Comparison between conditional and marginal maximum likelihood for a class of item response models

Francesco Bartolucci, University of Perugia (IT)
Silvia Bacci, University of Perugia (IT)
Claudia Pigini, University of Perugia (IT)

ASMOD 2013
Napoli - November, 25-26, 2013

Outline

Motivation and purpose

- In the literature on *latent variable models*, there is a considerable interest in estimation methods that do not require parametric assumptions on the latent distribution
- We focus on an Item Response Theory model for *ordinal responses* which is known as Graded Response Model
- We introduce a *conditional likelihood estimator* which requires no assumptions on the latent distribution and is very simple to implement
- The method also allows us to implement a *Hausman test* for a parametric assumption (e.g., normal distribution) on the latent distribution

Graded Response Model (GRM)

- For a *questionnaire* of r items, let X_j denote the response variable for the j-th item ($j = 1, \dots, r$), which is assumed to have l_j categories, indexed from 0 to $l_j - 1$
- *Assumptions* of the GRM model (Samejima, 1969):
  - *unidimensionality*: the test items contribute to measure a single latent trait Θ corresponding to a type of ability in education
  - *local independence*: the response variables $X_1, \dots, X_r$ are conditionally independent given Θ:

$$p(x_1, \dots, x_r|\theta) = \prod_{j=1}^{r} p(x_j|\theta)$$

  - *monotonicity*: $p(X_j \geq x|\theta)$ is nondecreasing in θ for all j:

$$\log \frac{p(X_j \geq x|\theta)}{p(X_j < x|\theta)} = \gamma_j(\theta - \beta_{jx}), \quad x = 1, \dots, l_j - 1$$

- γ_j identifies the *discriminating power* of item j (typically $\gamma_j > 0$)
- β_{jx} denotes the *difficulty level* for item j and category x, ordered as $\beta_{j1} < \ldots < \beta_{j,l_j-1}$
- We focus on a special case of GRM (1P-GRM) in which *all the items discriminate* in the same way (van der Ark, 2001):

$$\gamma_1 = \cdots = \gamma_r = 1$$

- We also consider a further special case (1P-RS-GRM) based on the *rating scale parametrization* (items have the same number of response categories):

$$\beta_{jx} = \beta_j + \tau_x, \quad j = 1, \ldots, r, \; x = 1, \ldots, l-1,$$

where β_j represents the difficulty of item j and τ_x are cut-points common to all items

Maximum likelihood estimation

- Given a sample of observations x_{ij}, $i = 1, \dots, n$, $j = 1, \dots, r$, different *maximum likelihood estimation methods* may be used
- Under a fixed-effects formulation, the model may be estimated by the *Joint Maximum Likelihood* (JML) method based on:

$$\ell_J(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \log \prod_{j=1}^{r} p(x_{ij}|\theta_i) = \sum_{i=1}^{n} \sum_{j=1}^{r} \log p(x_{ij}|\theta_i)$$

  with the parameter vector $\boldsymbol{\lambda}$ also including the ability parameters θ_i

- The JML method is simple to implement but *it does not ensure consistency* of the parameter estimates and may suffer from instability problems

- Under a random-effects formulation, with the latent trait assumed to have a normal distribution, we can use the *Marginal Maximum Likelihood* (MML) method based on:

$$\ell_M(\boldsymbol{\eta}) = \sum_{i=1}^{n} \log \int \phi(\theta_i; 0, \sigma^2) \prod_{j=1}^{r} p(x_{ij}|\theta_i) d\theta_i$$

  with $\phi(\theta_i; 0, \sigma^2)$ denoting the density function of $N(0, \sigma^2)$ and the parameter vector $\boldsymbol{\eta}$ containing the item parameters and σ^2

- The MML method is *more complex to implement* (requires a quadrature for the integral) and the parameter estimates are consistent under the hypothesis of normality of the latent trait

- In order to reduce the dependence of the parameter estimates on parametric assumptions on the latent distribution, we can use a *semi-parametric method* (MML-LC) based on the assumption that the latent trait has a discrete distribution with k support points (latent classes)

- The MML-LC method is based on the *marginal log-likelihood function*:

$$\ell_{LC}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \sum_{c=1}^{k} \pi_c \prod_{j=1}^{r} p(x_{ij}|\theta_i = \xi_c)$$

  with $\xi_1, \ldots, \xi_k$ being the support points and $\pi_1, \ldots, \pi_k$ the corresponding mass probabilities; these are included in the parameter vector $\boldsymbol{\psi}$ together with the item parameters

