METRON - *International Journal of Statistics*
2010, vol. LXVIII, n. 1, pp. 87-94

MARIA IANNARIO

On the identifiability of a mixture model for ordinal data

Summary - In this article we discuss the identifiability of a probability model which has been proven useful for capturing the main features of ordinal data generated by rating surveys. Specifically, we show that the mixture of a shifted Binomial and a Uniform discrete distribution is identifiable when the number of categories is greater than three.

Key Words - Identifiability criteria; CUB models; Ordinal data.

1. Introduction

The statistical approach for modelling ordinal data is mainly based on Generalized Linear Models (GLM), proposed by McCullagh (1980) and McCullagh and Nelder (1989). When data are collected as ordered responses to a sequence of items (concerning preferences, evaluations, proficiency, etc.), current literature has developed a vast amount of results known as *Item Response Theory* (see Bock and Moustaki (2007) for an updated review).

In the last few years, a different approach has been introduced for explaining the behaviour of respondents when faced to a single item characterized by ordinal choices (Piccolo, 2003; D'Elia and Piccolo, 2005). In this context, a class of statistical models known as CUB has been derived from the psychological mechanism of respondents and their parameters are interpreted with reference to unobserved components. In fact, the acronym CUB stems from the presence of ***C***ovariates in ***U***niform and shifted ***B***inomial mixtures, and this paper will be concerned with CUB models without covariates, generally denoted as CUB $(0, 0)$.

Asymptotic statistical inference for these models([1]), without and with covariates, has been developed by Piccolo (2006), an effective E-M procedure for

([1]) CUB models have been successfully applied in several fields including Linguistics (Balirano and Corduas, 2008; Cappelli and D'Elia, 2004), Medicine (D'Elia, 2008), Marketing (Iannario and Piccolo, 2010), University services performance (Corduas *et al.*, 2009), Ranking preference (Iannario, 2007), Sensometrics (Piccolo and D'Elia, 2008), Qualitative analysis (Piccolo, 2008; Piccolo and Iannario, 2008).

Received January, 2009 and revised February, 2009.

maximum likelihood estimators has been implemented and a related software is freely available (Iannario and Piccolo, 2009).

Hitherto there is no explicit statement about the identifiability of CUB models although this point is important for both theoretical and applied points of view. As discussed by Teicher (1963); Titterington *et al.* (1985), 35-42; McLachlan and Peel (2000), 26-28; 43-44, this problem is a peculiar issue for the specification and inference of mixture distributions. Most of the published results are relevant for mixtures of the same family of distributions (Binomial, Gaussian, Gamma, etc.) whereas Yakowitz and Spragins (1968) discuss mixtures of Gaussian and Gamma distributions and Atienza *et al.* (2006) are concerned with general finite mixtures. Since CUB models are obtained by mixing a shifted Binomial and a discrete Uniform distribution, we explicitly look for a straightforward proof to show their identifiability.

The paper is organized as follows: in the next section we will briefly introduce notations for CUB models and in Section 3 we will set some preliminary lemmas. Then, in Section 4 we will prove the main theorem about identifiability of these models. Some concluding remarks end the paper.

2. Notations for CUB models

In CUB models, ratings are interpreted as the result of a cognitive process, where the judgement is intrinsically continuous but it is expressed in a discrete way within a prefixed scale of m categories. The rationale of this approach stems from the interpretation of the final choices of respondents as a weighted combination of a personal *agreement/feeling* and some intrinsic *uncertainty/fuzziness*.

The first component is expressed by a shifted Binomial random variable, which is the discrete version over the support $\{1, 2, \ldots, m\}$ of an unobserved unimodal distribution. This random variable is generally adequate to model expressed choices since it maps a continuous latent variable into a discrete set of values.

The second component is expressed by a Uniform random variable, which is the extreme solution for a totally indifferent choice. If conveniently weighted, it is able to describe the inherent uncertainty of any evaluation process constrained to fit into discrete categories.

