Model selection in generalized linear finite mixture regression models by Hausman testing

Francesco Bartolucci

bart@stat.unipg.it
Department of Economics - University of Perugia (IT)

MBC - Catania, September 2014

Joint work with Silvia Bacci and Claudia Pigini (Department of Economics - UNIPG)
Outline

1 Motivation

2 Class of models of interest
   - Base-line model
   - Extended models

3 Estimation methods
   - Marginal Maximum Likelihood (MML)
   - Conditional Maximum Likelihood (CML)

4 Hausman-type test of misspecification

5 Simulation study

6 Applications
   - Example in IRT (educational NAEP data)
   - Multilevel data (contraceptive use in Bangladesh)
   - Longitudinal data (HRS data)

7 Conclusions
**Motivation**

- **Generalized Linear Mixed Models** (GLMMs; Zeger and Karim, 1991; McCulloch et al., 2008) represent a very useful instrument for the analysis of clustered data.

- **Applications**:
  - Item Response Theory (IRT)
  - Multilevel data where individuals are collected in groups
  - Longitudinal/panel data (repeated responses)

- We focus on the relevant case of *binary responses* and then on the (random-effects) logistic regression model (Stiratelli et al., 1984) and the extension of this model to deal with *ordinal data* (McCullagh, 1980).

- The random-effects included in a GLMM are typically assumed to have a normal distribution.
The study of the *consequences of the normality assumption* has considerable attention especially for the logistic regression model (less attention on linear models).

Some studies (e.g. Neuhaus et al., 1992) report that the effect of the normality assumption is *moderate* when this assumption is not true.

More recent studies conclude that the impact *may be considerable* on the quality of the estimates and random-effects prediction (e.g. Heagerty, 1999; Rabe-Hesketh et al., 2003; Agresti et al., 2004).

A flexible way to formulate the distribution of the random-effects is based on assuming a *discrete distribution* that leads to a finite mixture model.

This approach is seen as *semiparametric* and it is strongly related to the nonparametric maximum likelihood approach (Kiefer and Wolfowitz, 1956; Laird, 1978; Lindsay, 1983).
Motivation

**Relevant applications:**
- Lindsay et al. (1991) in the IRT context
- Aitkin (1999) in the general context of clustered data
- Vermunt (2003) specifically in the context of multilevel data
- Heckman and Singer (1984) for a flexible model for survival data
- Aitkin (1996) to create overdispersion in a generalized linear model

**Other pros** of the finite mixture approach for GLMMs:
- it avoids complex computational methods to integrate out the random-effects
- it leads to a natural clustering of sample units that may be of main interest for certain relevant applications (e.g., Deb, 2001) as in a latent class model (Lazarsfeld and Henry, 1968; Goodman, 1974)

**Cons:**
- difficult interpretation in certain contexts (when random-effects represent missing covariates seen as continuous)
- need to choose the number of mixture components
- some instability problems in estimation also due to the multimodality of the likelihood function that often arises
The **finite mixture approach** is the main alternative to the normal approach to formulate the distribution of the random-effects for GLMMs (in particular for logistic regression models).

**Testing** the hypothesis that the mixing distribution is normal has attracted considerable attention in the recent statistical literature.

**Available approaches:**
- empirical Bayes estimates of the individual effects (Lange and Ryan, 1989), but criticized for the lack of power
- method based on residuals (Ritz, 2004; Pan and Lin, 2005)
- simulating the random-effects from their posterior distribution given the observed data (Waagepetersen, 2006)
- comparing marginal and conditional maximum likelihood estimates (Tchetgen and Coull, 2006)
- methods based on the covariance matrix of the parameter estimates and the information matrix (Alonso et al., 2008, 2010)
- method based on the gradient function (Verbeke and Molenberghs, 2013)

**No approaches** seem to be tailored to the case of finite mixture GLMMs.
We develop the approach of Tchetgen and Coull (2006) for logistic models, for binary and ordinal responses, with random-effects assumed to have a discrete distribution (finite mixture).

The approach is based on the comparison of conditional and marginal maximum likelihood estimates for the fixed effects, as in the Hausman’s test (Hausman, 1978).

Since none of the two estimators compared is ensured to be fully efficient, we use a generalized estimate of the variance-covariance matrix of the difference between the two estimators (Bartolucci et al., 2014).

The proposed test may also be used to select the number of support points of the discrete distribution (or mixture components).

