Selection of the number of latent classes

Silvia Bacci

silvia.bacci@unipg.it
Dipartimento di Economia - Università degli Studi di Perugia (IT)

Outline

- Problem: a crucial point with LC models is represented by the selection of the number of latent classes
- In the following two different studies are illustrated
  - Study 1: comparison among several information criteria in the frame of multivariate LM models
  - Study 2: proposal of a new Hausman-type test in the frame of GLMMs with discrete random effects

Model selection criteria

- Usually, the selection of the number of components in LC models relies on the information criteria, consisting in penalized versions of the maximum log-likelihood, where the penalization term accounts for the number of parameters
- Information criteria represent a compromise between goodness-of-fit and model parsimony
- The optimal number of components is that corresponding to the minimum value of the corresponding index
- In practice, we fit a given LC model for increasing values of k until the index does not start to increase and we select the previous k as the optimal number of components

- Akaike's Information Criterion (AIC - Akaike, 1973)

$$\mathrm{AIC} = -2\,\hat{\ell} + 2 \cdot \#\mathrm{par}$$

- Bayesian Information Criterion (BIC - Schwarz, 1978)

$$\mathrm{BIC} = -2\,\hat{\ell} + \#\mathrm{par} \cdot \log(n)$$

- Consistent AIC (Bozdogan, 1987)

$$\mathrm{CAIC} = -2\,\hat{\ell} + \#\mathrm{par} \cdot (\log(n) + 1)$$

- AIC_3 (Bozdogan, 1993)

$$\mathrm{AIC}_3 = -2\,\hat{\ell} + 3 \cdot \#\mathrm{par}$$

- HT-AIC (Hurvich and Tsai, 1989)

$$\mathrm{HT-AIC} = -2\,\hat{\ell} + 2\#\mathrm{par} + \frac{2(\#\mathrm{par}+1)(\#\mathrm{par}+2)}{n-\#\mathrm{par}-2}$$

- AIC_c (Hurvich and Tsai, 1993)

$$\mathrm{AIC}_c = -2\,\hat{\ell} + 2\frac{\#\mathrm{par}(\#\mathrm{par}-1)}{n-\#\mathrm{par}-1}$$

- Adjusted BIC (Sclove, 1987)

$$\mathrm{BIC}^* = -2\,\hat{\ell} + \#\mathrm{par}\log\frac{n+2}{24}$$

- Adjusted CAIC (Yang and Yang, 2007)

$$\mathrm{CAIC}^* = -2\,\hat{\ell} + \#\mathrm{par}\left(\log\frac{n+2}{24}+1\right)$$

Classification-based criteria

In the context of multivariate LM models, we propose a comparison with a different type of criteria developed in the context of the classification likelihood approach, based on the relation

$$\ell^*(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \mathrm{EN}$$

where EN is the entropy, which takes explicitly into account the partition of observations in different latent states and it denotes a penalization term which measures the quality of the partition and it is defined as (Hernando et al., 2005)

$$\begin{aligned}\mathrm{EN} &= -\sum_{u_1}\dots\sum_{u_T} f_{u_1,\dots u_T|\boldsymbol{y}} \log(f_{u_1,\dots u_T|\boldsymbol{y}}) = \\ &= -\sum_{u_1}\dots\sum_{u_T} f^{(1)}_{u_1|\boldsymbol{y}} \cdot f^{(2)}_{u_2|u_1,\boldsymbol{y}} \cdot \dots \cdot f^{(t)}_{u_t|u_{t-1},\boldsymbol{y}} \cdot \dots \cdot f^{(T)}_{u_T|u_{T-1},\boldsymbol{y}} \cdot \\ &\cdot [\log(f^{(1)}_{u_1|\boldsymbol{y}}) + \log(f^{(2)}_{u_2|u_1,\boldsymbol{y}}) + \dots + \log(f^{(t)}_{u_t|u_{t-1},\boldsymbol{y}}) + \dots + \log(f^{(T)}_{u_T|u_{T-1},\boldsymbol{y}})]\end{aligned}$$

with

$$f_{u|\boldsymbol{y}}^{(t)} = \frac{q_{u,\boldsymbol{y}}^{(t)} \cdot \bar{q}_{u,\boldsymbol{y}}^{(t)}}{p(\boldsymbol{Y} = \boldsymbol{y})}$$

