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

Francesco Bartolucci[1]

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

MBC^2 - Catania, September 2014

[1] 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)

- *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
  - $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ_i})$: binary observations for cluster i
  - $\boldsymbol{x}_i$: column vector of cluster-specific covariates
  - $\boldsymbol{z}_{ij}$: column vector of observation-specific covariates
- *Base-line model*:

$$\log \frac{p(y_{ij}=1|\alpha_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})}{p(y_{ij}=0|\alpha_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})} = \alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{\gamma}, \quad i = 1, \ldots, n,\ j = 1, \ldots, J_i$$

- α_i are *random-effects* that in the standard case have a normal distribution with unknown variance σ^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)

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

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

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

$$\log \frac{p(y_{ij} \geq l | \boldsymbol{\alpha}_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})}{p(y_{ij} < l | \boldsymbol{\alpha}_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})} = \alpha_{il} + \boldsymbol{x}_i' \boldsymbol{\beta} + \boldsymbol{z}_{ij}' \boldsymbol{\gamma}, \quad l = 1, \ldots, L-1,$$

with cutpoints $\alpha_{i1} > \cdots > \alpha_{i,L-1}$ collected in the vectors $\boldsymbol{\alpha}_i$

- The first two models may be interpreted in terms of an *underlying continuous variable* and a suitable observation rule:

$$y_{ij} = G(y_{ij}^*), \quad y_{ij}^* = \alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{\gamma} + \varepsilon_{ij},$$

  with ε_{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 α_i on a vector of cluster-specific covariates $\boldsymbol{w}_i$ (to face *endogeneity*)
- *1st possible extension*: an interaction term is included as

$$\log \frac{p(y_{ij}=1|\alpha_i,\boldsymbol{w}_i,\boldsymbol{x}_i,\boldsymbol{z}_{ij})}{p(y_{ij}=0|\alpha_i,\boldsymbol{w}_i,\boldsymbol{x}_i,\boldsymbol{z}_{ij})} = \boldsymbol{w}_i'\boldsymbol{\alpha}_i + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{\gamma}, \quad i=1,\ldots,n,\ j=1,\ldots,J_i,$$

  with the vectors of random-effects $\boldsymbol{\alpha}_i$ having k support points $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k$ and mass probabilities $\pi_h = p(\boldsymbol{\alpha}_i = \boldsymbol{\xi}_h)$
- *2nd possible extension*: the mass probabilities depend on the covariates by a multinomial logit parameterization:

$$\log \frac{p(\alpha_i=\xi_{h+1}|\boldsymbol{w}_i)}{p(\alpha_i=\xi_1|\boldsymbol{w}_i)} = \phi_h + \boldsymbol{w}_i'\boldsymbol{\psi}_h, \quad h=1,\ldots,k-1,\ i=1,\ldots,n,$$

  or alternative parametrizations when the support points are ordered

Marginal Maximum Likelihood (MML)

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

$$p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{x}_i, \boldsymbol{Z}_i) = \prod_j p(y_{ij}|\alpha_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij})$$

with $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iJ_i})$ being the matrix of covariates varying within cluster

- The *manifest distribution* is

$$p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{Z}_i) = \sum_h \left[\prod_j p(y_{ij}|\xi_h, \boldsymbol{x}_i, \boldsymbol{z}_{ij}) \right] \pi_h$$

- The *marginal log-likelihood function* is

$$\ell_M(\boldsymbol{\theta}) = \sum_i \log p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{Z}_i) = \sum_i \log \sum_h \left[\prod_j p(y_{ij}|\xi_h, \boldsymbol{x}_i, \boldsymbol{z}_{ij}) \right] \pi_h$$

with $\boldsymbol{\theta}$ denoting the overall vector of parameters

- *Maximization* of $\ell_M(\boldsymbol{\theta})$ 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

$$\ell_M^*(\boldsymbol{\theta}) = \sum_i a_{hi} \left[\log \pi_h + \sum_j \log p(y_{ij}|\xi_h, \boldsymbol{x}_i, \boldsymbol{z}_{ij}) \right],$$

  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 $\hat{a}_{hi} = p(\alpha_i = \xi_h|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{Z}_i)$
  - **M-step**: maximize the function $\ell_M^*(\boldsymbol{\theta})$ with each a_{hi} substituted by $\hat{a}_{hi}$