With longitudinal data, the proposed test can be used in connection with that proposed by Bartolucci et al. (2014) to test the assumption that the random-effects are time constant rather than time varying.
Base-line model

**Basic notation:**
- \( n \): number of clusters (individuals in the case of longitudinal studies or IRT)
- \( J_i \): number of observations for cluster \( i \)
- \( y_i = (y_{i1}, \ldots, y_{iJ_i}) \): binary observations for cluster \( i \)
- \( x_i \): column vector of cluster-specific covariates
- \( z_{ij} \): column vector of observation-specific covariates

**Base-line model:**

\[
\log \frac{p(y_{ij} = 1|\alpha_i, x_i, z_{ij})}{p(y_{ij} = 0|\alpha_i, x_i, z_{ij})} = \alpha_i + x_i'\beta + z_{ij}'\gamma, \quad i = 1, \ldots, n, \ j = 1, \ldots, J_i
\]

\( \alpha_i \) are random-effects that in the standard case have a normal distribution with unknown variance \( \sigma^2 \)

We assume that the random-effects have a **discrete distribution** with:
- \( k \) support points \( \xi_1, \ldots, \xi_k \)
- mass probabilities \( \pi_1, \ldots, \pi_k \), where \( \pi_h = p(\alpha_i = \xi_h) \)
Local independence is also assumed: conditional independence between the responses given the random-effects and the covariates.

With ordinal response variables $y_{ij}$ having $L$ categories ($0, \ldots, L - 1$), the model is formulated as (Model-ord1):

$$\log \frac{p(y_{ij} \geq l | \alpha_i, x_i, z_{ij})}{p(y_{ij} < l | \alpha_i, x_i, z_{ij})} = \alpha_i + \delta_l + x_i'\beta + z_{ij}'\gamma, \quad l = 1, \ldots, L - 1,$$

with cutpoints $\delta_1 > \cdots > \delta_{L-1}$

A more general formulation is based on individual-specific cutpoints (Model-ord2):

$$\log \frac{p(y_{ij} \geq l | \alpha_i, x_i, z_{ij})}{p(y_{ij} < l | \alpha_i, x_i, z_{ij})} = \alpha_{il} + x_i'\beta + z_{ij}'\gamma, \quad l = 1, \ldots, L - 1,$$

with cutpoints $\alpha_{i1} > \cdots > \alpha_{i,L-1}$ collected in the vectors $\alpha_i$. 
The first two models may be interpreted in terms of an underlying continuous variable and a suitable observation rule:

\[ y_{ij} = G(y^*_ij), \quad y^*_ij = \alpha_i + x'_i \beta + z'_{ij} \gamma + \varepsilon_{ij}, \]

with \( \varepsilon_{ij} \) being i.i.d. error terms with standard logistic distribution.

With binary responses, the observation rule is

\[ G(y^*_ij) = I\{y^*_ij > 0\}, \]

where \( I\{\cdot\} \) is an indicator function.

With ordinal responses (Model-ord1), the observation rule is

\[
G(y^*_ij) = \begin{cases} 
0 & y^*_ij \leq \tilde{\delta}_1, \\
1 & \tilde{\delta}_1 < y^*_ij \leq \tilde{\delta}_2, \\
\vdots & \vdots \\
L-1 & y^*_ij > \tilde{\delta}_{L-1} 
\end{cases}
\]
Extended models

- The model may be *extended* also to account for the dependence of each $\alpha_i$ on a vector of cluster-specific covariates $w_i$ (to face *endogeneity*).

- **1st possible extension**: an interaction term is included as

$$
\log \frac{p(y_{ij} = 1|\alpha_i, w_i, x_i, z_{ij})}{p(y_{ij} = 0|\alpha_i, w_i, x_i, z_{ij})} = w_i'\alpha_i + x_i'\beta + z_{ij}'\gamma, \quad i = 1, \ldots, n, \ j = 1, \ldots, J_i,
$$

with the vectors of random-effects $\alpha_i$ having $k$ support points $\xi_1, \ldots, \xi_k$ and mass probabilities $\pi_h = p(\alpha_i = \xi_h)$.

- **2nd possible extension**: the mass probabilities depend on the covariates by a multinomial logit parameterization:

$$
\log \frac{p(\alpha_i = \xi_{h+1}|w_i)}{p(\alpha_i = \xi_1|w_i)} = \phi_h + w_i'\psi_h, \quad h = 1, \ldots, k - 1, \ i = 1, \ldots, n,
$$

or alternative parametrizations when the support points are ordered.
Marginal Maximum Likelihood (MML)

- For the base-line model, the assumption of *local independence* implies

\[ p(y_i | \alpha_i, x_i, Z_i) = \prod_j p(y_{ij} | \alpha_i, x_i, z_{ij}) \]

with \( Z_i = (z_{i1}, \ldots, z_{iJ_i}) \) being the matrix of covariates varying within cluster