- The *EM algorithm* (Dempster et al., 1977) is typically used for the maximization of $\ell_{LC}(\boldsymbol{\psi})$

- A drawback of the method is the greater *numerical complexity* and the need to *choose k properly* (AIC and BIC may be used in this regard)

- Some *instability problems* may arise with large values of k

Conditional maximum likelihood method

- We suggest a *Conditional Maximum Likelihood* (CML) method based on considering all the possible dichotomizations of the response variables (Baetschmann et al., 2011)
- For the case in which the response variables have the *same number l of response categories*:
  1. we consider the $l-1$ dichotomizations indexed by $d = 1, \dots, l-1$
  2. for each dichotomization d we transform the response variables X_j (for every unit) in the binary variables
  $$Y_j^{(d)} = 1\{X_j \geq d\}, \quad j = 1, \dots, r,$$
  with $1\{\cdot\}$ being the indicator function
  3. we maximize the function given by the *sum of the conditional log-likelihood functions* (Anderson, 1973) corresponding to each dichotomization:
  $$\ell_C^*(\boldsymbol{\beta}) = \sum_{d=1}^{l-1} \log p(y_{i1}^{(d)}, \dots, y_{ir}^{(d)} | y_{i+}^{(d)}), \quad y_{i+}^{(d)} = \sum_{j=1}^{r} y_{ij}^{(d)}$$

- The method relies on the fact that the dichotomized variable distributions satisfy the *Rasch (1961) model*:

$$\log \frac{p(Y_j^{(d)} = 1|\theta)}{p(Y_j^{(d)} = 0|\theta)} = \theta - \beta_{jd}, \quad j = 1, \ldots, r, \; d = 1, \ldots, l-1$$

- The total score $Y_+^{(d)} = \sum_{j=1}^r Y_j^{(d)}$ is a *sufficient statistic* for the ability parameter θ

- The resulting *conditional probability* involved in $\ell_C^*(\boldsymbol{\beta})$ has expression:

$$p(y_{i1}^{(d)}, \ldots, y_{ir}^{(d)} | y_{i+}^{(d)}) = \frac{\exp\left(-\sum_{j=1}^r y_{ij}^{(d)} \beta_{jx}\right)}{\sum_{\mathbf{z}: z_+ = y_{i+}^{(d)}} \exp\left(-\sum_{j=1}^r z_j \beta_{jx}\right)}$$

with $\sum_{\mathbf{z}: z_+ = y_{i+}^{(d)}}$ extended to all binary vectors $\mathbf{z}$ of dimension r with elements summing up to $y_{i+}^{(d)}$

- The likelihood function $\ell^*_C(\boldsymbol{\beta})$ *depends only on the item parameters* (β_{jx} or β_j) collected in $\boldsymbol{\beta}$:
  - under 1P-GRM the identifiable parameters are β_{jx} for $j = 2, \dots, r$ and $x = 1, \dots, l-1$ (we use the constraint $\beta_{1x} = 0$, $x = 1, \dots, l-1$)
  - under 1P-RS-GRM the identifiable parameters are β_j for $j = 2, \dots, r$ (we use the constraint $\beta_1 = 0$), whereas the cut-points τ_x are not identified
- This function may be simply maximized by a *Newton-Raphson algorithm* based on:
  - pseudo *score vector*:

$$\mathbf{s}^*_C(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{s}^*_{C,i}(\boldsymbol{\beta}), \quad \mathbf{s}^*_{C,i}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log p(y_{i1}^{(d)}, \dots, y_{ir}^{(d)} | y_{i+}^{(d)})$$

  - pseudo *observed information matrix*:

$$\mathbf{H}^*_C(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \log p(y_{i1}^{(d)}, \dots, y_{ir}^{(d)} | y_{i+}^{(d)})$$

- The *asymptotic variance-covariance matrix* may be obtained by the sandwich formula:

$$\hat{V}_C^*(\hat{\boldsymbol{\beta}}_C^*) = \mathbf{H}_C^*(\hat{\boldsymbol{\beta}}_C^*)^{-1}\mathbf{S}(\hat{\boldsymbol{\beta}}_C^*)\mathbf{H}_C^*(\hat{\boldsymbol{\beta}}_C^*)^{-1}$$

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{s}_{C,i}^*(\boldsymbol{\beta})[\mathbf{s}_{C,i}^*(\boldsymbol{\beta})]'$$