- The *asymptotic variance-covariance matrix* of the MML estimator $\hat{\boldsymbol{\theta}}_M$ may be estimated by the sandwich formula

$$
\begin{aligned}
\hat{V}_M(\hat{\boldsymbol{\theta}}_M) &= \boldsymbol{H}_M(\hat{\boldsymbol{\theta}}_M)^{-1}\, \boldsymbol{V}_M(\hat{\boldsymbol{\theta}}_M)\boldsymbol{H}_M(\hat{\boldsymbol{\theta}}_M)^{-1} \\
\boldsymbol{u}_{M,i}(\boldsymbol{\theta}) &= \frac{\partial \log p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{x}_i, \boldsymbol{Z}_i)}{\partial \boldsymbol{\theta}} \\
\boldsymbol{H}_M(\boldsymbol{\theta}) &= \sum_i \frac{\partial^2 \log p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{Z}_i)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \\
\boldsymbol{V}_M(\boldsymbol{\theta}) &= \sum_i \boldsymbol{u}_{M,i}(\boldsymbol{\theta})[\boldsymbol{u}_{M,i}(\boldsymbol{\theta})]'
\end{aligned}
$$

- 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 γ for the covariates in $\boldsymbol{Z}_i$ under mild assumptions (mainly time-constant individual effects)
- For binary data, the *conditional log-likelihood function* has expression

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

with

$$p(\boldsymbol{y}_i | \boldsymbol{Z}_i, y_{i+}) = \frac{\exp\left(\sum_j y_{ij}\, \boldsymbol{z}_{ij}'\gamma\right)}{\sum_{\boldsymbol{s} \in \mathcal{S}_{J_i}(y_{i+})} \exp\left(\sum_j s_j\, \boldsymbol{z}_{ij}'\gamma\right)},$$

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

- $p(\boldsymbol{y}_i | \boldsymbol{Z}_i, y_{i+})$ *does not depend* anymore on α_i and $\boldsymbol{x}_i$ (and possibly $\boldsymbol{w}_i$)

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

$$\boldsymbol{u}_C(\boldsymbol{\gamma}) = \sum_i \boldsymbol{u}_{C,i}(\boldsymbol{\gamma}), \quad \boldsymbol{u}_{C,i}(\boldsymbol{\gamma}) = \frac{\partial \log p(\boldsymbol{y}_i | y_{i+}, \boldsymbol{Z}_i)}{\partial \boldsymbol{\gamma}}$$

and Hessian matrix

$$\boldsymbol{H}_C(\boldsymbol{\gamma}) = \sum_i \frac{\partial^2 \log p(\boldsymbol{y}_i | y_{i+}, \boldsymbol{Z}_i)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'}$$

- The *asymptotic variance-covariance matrix* may be obtained as

$$\begin{aligned} \hat{V}_C(\hat{\boldsymbol{\gamma}}_C) &= \boldsymbol{H}_C(\hat{\boldsymbol{\gamma}}_C)^{-1} \boldsymbol{V}_C(\hat{\boldsymbol{\gamma}}_C) \boldsymbol{H}_C(\hat{\boldsymbol{\gamma}}_C)^{-1} \\ \boldsymbol{V}_C(\boldsymbol{\gamma}) &= \sum_i \boldsymbol{u}_{C,i}(\boldsymbol{\gamma}) [\boldsymbol{u}_{C,i}(\boldsymbol{\gamma})]' \end{aligned}$$

- 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 $\boldsymbol{y}_i^{(l)} = (y_{i1}^{(l)}, \ldots, y_{iJ_i}^{(l)})$

- The corresponding *pseudo log-likelihood* function is

$$\ell_C(\gamma) = \sum_i \sum_l \log p(\boldsymbol{y}_i^{(l)} | y_{i+}^{(l)}, \boldsymbol{Z}_i), \quad y_{i+}^{(l)} = \sum_{j=1}^{J} y_{ij}^{(l)},$$

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 γ as in Tchetgen and Coull (2006) and Bartolucci et al. (2014)

- The test exploits the *asymptotic distribution*

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

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

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

with $\boldsymbol{E}$ defined so that $\hat{\gamma}_M = \boldsymbol{E}\, \hat{\boldsymbol{\theta}}_M$