- The *manifest distribution* is

\[ p(y_i | x_i, Z_i) = \sum_h \left[ \prod_j p(y_{ij} | \xi_h, x_i, z_{ij}) \right] \pi_h \]

- The *marginal log-likelihood function* is

\[ \ell_M(\theta) = \sum_i \log p(y_i | x_i, Z_i) = \sum_i \log \sum_h \left[ \prod_j p(y_{ij} | \xi_h, x_i, z_{ij}) \right] \pi_h \]

with \( \theta \) denoting the overall vector of parameters
Maximization of $\ell_M(\theta)$ may be efficiently performed by an Expectation Maximization (EM) algorithm (Dempster et al., 1977)

The EM algorithm is based on the complete-data log-likelihood function

$$\ell^*_M(\theta) = \sum_i a_{hi} \left[ \log \pi_h + \sum_j \log p(y_{ij} | \xi_h, x_i, z_{ij}) \right],$$

with $a_{hi}$ being an indicator variable equal to 1 if $\alpha_i = \xi_h$ and to 0 otherwise.

The **algorithm** alternates two steps until convergence:

- **E-step**: compute the posterior expected value of each $a_{hi}$ which is equal to the posterior probability $\hat{a}_{hi} = p(\alpha_i = \xi_h | x_i, y_i, Z_i)$
- **M-step**: maximize the function $\ell^*_M(\theta)$ with each $a_{hi}$ substituted by $\hat{a}_{hi}$
The **asymptotic variance-covariance matrix** of the MML estimator $\hat{\theta}_M$ may be estimated by the sandwich formula

\[
\hat{V}_M(\hat{\theta}_M) = H_M(\hat{\theta}_M)^{-1} V_M(\hat{\theta}_M) H_M(\hat{\theta}_M)^{-1}
\]

\[
u_{M,i}(\theta) = \frac{\partial \log p(y_i|\alpha_i, x_i, Z_i)}{\partial \theta}
\]

\[
H_M(\theta) = \sum_i \frac{\partial^2 \log p(y_i|x_i, Z_i)}{\partial \theta \partial \theta'}
\]

\[
V_M(\theta) = \sum_i \nu_{M,i}(\theta) [\nu_{M,i}(\theta)]'
\]

The MML approach is easily adapted to estimate **extended models** with endogeneity.
 Conditional Maximum Likelihood (CML)

- The CML method (Andersen, 1970; Chamberlain, 1980) may be used to consistently estimate the parameters $\gamma$ for the covariates in $Z_i$ under mild assumptions (mainly time-constant individual effects).

- For binary data, the conditional log-likelihood function has expression

$$\ell_C(\gamma) = \sum_i \log p(y_i|y_{i+}, Z_i), \quad y_{i+} = \sum_{j=1}^{J} y_{ij},$$

with

$$p(y_i|Z_i, y_{i+}) = \frac{\exp \left( \sum_j y_{ij} z_{ij}' \gamma \right)}{\sum_{s \in S_{J_i}(y_{i+})} \exp \left( \sum_j s_j z_{ij}' \gamma \right)},$$

where the sum $\sum_{s \in S_{J_i}(y_{i+})}$ is extended to all binary vectors $s = (s_1, \ldots, s_{J_i})$ with sum equal to $y_{i+}$.

- $p(y_i|Z_i, y_{i+})$ does not depend anymore on $\alpha_i$ and $x_i$ (and possibly $w_i$).
Estimation methods

- $\ell_C(\beta)$ is simply maximized by a \textit{Newton-Raphson algorithm} based on the score vector

$$u_C(\gamma) = \sum_i u_{C,i}(\gamma), \quad u_{C,i}(\gamma) = \frac{\partial \log p(y_i|y_{i+}, Z_i)}{\partial \gamma}$$

and Hessian matrix

$$H_C(\gamma) = \sum_i \frac{\partial^2 \log p(y_i|y_{i+}, Z_i)}{\partial \gamma \partial \gamma'}$$

- The \textit{asymptotic variance-covariance matrix} may be obtained as

$$\hat{V}_C(\hat{\gamma}_C) = H_C(\hat{\gamma}_C)^{-1} V_C(\hat{\gamma}_C) H_C(\hat{\gamma}_C)^{-1}$$

$$V_C(\gamma) = \sum_i u_{C,i}(\gamma)[u_{C,i}(\gamma)]'$$
With **ordinal variables**, CML estimation is based on all the possible dichotomizations of the response variables (Baetschmann et al., 2011):

\[ y_{ij}^{(l)} = I\{y_{ij} \geq l\}, \quad j = l, \ldots, L - 1, \]

with \( y_{i}^{(l)} = (y_{i1}^{(l)}, \ldots, y_{ij}^{(l)}) \)