$$f_{u|v,\boldsymbol{y}}^{(t)} = \frac{f_{v,u|\boldsymbol{y}}^{(t-1,t)}}{f_{v|\boldsymbol{y}}^{(t-1)}} = \frac{q_{v,\boldsymbol{y}}^{(t-1)} \pi_{u|v}^{(t)} \phi_{\boldsymbol{y}^{(t)}|u} \bar{q}_{u,\boldsymbol{y}}^{(t)}}{p(\boldsymbol{Y} = \boldsymbol{y})} \cdot \frac{p(\boldsymbol{Y} = \boldsymbol{y})}{q_{v,\boldsymbol{y}}^{(t-1)} \bar{q}_{v,\boldsymbol{y}}^{(t-1)}} =$$

$$= \pi_{u|v}^{(t)} \phi_{\boldsymbol{y}^{(t)}|u} \cdot \frac{\bar{q}_{u,\boldsymbol{y}}^{(t)}}{\bar{q}_{v,\boldsymbol{y}}^{(t-1)}}$$

We may also formulate an approximation for EN, under the assumption that $u^{(t)}$ are independent given $\boldsymbol{Y}$:

- $\mathrm{EN}_1 = -\sum_{u_1} \dots \sum_{u_T} f_{u|\mathbf{y}}^{(t)} \log(f_{u|\mathbf{y}}^{(t)})$
- or a possible variant of EN_1 given by $\mathrm{EN}_2 = -\sum_{u_1} \dots \sum_{u_T} f_{u|\mathbf{y}}^{(t)} \log(f_{u|\mathbf{y}}^{(t)})/T$
- Example: T=3

$$
\begin{aligned}
\mathrm{EN} &= -\sum_u \sum_v \sum_z f_{u,v,z|\mathbf{y}} \log(f_{u,v,z|\mathbf{y}}) = \\
&= f_{z|v,\mathbf{y}}^{(3|2)} \cdot f_{v|u,\mathbf{y}}^{(2|1)} \cdot f_{u|\mathbf{y}}^{(1)} \cdot \\
&\cdot [\log(f_{z|v,\mathbf{y}}^{(3|2)}) + \log(f_{v|u,\mathbf{y}}^{(2|1)}) + \log(f_{u|\mathbf{y}}^{(1)})] \\
\mathrm{EN}_1 &= -[f_{u|\mathbf{y}}^{(1)} \cdot \log(f_{u|\mathbf{y}}^{(1)}) + f_{v|\mathbf{y}}^{(2)} \cdot \log(f_{v|\mathbf{y}}^{(2)}) + f_{z|\mathbf{y}}^{(3)} \cdot \log(f_{z|\mathbf{y}}^{(3)})] \\
\mathrm{EN}_2 &= \frac{1}{3}\mathrm{EN}_1
\end{aligned}
$$

Some classification-based criteria are (McLachlan and Peel, Chap. 6)

- Classification Likelihood information Criterion (CLC)

$$\mathrm{CLC} = -2\ell(\boldsymbol{\theta}) + 2 \cdot \mathrm{EN}$$

- Approximated Integrated Classification Likelihood criterion (ICL-BIC)

$$\mathrm{ICL} - \mathrm{BIC} = \mathrm{BIC} + 2 \cdot \mathrm{EN}$$

- Normalized Entropy Criterion (NEC)

$$\mathrm{NEC} = \frac{\mathrm{EN}}{\ell(\boldsymbol{\theta}) - \ell_1(\boldsymbol{\theta})} \quad k \geq 2$$

  where $\ell_1(\boldsymbol{\theta})$ is the maximum log-likelihood in case of $k = 1$, and $\mathrm{NEC} = 1$ if $k = 1$

- Approximated NECs:

$$\mathrm{NEC}_1 = \frac{\mathrm{EN}_1}{\ell(\boldsymbol{\theta}) - \ell_1(\boldsymbol{\theta})} \quad k \geq 2$$

$$\mathrm{NEC}_2 = \frac{\mathrm{EN}_2}{\ell(\boldsymbol{\theta}) - \ell_1(\boldsymbol{\theta})} \quad k \geq 2$$

Monte Carlo simulation study

- We compare
  - AIC, CAIC, AIC3, BIC
  - CLC, ICL-BIC, NEC, NEC_1, NEC_2
- 100 samples with a given size n and coming from a multivariate LM model, characterized by r binary $(y = 0, 1)$ response variables observed in T time occasions, k latent states, and given values of initial probabilities π_u, transition probabilities $\pi_{u|v}^{(t)}$, conditional response probabilities $\phi_{jy|u}^{(t)}$
- $n = 250, 500, 1000$
- $r = 1, 3, 5$
- $T = 5, 10$
- $k = 2, 3$
- all analyses are implemented in R software