- The *variance-covariates matrix* of $(\hat{\theta}_M, \hat{\gamma}_C)$ is obtained as

$$\hat{V}(\hat{\theta}_M, \hat{\gamma}_C) = \begin{pmatrix} \boldsymbol{H}_M(\hat{\theta}_M) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_C(\hat{\gamma}_C) \end{pmatrix}^{-1} \boldsymbol{S}\left(\hat{\theta}_M, \hat{\gamma}_C\right) \begin{pmatrix} \boldsymbol{H}_M(\hat{\theta}_M) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_C(\hat{\gamma}_C) \end{pmatrix}^{-1}$$

$$\boldsymbol{S}\left(\hat{\theta}_M, \hat{\gamma}_C\right) = \sum_i \begin{pmatrix} \boldsymbol{u}_{M,i}(\hat{\theta}_M) \\ \boldsymbol{s}_{C,i}(\hat{\gamma}_C) \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_{M,i}(\hat{\theta}_M)' & \boldsymbol{s}_{C,i}(\hat{\gamma}_C)' \end{pmatrix}$$

- The *test statistic* is

$$T = n(\hat{\gamma}_M - \hat{\gamma}_C)' \widehat{\boldsymbol{W}}^{-1} (\hat{\gamma}_M - \hat{\gamma}_C)$$

that has asymptotic null distribution of χ^2_g, with g being the dimension of γ (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

$$\log \frac{p(y_{ij}=1|\alpha_i, x_i, \boldsymbol{z}_i)}{p(y_{ij}=0|\alpha_i, x_i, \boldsymbol{z}_i)} = \alpha_i + x_i\beta + z_{ij}\gamma$$

with $\beta = \gamma = 1$, α_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:

	$J = 5$			$J = 10$		
n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
500	0.103	0.057	0.018	0.101	0.049	0.012
1000	0.097	0.060	0.018	0.099	0.052	0.011

- 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 γ_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:

$$\text{AIC} = -2\,\ell_M(\hat{\boldsymbol{\theta}}_M) + 2\#\text{par}$$

$$\text{BIC} = -2\,\ell_M(\hat{\boldsymbol{\theta}}_M) + \#\text{par}\log(n)$$

$$\text{AIC}_3 = -2\,\ell_M(\hat{\boldsymbol{\theta}}_M) + 3\#\text{par}$$

$$\text{CAIC} = -2\,\ell_M(\hat{\boldsymbol{\theta}}_M) + \#\text{par}(\log(n)+1)$$

$$\text{HT-AIC} = -2\,\ell_M(\hat{\boldsymbol{\theta}}_M) + 2\#\text{par} + \frac{2(\#\text{par}+1)(\#\text{par}+2)}{n-\#\text{par}-2}$$

$$\text{AIC}_c = -2\,\ell_M(\hat{\boldsymbol{\theta}}_M) + 2\frac{\#\text{par}(\#\text{par}-1)}{n-\#\text{par}-1}$$

$$\text{BIC}^* = -2\,\ell_M(\hat{\boldsymbol{\theta}}_M) + \#\text{par}\log\frac{n+2}{24}$$

$$\text{CAIC}^* = -2\,\ell_M(\hat{\boldsymbol{\theta}}_M) + \#\text{par}\left(\log\frac{n+2}{24}+1\right)$$

- 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:

k	Haus (10%)	Haus (5%)	Haus (1%)	AIC	BIC	AIC_3	CAIC	HT-AIC	AIC_c	BIC*	CAIC*
1	0	0	2	0	0	0	0	0	0	0	0
2	**415**	**597**	**841**	5	152	17	223	5	0	17	45
3	398	297	122	**627**	**815**	**764**	**757**	**635**	124	**774**	**831**
4	59	40	19	355	33	216	20	347	**550**	206	124
5	64	35	11	13	0	3	0	13	249	3	0
6	40	20	5	0	0	0	0	0	77	0	0