The corresponding **pseudo log-likelihood** function is

\[ \ell_C(\gamma) = \sum_i \sum_l \log p(y_{i}^{(l)} | y_{i+}, Z_i), \quad y_{i+} = \sum_{j=1}^{J} y_{ij}^{(l)}, \]

that may be maximized by a simple extension of the Newton-Raphson algorithm implemented for the binary case
Hausman-type test of misspecification

The test is based on the *comparison between the MML and the CML estimators* of $\gamma$ as in Tchetgen and Coull (2006) and Bartolucci et al. (2014)

The test exploits the *asymptotic distribution*

$$\sqrt{n}(\hat{\gamma}_M - \hat{\gamma}_C) \xrightarrow{d} N(0, W)$$

where $\hat{\gamma}_M$ is taken from $\hat{\theta}_M$ and the variance-covariate matrix $W$ is consistently estimated as

$$\hat{W} = n D \hat{V} (\hat{\theta}_M, \hat{\gamma}_C) D', \quad D = (E, -I),$$

with $E$ defined so that $\hat{\gamma}_M = E \hat{\theta}_M$
The variance-covariates matrix of \((\hat{\theta}_M, \hat{\gamma}_C)\) is obtained as

\[
\hat{V}(\hat{\theta}_M, \hat{\gamma}_C) = \left( \begin{array}{cc} H_M(\hat{\theta}_M) & 0 \\ 0 & H_C(\hat{\gamma}_C) \end{array} \right)^{-1} S(\hat{\theta}_M, \hat{\gamma}_C) \left( \begin{array}{cc} H_M(\hat{\theta}_M) & 0 \\ 0 & H_C(\hat{\gamma}_C) \end{array} \right)^{-1}
\]

\[
S(\hat{\theta}_M, \hat{\gamma}_C) = \sum_i \begin{pmatrix} u_{M,i}(\hat{\theta}_M) \\ s_{C,i}(\hat{\gamma}_C) \end{pmatrix} \begin{pmatrix} u_{M,i}(\hat{\theta}_M)' \\ s_{C,i}(\hat{\gamma}_C)' \end{pmatrix}
\]

The test statistic is

\[
T = n(\hat{\gamma}_M - \hat{\gamma}_C)' \hat{W}^{-1} (\hat{\gamma}_M - \hat{\gamma}_C)
\]

that has asymptotic null distribution of \(\chi^2_g\), with \(g\) being the dimension of \(\gamma\) (i.e., number of covariates varying within cluster)
The method extends the original method of Hausman (1978) because a generalized form for the variance-covariance matrix is used; this has advantages of stability and avoids to require that one of the two estimators is efficient (Vijverberg, 2011)

The proposed test may be simply used also to select the number of mixture components ($k$) when this number is unknown: $k$ is increased until the test does not stop to reject

We expect that the selection criterion for $k$ based on $T$ is more parsimonious with respect to available criteria when the random-effects are independent of the covariates
Simulation study

- Limited to the model for binary responses, we performed a simulation study for the case of the distribution correctly specified and for the case it is misspecified.

- The first model considered in the simulation is based on the assumption

\[
\log \frac{p(y_{ij} = 1 | \alpha_i, x_i, z_i)}{p(y_{ij} = 0 | \alpha_i, x_i, z_i)} = \alpha_i + x_i \beta + z_{ij} \gamma
\]

with \( \beta = \gamma = 1 \), \( \alpha_i \) having distribution

\[
\alpha_i = \begin{cases} 
-\sqrt{3}/2, & 0.25, \\
0, & 0.50, \\
\sqrt{3}/2, & 0.25, 
\end{cases}
\]

\( x_i \sim N(0, 1) \) and \( z_{ij} \) is generated as an AR(1) process with correlation coefficient \( \rho = 0.5 \) and variance \( \pi^2/3 \).
The proposed test for \( k = 3 \) support points has the expected behavior in terms of actual rejection rate:

<table>
<thead>
<tr>
<th></th>
<th>( J = 5 )</th>
<th></th>
<th>( J = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n = 500 )</td>
<td>( n = 1000 )</td>
<td></td>
</tr>
<tr>
<td>( \alpha = 0.10 )</td>
<td>0.103</td>
<td>0.097</td>
<td></td>
</tr>
<tr>
<td>( \alpha = 0.05 )</td>
<td>0.057</td>
<td>0.060</td>
<td></td>
</tr>
<tr>
<td>( \alpha = 0.01 )</td>
<td>0.018</td>
<td>0.018</td>
<td></td>
</tr>
</tbody>
</table>

The second model is a *Rasch model* based on the assumption

\[
\log \frac{p(y_{ij} = 1 | \alpha_i)}{p(y_{ij} = 0 | \alpha_i)} = \alpha_i - \gamma_j,
\]

with \( \gamma_j \) being the difficulty level of item \( j \); these parameters are taken as equidistant points in the interval \([-2, 2]\).