Scenery 1

- $n = 250$, $T = 5$, $k = 2$
- $\phi^{(t)}_{j0|u=1} = 0.8 = \phi^{(t)}_{j1|u=2}, \quad \phi^{(t)}_{j0|u=2} = 0.2 = \phi^{(t)}_{j1|u=1}$
- $\pi_1 = 0.5 = \pi_2$
- $\pi^{(t)}_{1|1} = 0.7 = \pi^{(t)}_{2|2}, \quad \pi^{(t)}_{1|2} = 0.3 = \pi^{(t)}_{2|1}$ (time homogenous assumption)
- $r = 1, 3, 5$

Scenery 1: Relative frequencies of k chosen on the basis of several criteria

k	BIC	AIC	AIC$_3$	CAIC	NEC	NEC$_1$	NEC$_2$	CLC	ICL-BIC
$r = 1$									
1	**0.52**	0.00	0.10	**0.63**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**
2	**0.48**	0.98	0.90	0.37	0.00	0.00	0.01	0.00	0.00
3	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	**0.88**	**0.92**	0.00	**0.88**	**0.95**
2	1.00	0.83	0.98	1.00	0.10	0.07	0.96	0.10	0.04
3	0.00	0.16	0.02	0.00	0.01	0.01	0.04	0.01	0.01
4	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	1.00	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Scenery 2

- $n = 250$, $T = 5$, $k = 2$
- $\phi^{(t)}_{j0|u=1} = 0.7 = \phi^{(t)}_{j1|u=2}, \quad \phi^{(t)}_{j0|u=2} = 0.3 = \phi^{(t)}_{j1|u=1}$
- $\pi_1 = 0.5 = \pi_2$
- $\pi^{(t)}_{1|1} = 0.9 = \pi^{(t)}_{2|2}, \quad \pi^{(t)}_{1|2} = 0.1 = \pi^{(t)}_{2|1}$ (time homogenous assumption)
- $r = 1, 3, 5$

Scenery 2: Relative frequencies of k chosen on the basis of several criteria

k	BIC	AIC	AIC_3	CAIC	NEC	NEC_1	NEC_2	CLC	ICL-BIC
$r = 1$									
1	0.35	0.01	0.02	**0.53**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
2	0.65	0.98	0.97	**0.47**	0.00	0.00	0.00	0.00	0.00
3	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	**1.00**	**1.00**	0.09	**1.00**	**1.00**
2	1.00	0.92	0.995	1.00	0.00	0.00	0.855	0.00	0.00
3	0.00	0.07	0.005	0.00	0.00	0.00	0.015	0.00	0.00
4	0.00	0.01	0.00	0.00	0.00	0.00	0.015	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.025	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.285	**0.77**	0.00	0.285	**0.55**
2	1.00	0.78	0.995	1.00	0.59	0.22	0.98	0.59	**0.445**
3	0.00	0.205	0.005	0.00	0.03	0.005	0.015	0.035	0.005
4	0.00	0.01	0.00	0.00	0.07	0.005	0.005	0.070	0.00
5	0.00	0.005	0.00	0.00	0.025	0.00	0.000	0.025	0.00

Scenery 3

- $n = 500$, $T = 5$, $k = 3$
- $\phi^{(t)}_{j0|u=1} = 0.9 = \phi^{(t)}_{j1|u=2}$, $\quad \phi^{(t)}_{j0|u=2} = 0.1 = \phi^{(t)}_{j1|u=1}$, $\quad \phi^{(t)}_{j0|u=3} = 0.4$, $\phi^{(t)}_{j1|u=3} = 0.6$
- $\pi_1 = \pi_2 = \pi_3 = 0.33$
- $\pi^{(t)}_{1|1} = \pi^{(t)}_{2|2} = \pi^{(t)}_{3|3} = 0.80$, $\quad \pi^{(t)}_{2|1} = 0.15 = \pi^{(t)}_{2|3}$, $\quad \pi^{(t)}_{3|1} = 0.05 = \pi^{(t)}_{1|3}$, $\pi^{(t)}_{1|2} = 0.10 = \pi^{(t)}_{3|2}$ (time homogenous assumption)
- $r = 1, 3, 5$