- *Standard errors* may be extracted in the usual way from $\hat{V}_C^*(\hat{\boldsymbol{\beta}}_C^*)$

- On the basis of the pseudo score vector and information we can also implement a *Hausman (1978) test* for the hypothesis of normality in which the estimate $\hat{\boldsymbol{\beta}}_C^*$ is compared with the corresponding estimate obtained from the MML method

Simulation study of the CML estimator

- We *simulated* 1,000 samples of size n from the 1P-RS-GRM model for r response variable with $l = 5$ categories:
  - $r = 5, 10$
  - $n = 1000, 2000$
  - cut-points (τ_x) equal to $-2, -0.5, 0.5, 2$
  - difficulty parameters (β_j) as r equally distant points in $[-2, 2]$
  - four different latent distributions (all are standardized):
    - Normal(0,1)
    - Gamma(2,2)
    - LC1: latent class model with symmetric distribution based on mass probabilities 0.25, 0.5, 0.25 for increasing and equally spaced support points
    - LC2: as in LC1 but with skewed distribution based on mass probabilities 0.4, 0.5, 0.1
- For *all samples we fit* 1P-GRM and 1P-RS-GRM by the MML, MML-LC (k chosen by BIC), and CML methods

Simulation results for 1P-GRM: average values of absolute bias and RMSE for the estimates of parameters β_{jx}

			CML		MML		MML-LC	
Distrib.	n	r	abs.bias	RMSE	abs.bias	RMSE	abs.bias	RMSE
$N(0,1)$	1000	5	0.0121	0.1646	0.0112	0.1575	0.0019	0.1569
$N(0,1)$	2000	5	0.0043	0.1134	0.0032	0.1080	0.0089	0.1081
$N(0,1)$	1000	10	0.0085	0.1549	0.0085	0.1521	0.0156	0.1514
$N(0,1)$	2000	10	0.0041	0.1086	0.0038	0.1069	0.0216	0.1083
$\Gamma(2,2)$	1000	5	0.0070	0.1640	0.0634	0.1721	0.0053	0.1568
$\Gamma(2,2)$	2000	5	0.0025	0.1139	0.0618	0.1306	0.0080	0.1098
$\Gamma(2,2)$	1000	10	0.0150	0.1573	0.0474	0.1639	0.0128	0.1543
$\Gamma(2,2)$	2000	10	0.0087	0.1088	0.0455	0.1189	0.0138	0.1074
LC1	1000	5	0.0109	0.1619	0.0221	0.1586	0.0071	0.1572
LC1	2000	5	0.0068	0.1126	0.0183	0.1101	0.0059	0.1077
LC1	1000	10	0.0056	0.1553	0.0144	0.1545	0.0059	0.1526
LC1	2000	10	0.0031	0.1068	0.0099	0.1063	0.0031	0.1050
LC2	1000	5	0.0115	0.1650	0.0305	0.1634	0.0080	0.1587
LC2	2000	5	0.0044	0.1157	0.0251	0.1163	0.0039	0.1116
LC2	1000	10	0.0089	0.1569	0.0199	0.1573	0.0084	0.1544
LC2	2000	10	0.0033	0.1104	0.0174	0.1117	0.0034	0.1089

Simulation results for 1P-RS-GRM: average values of absolute bias and RMSE for the estimates of parameters β_j

Distrib.	n	r	CML abs.bias	CML RMSE	MML abs.bias	MML RMSE	MML-LC abs.bias	MML-LC RMSE
$N(0,1)$	1000	5	0.0042	0.1005	0.0007	0.0955	0.0055	0.0960
$N(0,1)$	2000	5	0.0012	0.0693	0.0030	0.0645	0.0078	0.0653
$N(0,1)$	1000	10	0.0022	0.0923	0.0040	0.0936	0.0168	0.0902
$N(0,1)$	2000	10	0.0013	0.0637	0.0030	0.0603	0.0199	0.0647
$\Gamma(2,2)$	1000	5	0.0000	0.0988	0.0130	0.0945	0.0075	0.0940
$\Gamma(2,2)$	2000	5	0.0015	0.0690	0.0125	0.0648	0.0105	0.0663
$\Gamma(2,2)$	1000	10	0.0078	0.0920	0.0072	0.0861	0.0109	0.0890
$\Gamma(2,2)$	2000	10	0.0046	0.0648	0.0108	0.0644	0.0154	0.0640
LC1	1000	5	0.0000	0.0978	0.0043	0.0905	0.0020	0.0945
LC1	2000	5	0.0037	0.0693	0.0040	0.0640	0.0025	0.0650
LC1	1000	10	0.0021	0.0947	0.0069	0.0968	0.0019	0.0801
LC1	2000	10	0.0011	0.0646	0.0036	0.0647	0.0012	0.0620
LC2	1000	5	0.0040	0.1003	0.0095	0.0955	0.0008	0.0953
LC2	2000	5	0.0028	0.0718	0.0082	0.0705	0.0038	0.0678
LC2	1000	10	0.0038	0.0951	0.0063	0.0844	0.0032	0.0819
LC2	2000	10	0.0007	0.0662	0.0044	0.0608	0.0011	0.0638