Even in this case the *nominal significance level* is attained with very similar results as for the first model.
We repeated the simulation study under the assumption $\alpha_i \sim N(0, 3)$ and compared the proposed criterion for choosing $k$ based on the proposed test with:

\begin{align*}
\text{AIC} &= -2 \ell_M(\hat{\theta}_M) + 2\#\text{par} \\
\text{BIC} &= -2 \ell_M(\hat{\theta}_M) + \#\text{par} \log(n) \\
\text{AIC}_3 &= -2 \ell_M(\hat{\theta}_M) + 3\#\text{par} \\
\text{CAIC} &= -2 \ell_M(\hat{\theta}_M) + \#\text{par} (\log(n) + 1) \\
\text{HT-AIC} &= -2 \ell_M(\hat{\theta}_M) + 2\#\text{par} + \frac{2(\#\text{par} + 1)(\#\text{par} + 2)}{n - \#\text{par} - 2} \\
\text{AIC}_c &= -2 \ell_M(\hat{\theta}_M) + 2\frac{\#\text{par}(\#\text{par} - 1)}{n - \#\text{par} - 1} \\
\text{BIC}^* &= -2 \ell_M(\hat{\theta}_M) + \#\text{par} \log \frac{n + 2}{24} \\
\text{CAIC}^* &= -2 \ell_M(\hat{\theta}_M) + \#\text{par} \left( \log \frac{n + 2}{24} + 1 \right)
\end{align*}
The proposed procedure leads to a more parsimonious model with respect to the other criteria; for instance, on 1,000 samples generated from the first model with \( n = 500 \) and \( J = 5 \), we have:

<table>
<thead>
<tr>
<th>( k )</th>
<th>Haus (10%)</th>
<th>Haus (5%)</th>
<th>Haus (1%)</th>
<th>AIC</th>
<th>BIC</th>
<th>AIC(_3)</th>
<th>CAIC</th>
<th>HT-AIC</th>
<th>AIC(_c)</th>
<th>BIC*</th>
<th>CAIC*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>415</td>
<td>597</td>
<td>841</td>
<td>5</td>
<td>152</td>
<td>17</td>
<td>223</td>
<td>5</td>
<td>0</td>
<td>17</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>398</td>
<td>297</td>
<td>122</td>
<td>627</td>
<td>815</td>
<td>764</td>
<td>757</td>
<td>635</td>
<td>124</td>
<td>774</td>
<td>831</td>
</tr>
<tr>
<td>4</td>
<td>59</td>
<td>40</td>
<td>19</td>
<td>355</td>
<td>33</td>
<td>216</td>
<td>20</td>
<td>347</td>
<td>550</td>
<td>206</td>
<td>124</td>
</tr>
<tr>
<td>5</td>
<td>64</td>
<td>35</td>
<td>11</td>
<td>13</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>13</td>
<td>249</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The same tendency is confirmed in many other cases and even when the true distribution is discrete.
The last simulation concerns the case of *endogeneity* in which we generate continuous random-effects as

\[ \alpha_i = \tau \bar{z}_i + \eta_i \sqrt{1 - \tau^2}, \]

where \( \bar{z}_i = \left(\frac{1}{J_i}\right) \sum_j z_{ij} \), \( \eta_i \sim N(0, 1) \), and \( \tau = 0, 0.5, 0.8 \).

The results of the test for \( k = 3 \) confirms the power of the test already with small sample sizes for \( \alpha = 0.1 \) (solid line), \( \alpha = 0.05 \) (dashed line), and \( \alpha = 0.01 \) (dotted line).
We considered *three empirical examples* in different fields:

- **IRT data**: the number of support points chosen by BIC is confirmed
- **multilevel data**: a smaller number of support points is chosen with respect to BIC
- **longitudinal data**: more support points and a different model specification are chosen with respect to BIC
Example in IRT (educational NAEP data)

- Data referred to a sample of 1510 examinees who responded to 12 binary items on Mathematics; source: National Assessment of Educational Progress (NAEP), 1996

- The test confirms the choice of \( k = 3 \) classes for the Rasch model suggested by BIC and other criteria:

<table>
<thead>
<tr>
<th></th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 3 )</th>
<th>( k = 4 )</th>
<th>( k = 5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausman ( T )</td>
<td>414.850</td>
<td>90.071</td>
<td>6.721</td>
<td>2.895</td>
<td>1.639</td>
</tr>
<tr>
<td>Hausman ( p )-value</td>
<td>0.000</td>
<td>0.000</td>
<td><strong>0.821</strong></td>
<td>0.992</td>
<td>0.999</td>
</tr>
<tr>
<td>AIC</td>
<td>22042.3</td>
<td>20511.4</td>
<td>20364.6</td>
<td><strong>20361.8</strong></td>
<td>20365.0</td>
</tr>
<tr>
<td>BIC</td>
<td>22106.2</td>
<td>20585.9</td>
<td><strong>20449.7</strong></td>
<td>20457.6</td>
<td>20471.4</td>
</tr>
<tr>
<td>AIC(_c)</td>
<td>22054.3</td>
<td>20525.4</td>
<td>20380.6</td>
<td><strong>20379.8</strong></td>
<td>20385.0</td>
</tr>
<tr>
<td>CAIC</td>
<td>22118.2</td>
<td>20599.9</td>
<td><strong>20465.7</strong></td>
<td>20475.6</td>
<td>20491.4</td>
</tr>
<tr>
<td>HTAIC</td>
<td>22042.6</td>
<td>20511.7</td>
<td>20365.0</td>
<td><strong>20362.3</strong></td>
<td>20365.6</td>
</tr>
<tr>
<td>AIC(_c^*)</td>
<td>22018.5</td>
<td>20483.6</td>
<td>20332.9</td>
<td>20326.2</td>
<td>20325.5</td>
</tr>
<tr>
<td>BIC(^*)</td>
<td>22068.1</td>
<td>20541.4</td>
<td><strong>20398.9</strong></td>
<td>20400.4</td>
<td>20407.8</td>
</tr>
<tr>
<td>CAIC(^*)</td>
<td>22080.1</td>
<td>20555.4</td>
<td><strong>20414.9</strong></td>
<td>20418.4</td>
<td>20427.8</td>
</tr>
</tbody>
</table>
Intuitively, the explanation is that with \( k = 3 \) classes the item estimates by MML are already very close to those obtained with CML:

|      | CML       | MML  
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( k = 1 )</td>
<td>( k = 2 )</td>
<td>( k = 3 )</td>
<td>( k = 4 )</td>
<td>( k = 5 )</td>
<td></td>
</tr>
<tr>
<td>Item 1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Item 2</td>
<td>-0.047</td>
<td>-0.038</td>
<td>-0.045</td>
<td>-0.047</td>
<td>-0.047</td>
<td>-0.047</td>
</tr>
<tr>
<td>Item 3</td>
<td>0.691</td>
<td>0.549</td>
<td>0.670</td>
<td>0.689</td>
<td>0.691</td>
<td>0.691</td>
</tr>
<tr>
<td>Item 4</td>
<td>-1.040</td>
<td>-0.855</td>
<td>-0.984</td>
<td>-1.032</td>
<td>-1.037</td>
<td>-1.040</td>
</tr>
<tr>
<td>Item 5</td>
<td>1.521</td>
<td>1.207</td>
<td>1.478</td>
<td>1.518</td>
<td>1.521</td>
<td>1.521</td>
</tr>
<tr>
<td>Item 6</td>
<td>0.013</td>
<td>0.010</td>
<td>0.012</td>
<td>0.013</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>Item 7</td>
<td>0.662</td>
<td>0.527</td>
<td>0.642</td>
<td>0.661</td>
<td>0.662</td>
<td>0.662</td>
</tr>
<tr>
<td>Item 8</td>
<td>1.191</td>
<td>0.945</td>
<td>1.158</td>
<td>1.189</td>
<td>1.191</td>
<td>1.191</td>
</tr>
<tr>
<td>Item 9</td>
<td>0.334</td>
<td>0.267</td>
<td>0.323</td>
<td>0.333</td>
<td>0.334</td>
<td>0.334</td>
</tr>
<tr>
<td>Item 10</td>
<td>0.525</td>
<td>0.418</td>
<td>0.508</td>
<td>0.524</td>
<td>0.525</td>
<td>0.525</td>
</tr>
<tr>
<td>Item 11</td>
<td>2.427</td>
<td>1.945</td>
<td>2.339</td>
<td>2.418</td>
<td>2.427</td>
<td>2.427</td>
</tr>
<tr>
<td>Item 12</td>
<td>2.474</td>
<td>1.984</td>
<td>2.383</td>
<td>2.464</td>
<td>2.474</td>
<td>2.474</td>
</tr>
</tbody>
</table>
Multilevel data (contraceptive use in Bangladesh)

