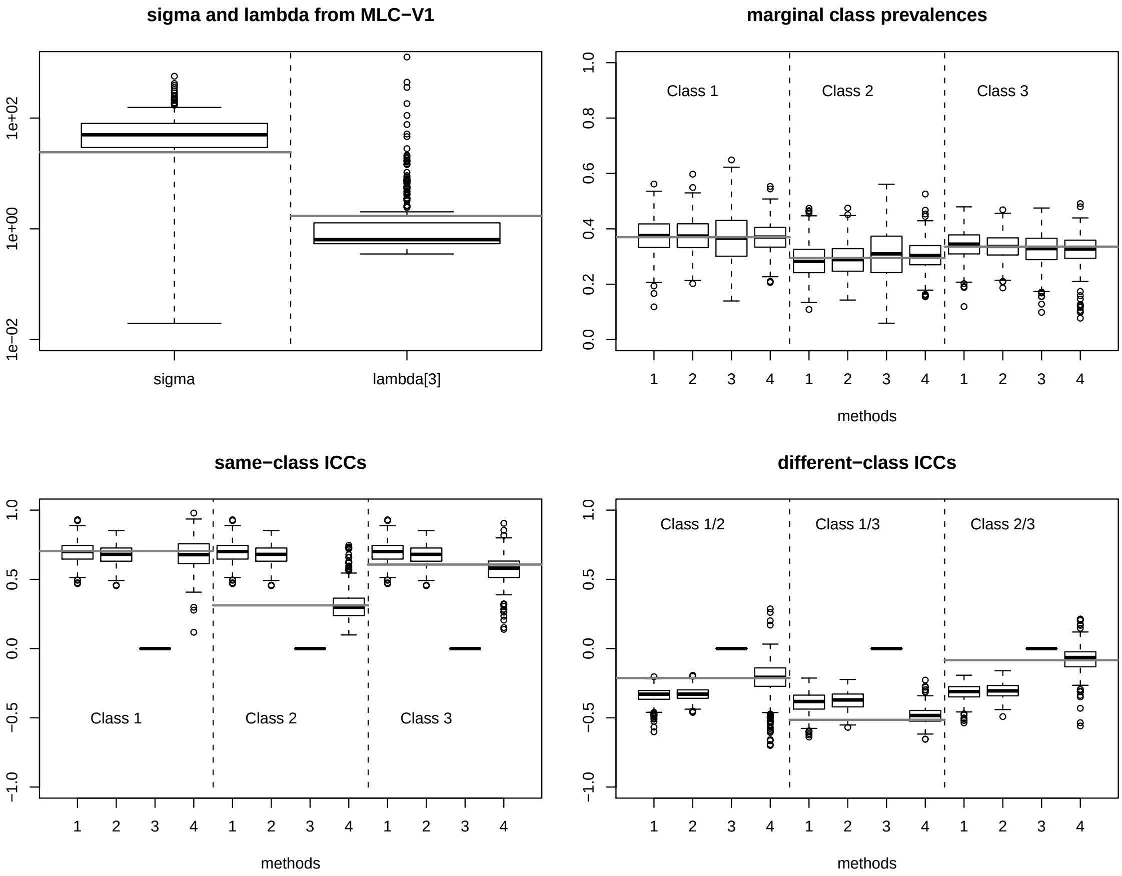

Figure 8. Mixing distribution model estimates when the MLC-V1 model is true. Methods 1, 2, 3 and 4 correspond to "ML-D", "MPL-D", "ML-S" and "ML-V1", respectively. The solid gray lines represent true values. In the first row, the left panel displays boxplots of σ and λ_3 estimates (only applicable to "ML-V1"; note that $\lambda_1 = 0$ and $\lambda_2 = 1$ are not free parameters for a 3-class model) and the right panel displays marginal class prevalences for three classes, respectively. The second row contain boxplots for ICC estimates, including three same-class ICCs and three different-class ICCs.

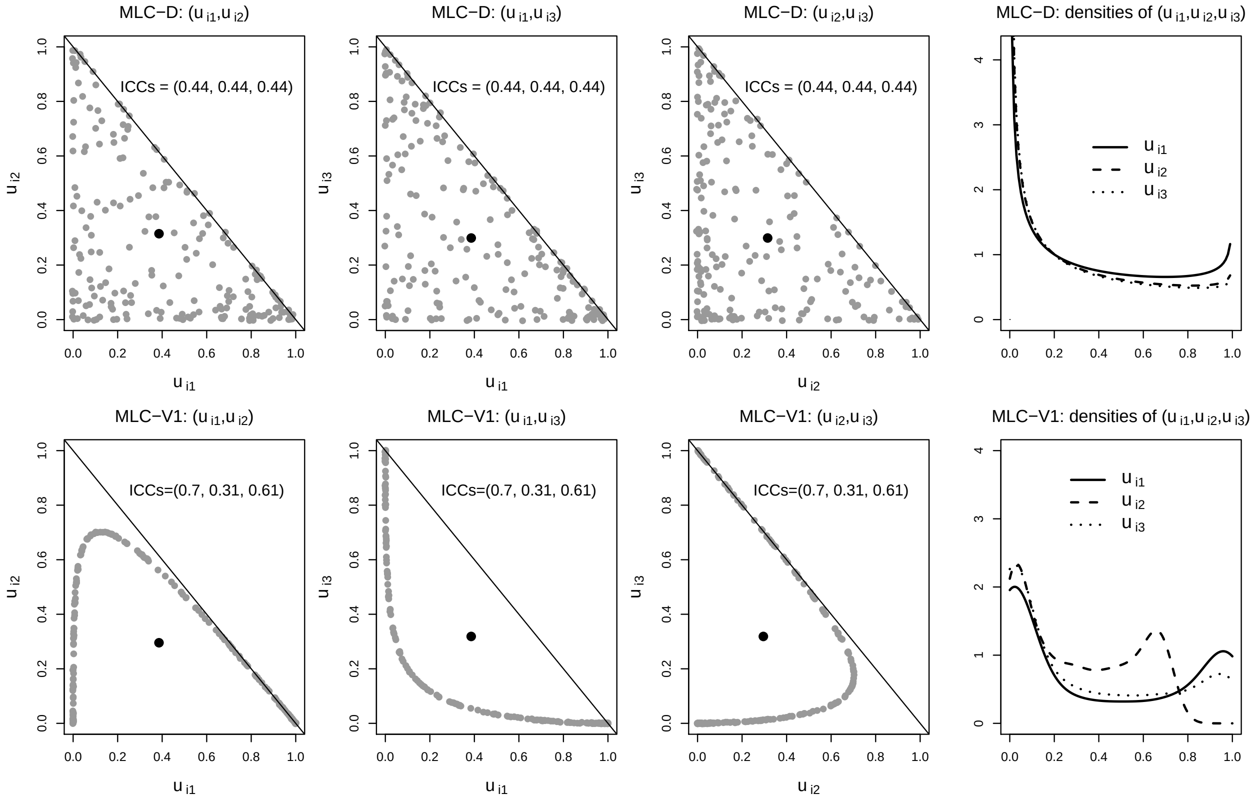

Figure 9. OCD data: distribution of class mixing random effects $u_i = (u_{i1}, u_{i2}, u_{i3})$ from both MLC-D and MLC-V1 models. In each row, the first three panels display 200 randomly generated samples for pairs (u_{i1}, u_{i2}), (u_{i1}, u_{i3}) and (u_{i2}, u_{i3}), respectively, and the fourth panel shows marginal probability density functions for u_{i1}, u_{i2} and u_{i3}, respectively.