Scenery 3: Relative frequencies of k chosen on the basis of several criteria

k	BIC	AIC	AIC_3	CAIC	NEC	NEC_1	NEC_2	CLC	ICL-BIC
$r = 1$									
1	0.00	0.00	0.00	0.00	**1.00**	**1.00**	**0.92**	**1.00**	**1.00**
2	**1.00**	**0.98**	**0.99**	**1.00**	0.00	0.00	0.07	0.00	0.00
3	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.03	0.00	0.00	0.10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
3	0.97	0.81	1.00	0.90	0.00	0.00	0.00	0.00	0.00
4	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
3	1.00	0.78	0.99	1.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.20	0.01	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Conclusions

- We compared several criteria for the selection of the number of latent states in the LM models
  - AIC, BIC and their variants present a better general behavior with respect to the classification-based criteria
  - classification-based criteria tend to underestimate the true number of latent states, mainly for the univariate case
  - the behavior of classification-based criteria improves by increasing the number of observed response variables
  - by increasing the number k of latent states the performance of all considered criteria gets worse

Main references

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, Second International symposium of information theory, pages 267-281, Budapest. Akademiai Kiado.
- Bacci S., Pandolfi S., Pennoni F. (2014). A comparison of some criteria for states selec- tion in the latent Markov model for longitudinal data, Advances in Data Analysis and Classification, 8, 125-145.
- Hernando, D., Crespi, V., and Cybenko, G. (2005). Efficient computation of the hidden Markov model entropy for a given observation sequence. IEEE Transactions on Information Theory, 51(7), 2681-2685.
- McLachlan, G. and Peel, D. (2000). Finite Mixture Models. Chap. 6. Wiley.
- Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Motivation

- *Generalized Linear Mixed Models* (GLMMs) represent a very useful instrument for the analysis of clustered data
- *Applications*:
    - Item Response Theory (IRT)
    - multilevel data (individuals 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 and the extension of this model to deal with *ordinal data*
- 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 (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

- Testing the hypothesis that the mixing distribution is normal has attracted considerable attention in the recent statistical literature
- Among the available approaches we recall the Hausman's test (Hausman, 1978)
- 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, to test the hypothesis that the mixing distribution of random-effects is discrete (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)

Basic notation

- For other details we rely on Lecture 1, section "LC models with covariates" and "Example 3"
- n: number of clusters (individuals in the case of longitudinal studies or IRT)
- J_i: number of observations for cluster i
- $\boldsymbol{y}_{ij}$: binary ($y_{ij} = 0, 1$) or ordered ($y_{ij} = 0, \dots, L-1$) response of unit j belonging to cluster i
- $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iJ_i})$: vector of binary or ordered responses for cluster i
- $\boldsymbol{x}_i$: column vector of cluster-specific covariates
- z_{ij}: column vector of unit-specific covariates

Base-line models

- In case of binary responses, the following random intercept logit model follows

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

  - $\boldsymbol{\beta}$ is the vector of regression parameters for the cluster-specific covariates
  - $\boldsymbol{\gamma}$ is the vector of regression parameters for the unit-specific covariates
  - α_i are *cluster-specific 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 $\boldsymbol{y}_i$ given the random-effects α_i and the covariates $\boldsymbol{x}_i$ and $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{iJ_i})$

- With ordinal response variables, the model may be formulated on the basis of global logits as (Model-ord1)

$$\log \frac{p(y_{ij} \geq l|\alpha_i, \boldsymbol{x}_i, z_{ij})}{p(y_{ij} < l|\alpha_i, \boldsymbol{x}_i, z_{ij})} = \alpha_i + \delta_y + \boldsymbol{x}_i'\boldsymbol{\beta} + z_{ij}'\gamma, \quad l = 1, \ldots, L-1, \qquad (2)$$

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

- An alternative formulation is based on cluster-specific cutpoints (Model-ord2):

$$\log \frac{p(y_{ij} \geq l|\boldsymbol{\alpha}_i, \boldsymbol{x}_i, z_{ij})}{p(y_{ij} < l|\boldsymbol{\alpha}_i, \boldsymbol{x}_i, z_{ij})} = \alpha_{il} + \boldsymbol{x}_i'\boldsymbol{\beta} + z_{ij}'\gamma, \quad l = 1, \ldots, L-1, \qquad (3)$$

with $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{i,L-1})$ having multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ or a discrete distribution with support points $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k$ and corresponding probabilities $\pi_h = p(\boldsymbol{\alpha}_i = \boldsymbol{\xi}_h)$, $h = 1, \ldots, k$.