Main conclusions from the simulation study

- *Very similar performances* are observed in terms of efficiency under the normal distribution (the MML method is the most efficient, but the RMSE of the CML estimator is rather close)
- *A certain bias* arises for the MML method when the distribution is not normal (especially in the Gamma(2,2) case), whereas this bias is negligible for the CML method and the MML-LC method
- When the latent distribution is not normal, and then the MML estimator is biased, the CML method performs very similarly to the MML-LC method, with a *negligible loss of efficiency* of the CML method

Hausman test for normality of the latent trait

- The hypothesis of normality on which the MML method is based may be tested by a *Hausman test statistic*:

$$T = (\hat{\boldsymbol{\beta}}^*_M - \hat{\boldsymbol{\beta}}^*_C)'\hat{\mathbf{W}}^{-1}(\hat{\boldsymbol{\beta}}^*_M - \hat{\boldsymbol{\beta}}^*_C)$$

  with $\hat{\boldsymbol{\beta}}^*_M$ being the estimator based on the MML method under the constraint $\beta_{1x} = 0$, $x = 1, \ldots, l-1$

- $\hat{\mathbf{W}}$ is the *estimate of the variance-covariance matrix* of $\hat{\boldsymbol{\beta}}^*_M - \hat{\boldsymbol{\beta}}^*_C$ obtained starting from the sandwich formula ($\hat{\boldsymbol{\beta}}^*_M$ is a function of $\hat{\boldsymbol{\lambda}}_M$):

$$\hat{V}\begin{pmatrix}\hat{\boldsymbol{\lambda}}_M \\ \hat{\boldsymbol{\beta}}^*_C\end{pmatrix} = \begin{pmatrix}\mathbf{H}_M(\hat{\boldsymbol{\lambda}}_M) & \mathbf{O} \\ \mathbf{O} & \mathbf{H}^*_C(\hat{\boldsymbol{\beta}}^*_C)\end{pmatrix}^{-1} \mathbf{S}^*\begin{pmatrix}\hat{\boldsymbol{\lambda}}_M \\ \hat{\boldsymbol{\beta}}^*_C\end{pmatrix} \begin{pmatrix}\mathbf{H}_M(\hat{\boldsymbol{\lambda}}_M) & \mathbf{O} \\ \mathbf{O} & \mathbf{H}^*_C(\hat{\boldsymbol{\beta}}^*_C)\end{pmatrix}^{-1}$$

$$\mathbf{S}^*\begin{pmatrix}\hat{\boldsymbol{\lambda}}_M \\ \hat{\boldsymbol{\beta}}^*_C\end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix}\mathbf{s}_{M,i}(\hat{\boldsymbol{\lambda}}_M) \\ \mathbf{s}_{C,i}(\hat{\boldsymbol{\beta}}^*_C)\end{pmatrix} \begin{pmatrix}\mathbf{s}_{M,i}(\hat{\boldsymbol{\lambda}}_M)' & \mathbf{s}_{C,i}(\hat{\boldsymbol{\beta}}^*_C)'\end{pmatrix}$$

- Under the 1P-GRM model, the *asymptotic null distribution* of T is $\chi^2((r-1)(I-1))$

- Under the 1P-RS-GRM model, the *asymptotic null distribution* of T is $\chi^2(r-1)$

- If the hypothesis of normality is rejected, we estimate the model in a *semi-parametric way* by the MML-LC method