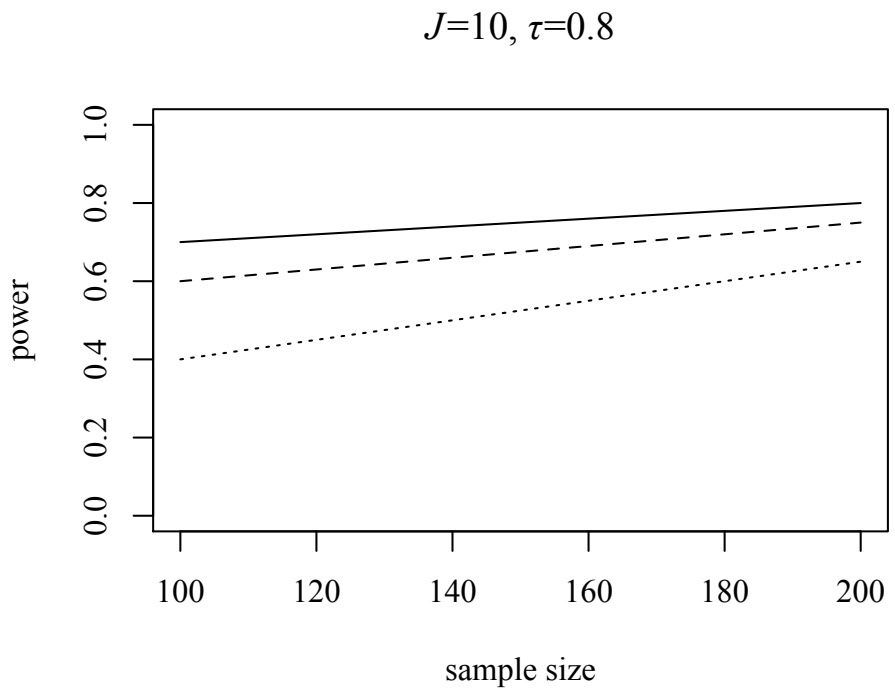
- The *same tendency* is confirmed in many other cases and even when the true distribution is discrete

- 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 $\bar{z}_i = (1/J_i) \sum_j z_{ij}$, $\eta_i \sim N(0, 1)$, 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)



Applications

- 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:

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Hausman T	414.850	90.071	6.721	2.895	1.639
Hausman p-value	0.000	0.000	**0.821**	0.992	0.999
AIC	22042.3	20511.4	20364.6	**20361.8**	20365.0
BIC	22106.2	20585.9	**20449.7**	20457.6	20471.4
AIC_3	22054.3	20525.4	20380.6	**20379.8**	20385.0
CAIC	22118.2	20599.9	**20465.7**	20475.6	20491.4
HTAIC	22042.6	20511.7	20365.0	**20362.3**	20365.6
AIC_c	22018.5	20483.6	20332.9	20326.2	20325.5
BIC*	22068.1	20541.4	**20398.9**	20400.4	20407.8
CAIC*	22080.1	20555.4	**20414.9**	20418.4	20427.8

- Intuitively, the explanation is that with $k = 3$ classes the item estimates by MML are already *very close* to those obtained with CML:

		MML				
	CML	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Item 1	**0.000**	0.000	0.000	**0.000**	0.000	0.000
Item 2	**-0.047**	-0.038	-0.045	**-0.047**	-0.047	-0.047
Item 3	**0.691**	0.549	0.670	**0.689**	0.691	0.691
Item 4	**-1.040**	-0.855	-0.984	**-1.032**	-1.037	-1.040
Item 5	**1.521**	1.207	1.478	**1.518**	1.521	1.521
Item 6	**0.013**	0.010	0.012	**0.013**	0.013	0.013
Item 7	**0.662**	0.527	0.642	**0.661**	0.662	0.662
Item 8	**1.191**	0.945	1.158	**1.189**	1.191	1.191
Item 9	**0.334**	0.267	0.323	**0.333**	0.334	0.334
Item 10	**0.525**	0.418	0.508	**0.524**	0.525	0.525
Item 11	**2.427**	1.945	2.339	**2.418**	2.427	2.427
Item 12	**2.474**	1.984	2.383	**2.464**	2.474	2.474

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:

	$k=1$	$k=2$	$k=3$	$k=4$
Hausman T	10.160	9.778	5.164	5.163
Hausman p-value	**0.071**	0.082	0.400	0.396
AIC	2469.1	**2427.2**	2430.0	2434.0
BIC	2481.7	**2444.1**	2451.1	2459.4
AIC_3	2475.1	**2435.2**	2440.0	2446.0
CAIC	2487.7	**2452.1**	2461.1	2471.4
HTAIC	2471.2	**2430.8**	2435.4	2441.8
AIC_c	2458.2	**2413.4**	2413.6	2415.5
BIC*	2462.8	**2418.9**	2419.7	2421.6
CAIC*	2468.8	**2426.9**	2429.7	2433.6

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:

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
Hausman T	75.483	59.095	29.917	31.736	30.996	28.494	28.558
Hausman p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AIC	14880	13345	12872	12665	**12581**	12583	12585
BIC	14949	13427	12968	12775	**12705**	12720	12736
AIC_3	14890	13357	12886	12681	**12599**	12603	12607
CAIC	14959	13439	12982	12791	**12723**	12740	12758
HTAIC	14880	13345	12872	12665	**12581**	12583	12585
AIC_c	14860	13321	12844	12633	12545	12543	**12541**
BIC*	14917	13389	12924	12724	**12647**	12656	12666
CAI*	14927	13401	12938	12740	**12665**	12676	12688