Extended models

- All the above models may be extended to deal with the dependence of the random effects on one or more cluster-specific covariates $\boldsymbol{w}_i$ (which may be a subset of $\boldsymbol{x}_i$), which may be seen as a form of endogeneity
  - First extension: an interaction term is included as (binary case)

$$\log \frac{p(y_{ij} = 1|\alpha_i, \boldsymbol{w}_i, \boldsymbol{x}_i, z_{ij})}{p(y_{ij} = 0|\alpha_i, \boldsymbol{w}_i, \boldsymbol{x}_i, z_{ij})} = \boldsymbol{w}_i'\boldsymbol{\alpha}_i + \boldsymbol{x}_i'\boldsymbol{\beta} + z_{ij}'\gamma, \quad i = 1, \ldots, n, j = 1, \ldots, J_i, \quad (4)$$

  - Second extension: the mass probabilities depend on the covariates by a multinomial logit parameterization (binary case):

$$\log \frac{p(\alpha_i = \xi_{h+1}|\boldsymbol{w}_i)}{p(\alpha_i = \xi_1|\boldsymbol{w}_i)} = \phi_h + \boldsymbol{w}_i'\boldsymbol{\psi}_h, \quad h = 1, \ldots, k-1, \quad (5)$$

    or alternative parameterizations when the support points are ordered

Discrete Marginal Maximum Likelihood (MML)

- The assumption of local independence implies

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

- The *manifest distribution* of $\boldsymbol{y}_i$ given the covariates is obtained by marginalizing $p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{x}_i, \boldsymbol{Z}_i)$ with respect to α_i

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

- 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 free parameters

- *Maximization* of $\ell_M(\boldsymbol{\theta})$ may be efficiently performed by an Expectation Maximization (EM) algorithm

- 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 (White, 1982)

$$\widehat{\boldsymbol{V}}_M(\hat{\boldsymbol{\theta}}_M) = \boldsymbol{H}_M(\hat{\boldsymbol{\theta}}_M)^{-1}\,\boldsymbol{S}_M(\hat{\boldsymbol{\theta}}_M)\boldsymbol{H}_M(\hat{\boldsymbol{\theta}}_M)^{-1}, \tag{6}$$

with

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

- 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

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

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

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

and Hessian matrix

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

- The asymptotic variance-covariance matrix may be obtained as

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

- With ordinal variables, CML estimation is based on all the possible dichotomizations of the response variables:

$$y_{ij}^{(l)} = I\{y_{ij} \geq l\}, \quad j = l, \dots, L-1,$$

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

- 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 relies on the traditional Hausman test, which is typically used to test the assumption of normality of the random effects in linear mixed models
- The traditional Hausman test is based on the comparison of two estimators (CML and MML) that under the null hypothesis of correct model specification (H_0) are both consistent, but if the model is misspecified (H_1) only one of them remains consistent (CML)
- Moreover, it is required that one of the two estimators is asymptotically efficient under H_0 (MML), so as to simplify the estimation of the variance-covariance matrix of the difference between them

- In the Hausman-type test here proposed, H_0 corresponds to a GLLM for binary data or for ordinal data, or their extended versions, in which the distribution of the random effects α_i is discrete with k support points
- The method 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 statistic is defined as

$$T_2 = n(\hat{\gamma}_M - \hat{\gamma}_C)' \widehat{\boldsymbol{W}}^{-1} (\hat{\gamma}_M - \hat{\gamma}_C)$$

- T_2 has an asymptotical distribution of type χ^2_c under H_0, where c is number of unit-specific covariates in $\boldsymbol{z}_{ij}$
- Traditional method to estimate the variance-covariance matrix:

$$\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{V}}_C(\hat{\gamma}_C) - \widehat{\boldsymbol{V}}_M(\hat{\gamma}_M)$$

Generalized variance-covariance matrix estimator

- The traditional formula for $\widehat{\boldsymbol{W}}$ presents, in the present context, stability problems with small samples
- To avoid instability problems and to avoid to require that one of the two estimators is efficient , we use a generalized form for the variance-covariance matrix (Bartolucci et al., 2014), so extending the original method of Hausman (1978):

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

with $\boldsymbol{I}$ being the identity matrix of dimension q and $\boldsymbol{E}$ a matrix such that $\hat{\boldsymbol{\gamma}}_M = \boldsymbol{E}\,\hat{\boldsymbol{\theta}}_M$