Application

- We consider a *dataset* (available in R) referred to a sample of $n = 392$ individuals from UK extracted from the Consumer Protection and Perceptions of Science and Technology section of the 1992 Euro-Barometer Survey
- The dataset is based on the responses to $r = 7$ items (with $l = 4$ *ordered categories*):
  - **Comfort** Science and technology are making our lives healthier, easier and more comfortable
  - **Environment** Scientific and technological research cannot play an important role in protecting the environment and repairing it
  - **Work** The application of science and new technology will make work more interesting
  - **Future** Thanks to science and technology, there will be more opportunities for the future generations
  - **Technology** New technology does not depend on basic scientific research
  - **Industry** Scientific and technological research do not play an important role in industrial development
  - **Benefit** The benefits of science are greater than any harmful effect it may have

Estimation results of CML and MML methods (under the constraint $\beta_{1x} = 0$, $x = 1, \ldots, I - 1$)

	1st cut-point	2nd cut-point	3rd cut-point
		CML	
Environment	1.966 (.487)	1.531 (.211)	-0.628 (.189)
Work	2.125 (.468)	1.688 (.208)	0.698 (.197)
Future	1.115 (.488)	1.051 (.198)	-0.121 (.183)
Technology	1.401 (.529)	1.395 (.202)	-0.598 (.195)
Industry	0.742 (.577)	0.514 (.220)	-1.121 (.189)
Benefit	1.580 (.425)	1.558 (.200)	0.203 (.185)
Log-lik.		-1734.413	
		MML	
Environment	1.885 (.486)	1.533 (.215)	-0.609 (.170)
Work	2.049 (.465)	1.716 (.213)	0.623 (.183)
Future	1.086 (.479)	1.076 (.203)	-0.116 (.168)
Technology	1.357 (.524)	1.394 (.207)	-0.576 (.176)
Industry	0.719 (.563)	0.499 (.227)	-1.013 (.167)
Benefit	1.524 (.424)	1.590 (.207)	0.169 (.171)
Log-lik.		-3014.706	

- The Hausman test leads to *reject the hypothesis of normality*:

$$T = 39.9106, \quad \text{Prob}\left(\chi^2_{18} > T\right) = 0.002146$$

- We then estimate the model by the *MML-LC method with $k = 3$ latent classes* obtaining:

c	$\hat{\xi}_c$	$\hat{\pi}_c$
1	-1.158	0.265
2	-0.073	0.548
3	1.851	0.187

- The latent distribution is standardized and *skewed* (skewness index $= 0.777$)

Estimation results from the MML-LC method with $k = 3$

	1st cut-point	2nd cut-point	3rd cut-point
Environment	1.848 (.537)	1.497 (.282)	-0.623 (.182)
Work	2.011 (.528)	1.682 (.293)	0.639 (.185)
Future	1.067 (.480)	1.050 (.225)	-0.116 (.164)
Technology	1.332 (.519)	1.371 (.262)	-0.582 (.212)
Industry	0.701 (.602)	0.493 (.203)	-1.030 (.219)
Benefit	1.506 (.479)	1.557 (.282)	0.174 (.158)
Log-lik.		-3010.826	

- The *estimates of the item parameters* are rather similar with respect to the MML method and the *log-likelihood is higher*
- The *influence on prediction* of the latent ability may be considerable (prediction for a certain subject on the basis of the sequence of responses he/she provided through a posterior expected value)

Conclusions

- The proposed method for estimating the parameters of a constrained version of GRM is *very simple to implement* and is *consistent* under any true distribution of the latent trait
- The method seems to provide an *efficient estimator* (efficiency close to the MML estimator under the normal distribution)
- It also allows us to implement a *Hausman test for the hypothesis of normality*
- When the hypothesis of normality is rejected, the *semi-parametric MML-LC method* is an interesting alternative to MML
- Even if significant differences are not observed in terms of estimates of the item parameters, the effect on *prediction of the ability* levels may be relevant

References

- Aitkin, M. and Bock, R. D. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, **46**, pp. 443-459.
- Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, **59**, pp. 19-35.
- Baetschmann, G., Staub, K. E. and Winkelmann, R. (2011). Consistent estimation of the fixed effects ordered logit model, *IZA Discussion Paper*, **5443**.
- 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**, pp. 1-38.
- Hausman, J. (1978). Specification Tests in Econometrics, *Econometrica*, **46**, pp. 1251–1271.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology, in *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, 321-333.
- Samejima, F. (1969). Estimation of ability using a response pattern of graded scores, *Psychometrika Monograph*, **17**.
- van der Ark, L.A. (2001). Relationships and properties of polytomous item response theory models, *Applied Psychological Measurement*, **25**, pp. 273–282.