- Data coming from a study in Bangladesh about the *knowledge and use of family planning methods* by ever-married women

- We considered subset of 1934 women nested in 60 administrative districts where the response of interest is a *binary variable* denoting whether the interviewed woman is currently using contraceptions

- **Covariates** (5 covariates varying within cluster):
  - geographical residence area (0= rural, 1=urban)
  - age
  - number of children (no child, a single child, two children, three or more children)
The proposed test chooses *only 1 support point* at 5%, whereas other criteria select 2 support points:

<table>
<thead>
<tr>
<th></th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 3$</th>
<th>$k = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausman $T$</td>
<td>10.160</td>
<td>9.778</td>
<td>5.164</td>
<td>5.163</td>
</tr>
<tr>
<td>Hausman $p$-value</td>
<td><strong>0.071</strong></td>
<td>0.082</td>
<td>0.400</td>
<td>0.396</td>
</tr>
<tr>
<td>AIC</td>
<td>2469.1</td>
<td><strong>2427.2</strong></td>
<td>2430.0</td>
<td>2434.0</td>
</tr>
<tr>
<td>BIC</td>
<td>2481.7</td>
<td><strong>2444.1</strong></td>
<td>2451.1</td>
<td>2459.4</td>
</tr>
<tr>
<td>AIC$_3$</td>
<td>2475.1</td>
<td><strong>2435.2</strong></td>
<td>2440.0</td>
<td>2446.0</td>
</tr>
<tr>
<td>CAIC</td>
<td>2487.7</td>
<td><strong>2452.1</strong></td>
<td>2461.1</td>
<td>2471.4</td>
</tr>
<tr>
<td>HTAIC</td>
<td>2471.2</td>
<td><strong>2430.8</strong></td>
<td>2435.4</td>
<td>2441.8</td>
</tr>
<tr>
<td>AIC$_c$</td>
<td>2458.2</td>
<td><strong>2413.4</strong></td>
<td>2413.6</td>
<td>2415.5</td>
</tr>
<tr>
<td>BIC*</td>
<td>2462.8</td>
<td><strong>2418.9</strong></td>
<td>2419.7</td>
<td>2421.6</td>
</tr>
<tr>
<td>CAIC*</td>
<td>2468.8</td>
<td><strong>2426.9</strong></td>
<td>2429.7</td>
<td>2433.6</td>
</tr>
</tbody>
</table>
Longitudinal data (HRS data)  

- Longitudinal data set about Self-Reported Health Status (SRHS) deriving from the Health and Retirement Study (HRS) about 1308 individuals who were asked to express opinions on their health status at 4 equally spaced time occasions, from 2000 to 2006.

- The response variable (SRHS) is measured on a Likert type scale based on 5 ordered categories (poor, fair, good, very good, and excellent).

- **Covariates** (2 time-varying covariates):
  - gender (0=male, 1 = female)
  - race (0=white, 1=nonwhite)
  - educational level (3 ordered categories)
  - age, age^2
The proposed test *rejects all* $k$ for Model-ord1 (constant shift in the cutpoints), despite most selection criteria tend to choose 5 components:

<table>
<thead>
<tr>
<th></th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 3$</th>
<th>$k = 4$</th>
<th>$k = 5$</th>
<th>$k = 6$</th>
<th>$k = 7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausman $T$</td>
<td>75.483</td>
<td>59.095</td>
<td>29.917</td>
<td>31.736</td>
<td>30.996</td>
<td>28.494</td>
<td>28.558</td>
</tr>
<tr>
<td>Hausman $p$-value</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>AIC</td>
<td>14880</td>
<td>13345</td>
<td>12872</td>
<td>12665</td>
<td>12581</td>
<td>12583</td>
<td>12585</td>
</tr>
<tr>
<td>BIC</td>
<td>14949</td>
<td>13427</td>
<td>12968</td>
<td>12775</td>
<td>12705</td>
<td>12720</td>
<td>12736</td>
</tr>
<tr>
<td>AIC$_3$</td>
<td>14890</td>
<td>13357</td>
<td>12886</td>
<td>12681</td>
<td><strong>12599</strong></td>
<td>12603</td>
<td>12607</td>
</tr>
<tr>
<td>CAIC</td>
<td>14959</td>
<td>13439</td>
<td>12982</td>
<td>12791</td>
<td><strong>12723</strong></td>
<td>12740</td>
<td>12758</td>
</tr>
<tr>
<td>HTAIC</td>
<td>14880</td>
<td>13345</td>
<td>12872</td>
<td>12665</td>
<td><strong>12581</strong></td>
<td>12583</td>
<td>12585</td>
</tr>
<tr>
<td>AIC$_c$</td>
<td>14860</td>
<td>13321</td>
<td>12844</td>
<td>12633</td>
<td>12545</td>
<td><strong>12543</strong></td>
<td><strong>12541</strong></td>
</tr>
<tr>
<td>BIC*</td>
<td>14917</td>
<td>13389</td>
<td>12924</td>
<td>12724</td>
<td><strong>12647</strong></td>
<td>12656</td>
<td>12666</td>
</tr>
<tr>
<td>CAI*</td>
<td>14927</td>
<td>13401</td>
<td>12938</td>
<td>12740</td>
<td><strong>12665</strong></td>
<td>12676</td>
<td>12688</td>
</tr>
</tbody>
</table>