- The joint variance-covariance matrix of $\hat{\gamma}_C$ and $\hat{\boldsymbol{\theta}}_M$ is obtained by the generalised sandwich formula

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

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

Use of the proposed Hausman-type test

The proposed test may be used:

- to investigate about the correct specification of a discrete GLLM
- to select the number of mixture components (k), when this number is unknown
  - sequential procedure: k is increased until the test does not stop to reject H_0

    The selection criterion based on T_2 is expected to be more parsimonious with respect to available criteria (i.e., AIC, BIC) provided that the assumptions about the dependence between the random effects and the covariates are correctly specified
  - absolute judgement: for a given k, a sufficiently high p-value leads to conclude for the correct specification of the model in the complex

    The other available criteria to select k only perform relative comparisons among differently specified models

Simulation study

- The study is based on the GLLM (1) for binary responses and (2) for ordinal responses
- Two scenarios: longitudinal setting (one cluster-specific covariate and one unit-specific covariate) and IRT setting (Rasch model)
- Several discrete distributions with $k = 3$ for α_i: symmetric, symmetric with shift, and asymmetric
- Two possible misspecifications: the true distribution of α_i is a normal one; presence of endogeneity
- The proposed test for choosing k is compared with some available criteria, such as AIC, BIC, and several variants

Simulation results

- If the number of classes is underspecified, the Hausman test rejection rate considerably increases when the distribution of the random effects is skewed
- If the random effects follow a continuous distribution, the proposed Hausman test chooses a more parsimonious model in comparison to standard model selection criteria
- The parsimony is greater for large values of units J, which usually leads to a clearer interpretation of the results, especially when the aim is data classification or when the interest in on the regression parameters
- In the presence of endogeneity, rejection rates are remarkably high, even in very small samples
- the power of the test increases with the correlation and the number of clusters, while an increasing number of units seems to only slightly affect the rejection rates

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

- A traditional Hausman test for the Rasch model based on the assumption of normality of the distribution of the random effects leads to accept the null hypothesis of correct model specification ($T_2 = 10.230$, $p = 0.510$)
- However, the normality assumption does not allow us to cluster subjects in homogeneous classes in an easy way, differently from the discreteness assumption:

Table : *Naep data, Rasch model with $k = 3$: estimated support points and weights (standard errors in brackets).*

	$h = 1$	$h = 2$	$h = 3$
$\hat{\xi}_h$	-0.647 (0.138)	0.967 (0.131)	2.430 (0.120)
$\hat{\pi}_h$	0.164 (–)	0.457 (0.154)	0.379 (0.251)

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 a 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 cut points) and for Model-ord2 (free cut points), despite most selection criteria tend to choose 5 components
- The model with *normal distributed random-effects* is strongly rejected with $T_2 = 32.158$ and p-value $= 0.000$
- Such results suggest that a possible problem with the data at issue may be due to the presence of endogeneity
- Then, we extend models Model-ord1 and Model-ord2 to account for a possible effect of age and squared age on the mixture components weights, as in (5)

Table : *Model-ord2 with endogeneity of type* (5)

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
T_2	75.483	59.454	19.484	22.274	13.767	9.003	5.994
p	0.000	0.000	0.000	0.000	0.001	0.011	**0.050**
AIC	14879.9	13355.1	12852.8	12636.9	12497.6	12486.4	12457.8
BIC	14948.6	13499.2	13072.5	12932.1	**12868.3**	12932.6	12979.4
AIC$_3$	14889.9	13376.1	12884.8	12679.9	12551.6	12551.4	**12533.8**
CAIC	14958.6	13520.2	13104.5	12975.1	**12922.3**	12997.6	13055.4
HTAIC	14880.0	13355.2	12853.2	12637.5	12498.5	12487.7	12459.5
AIC$_c$	14859.9	13313.2	12789.1	12551.4	12390.4	12357.6	12307.4
BIC*	14916.8	13432.5	12970.8	12795.4	**12696.7**	12726.0	12737.9
CAI*	14926.8	13453.5	13002.8	12838.4	**12750.7**	12791.0	12813.9

- model Model-ord2 (based on assumption (3)) with endogeneity of type (5) is not rejected with $k=7$
- BIC and several other information criteria do not recognize the misspecification of the model and tend again to choose $k=5$ components
- the traditional Hausman test recognizes the misspecification of the model, but does not detect a valid alternative

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, so detecting misspecification problems
- 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

Main 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.