- 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 $\sim 5\%$:

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
Hausman T	75.483	59.455	19.484	22.274	13.766	8.915	**6.088**
Hausman p-value	0.000	0.000	0.000	0.000	0.001	0.012	**0.048**
AIC	14880	13355	12853	12637	12498	12486	**12458**
BIC	14949	13499	13073	12932	**12868**	12933	12979
AIC_3	14890	13376	12885	12680	12552	12551	**12534**
CAIC	14959	13520	13105	12975	**12922**	12998	13055
HTAIC	14880	13355	12853	12637	12498	12488	**12460**
AIC_c	14860	13313	12789	12551	12390	12358	**12307**
BIC*	14917	13432	12971	12795	12697	**12726**	12738
CAI*	14927	13453	13003	12838	**12751**	12791	12814

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

	CML		MML (initial, $k=5$)		MML (ext., $k=7$)	
	est.	s.e.	est.	s.e.	est.	s.e.
age	-0.2235	0.0410	-0.1859	0.0320	-0.2042	0.0381
age^2	0.0013	0.0024	0.0013	0.0018	0.0016	0.0023

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*

References

Bartolucci, F., Belotti, F., and Peracchi, F. (2014). Testing for time-invariant unobserved heterogeneity in generalized linear models for panel data. *Journal of Econometrics,* in press.

Hausman, J. (1978). Specification tests in econometrics. *Econometrica*, 46:1251–1271.

Tchetgen, E. J. and Coull, B. A. (2006). A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika*, 93(4):1003–1010.

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, 47:639–653.

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and computing*, 6(3):251–262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalised linear models. *Biometrics*, 55:218–234.

Alonso, A., Litière, S., and Molenberghs, G. (2008). A family of tests to detect misspecifications in the random-effects structure of generalized linear mixed models. *Computational statistics & data analysis*, 52(9):4474–4486.

Alonso, A. A., Litière, S., and Molenberghs, G. (2010). Testing for misspecification in generalized linear mixed models. *Biostatistics*, 11(4):771–786.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of Royal Statistical Society, Series B*, 32:283–301.

Baetschmann, G., Staub, K. E., and Winkelmann, R. (2011). Consistent estimation of the fixed effects ordered logit model. Technical Report 5443, IZA.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, 72:141–157.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47:225–238.

Deb, P. (2001). A discrete random effects probit model with application to the demand for preventive care. *Health Economics*, 10:371–383.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55(3):688–698.

Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric model for duration data. *Econometrica*, 52:271–320.

Heiss, F. (2008). Sequential numerical integration in nonlinear state space models for microeconometric panel data. *Journal of Applied Econometrics*, 23:373–389.

Huq, M. N. and Cleland, J. (1990). Bangladesh fertility survey, 1989. Technical report, Main Report.

Juster, F. T. and Suzman, R. (1995). An overview of the health and retirement study. *The Journal of Human Resources*, 30:S7–S56.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixture distribution. *Journal of the American Statistical Association*, 73:805–811.

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, pages 624–642.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, Boston.

Lindsay, B., Clogg, C. C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86:96–107.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Annals of Statistics*, 11:86–94.

Mazharul Islam, M. and Mahmud, M. (1995). Contraceptions among adolescents in Bangladesh. *Asia Pacific Population Journal*, 10:21–38.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. Wiley.

Neuhaus, J. M., Hauck, W. W., and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, 79(4):755–762.

Pan, Z. and Lin, D. (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, 61(4):1000–1009.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. (2003). Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, 3:215–232.

Ritz, C. (2004). Goodness-of-fit tests for mixed models. *Scandinavian journal of statistics*, 31(3):443–458.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, pages 961–971.

Verbeke, G. and Molenberghs, G. (2013). The gradient function as an exploratory goodness-of-fit assessment of the random-effects distribution in mixed models. *Biostatistics*, 14(3):477–490.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, 33:213–239.

Vijverberg, W. P. (2011). Testing for IIA with the Hausman-McFadden Test. IZA Discussion Papers 5826, Institute for the Study of Labor (IZA).

Waagepetersen, R. (2006). A simulation-based goodness-of-fit test for random effects in generalized linear mixed models. *Scandinavian journal of statistics*, 33(4):721–731.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American statistical association*, 86(413):79–86.