The model with *normal distributed random-effects* is strongly rejected with $T = 32.165$ and $p$-value = 0.000
For Model-ord2 (free cutpoints) the proposed test leads to selecting $k = 7$ (BIC selects $k = 5$); the model is not rejected at a significance level $\sim 5\%$:

<table>
<thead>
<tr>
<th></th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 3$</th>
<th>$k = 4$</th>
<th>$k = 5$</th>
<th>$k = 6$</th>
<th>$k = 7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hausman $T$</td>
<td>75.483</td>
<td>59.455</td>
<td>19.484</td>
<td>22.274</td>
<td>13.766</td>
<td>8.915</td>
<td>6.088</td>
</tr>
<tr>
<td>Hausman p-value</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.012</td>
<td>0.048</td>
</tr>
<tr>
<td>AIC</td>
<td>14880</td>
<td>13355</td>
<td>12853</td>
<td>12637</td>
<td>12498</td>
<td>12486</td>
<td>12458</td>
</tr>
<tr>
<td>BIC</td>
<td>14949</td>
<td>13499</td>
<td>13073</td>
<td>12932</td>
<td>12868</td>
<td>12933</td>
<td>12979</td>
</tr>
<tr>
<td>AIC$_3$</td>
<td>14890</td>
<td>13376</td>
<td>12885</td>
<td>12680</td>
<td>12552</td>
<td>12551</td>
<td>12534</td>
</tr>
<tr>
<td>CAIC</td>
<td>14959</td>
<td>13520</td>
<td>13105</td>
<td>12975</td>
<td>12922</td>
<td>12998</td>
<td>13055</td>
</tr>
<tr>
<td>HTAIC</td>
<td>14880</td>
<td>13355</td>
<td>12853</td>
<td>12637</td>
<td>12498</td>
<td>12488</td>
<td>12460</td>
</tr>
<tr>
<td>AIC$_c$</td>
<td>14860</td>
<td>13313</td>
<td>12789</td>
<td>12551</td>
<td>12390</td>
<td>12358</td>
<td>12307</td>
</tr>
<tr>
<td>BIC*</td>
<td>14917</td>
<td>13432</td>
<td>12971</td>
<td>12795</td>
<td>12697</td>
<td>12726</td>
<td>12738</td>
</tr>
<tr>
<td>CAI*</td>
<td>14927</td>
<td>13453</td>
<td>13003</td>
<td>12838</td>
<td>12751</td>
<td>12791</td>
<td>12814</td>
</tr>
</tbody>
</table>

**Comparison** between the estimates under the different models:

<table>
<thead>
<tr>
<th></th>
<th>CML</th>
<th>MML (initial, $k = 5$)</th>
<th>MML (ext., $k = 7$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>est.</td>
<td>s.e.</td>
<td>est.</td>
</tr>
<tr>
<td>age</td>
<td>-0.2235</td>
<td>0.0410</td>
<td>-0.1859</td>
</tr>
<tr>
<td>age$^2$</td>
<td>0.0013</td>
<td>0.0024</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-0.2042</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0016</td>
</tr>
</tbody>
</table>
Conclusions

The approach is *easy to implement* and may be used to test the correct specification of the random-effects distribution and to select the number of support points.

It provides *reasonable results* on simulated and real data.

With respect to most used selection criteria (e.g., BIC), the method is expected to lead to *more parsimonious models* (when assumptions hold), but it may reject all models (with different values of $k$) of a certain type.

The applicability is *limited to certain models* (based on a canonical link function), whereas for linear and Poisson models we did not obtain interesting results; however, the case of binary/ordinal data is very relevant.

An interesting case to try with may be that of *survival data*.
Conclusions

References