Formally, for a given $m > 3$, we consider the expressed rating r as the realization of a random variable R, whose probability distribution is given by:

$$Pr(R = r) = \pi \binom{m-1}{r-1} \xi^{m-r} (1-\xi)^{r-1} + (1-\pi)\frac{1}{m}, \quad r = 1, 2, \ldots, m.$$

From a computational point of view, when $\xi > 0$, the whole probability distribution $p_r = Pr(R = r)$ is efficiently derived by recursive relationships:

$$\begin{cases} p_1 = \pi \left[\xi^{m-1} - \dfrac{1}{m}\right] + \dfrac{1}{m}; \\ p_{r+1} = \dfrac{p_r\,(1-\xi)(m-r) + [r - m(1-\xi)]\,(1-\pi)/m}{r\,\xi}, \\ r = 1, 2, \ldots, m-1. \end{cases}$$

In order to interpret this model, we are assuming that each subject acts *as if* his/her final choice would be generated with *propensities* (π) and $(1-\pi)$ from the two component distributions, respectively. Then, $(1-\xi)$ is a measure of agreement/feeling towards the item and $(1-\pi)$ is a measure of uncertainty that accompanies the choice. From a statistical viewpoint, ξ may be interpreted as mostly related to location measures and strongly determined by the skewness of responses as it increases when respondents select low ratings, and vice versa. Instead, π adds dispersion to the shifted Binomial distribution and thus it should be related to variability concepts; indeed, it increases frequencies in each category and modifies the heterogeneity of the distribution. Briefly, ξ is mainly related to the categories of the ordinal response whereas π is mostly related to the comparison among probabilities.

This model is fully specified by the parameter vector $\boldsymbol{\theta} = (\pi, \xi)'$ and the *parametric space* of this random variable is defined by:

$$\Omega(\boldsymbol{\theta}) = \Omega(\pi, \xi) = \{(\pi, \xi) : 0 < \pi \leq 1,\ 0 \leq \xi \leq 1\}.$$

Observe that the left border of the unit square is not included in the parametric space since $\pi = 0$ would imply the non-identifiability of ξ. Thus, the discrete Uniform random variable is a limiting distribution of this class.

This probability structure proved flexible for fitting real case studies as it accounts for skewness, intermediate modes and flat distributions (Piccolo, 2003). Moreover, a simple and parsimonious variant of these models is also able to accommodate to degenerate situations (Iannario, 2010; Corduas *et al.*, 2009).

An interesting consequence of the approach is the ability to identify a collection of estimated CUB models as a set of points located in the parametric space. This representation allows effective comparisons and interpretations of data and models, and it has been exploited by Corduas (2008) for clustering ordinal data.

In common applications m is a low integer value. For instance, $m = 5$ is a frequent choice for marketing surveys, $m = 7$ is a convenient option for finer evaluation questionnaires and $m = 9$ is generally required in sensometric studies when judges/consumers assess their preferences on some hedonic scale.

In this regard, the constraint $m > 3$ avoids considering degenerate ($m = 1$), indeterminate ($m = 2$) or saturated ($m = 3$) models, respectively.

Then, we will define as *admissible* a CUB model such that $m > 3$ and $\boldsymbol{\theta} \in \Omega(\boldsymbol{\theta})$.

3. Preliminary lemmas

In this section, we will set two useful lemmas. As a rule, we assume that -for any CUB model- the number m of categories is known in advance and fixed.

Lemma 1. *For any admissible CUB model, the real-valued function $\rho(\xi)$ defined on $\xi \in [0, 1]$ by:*

$$\rho(\xi) = \frac{p_m - p_1}{p_1 - 1/m} = \frac{(1-\xi)^{m-1} - \xi^{m-1}}{\xi^{m-1} - 1/m}, \qquad \xi \neq \xi_0\,,$$

where $\xi_0 = \left(\frac{1}{m}\right)^{\frac{1}{m-1}}$, does not depend on π. The function $\rho(\xi)$ is continuous on the admissible range of ξ (except for the singular value ξ_0) and it is therein monotonically increasing.

Function $\rho(\xi)$ is increasing from $\xi = 0$, where $\rho(0) = -m$, to $\xi = \xi_0^-$ (when it approaches $+\infty$); again, it is monotonically increasing from $\xi = \xi_0^+$ (when it approaches $-\infty$) to $\xi = 1$, where $\rho(1) = -m/(m-1)$. This monotonic behavior follows from the derivative:

$$\rho'(\xi) \propto \xi^{m-2} + (1-\xi)^{m-2} - m\,\xi^{m-2}\,(1-\xi)^{m-2} > 0, \quad \forall \xi \neq \xi_0.$$

As a matter of fact, both the minimum of the first two terms and the maximum of the last term are attained for $\xi = 1/2$. Then, the minimum of the first two terms ($= 2^{3-m}$) is strictly larger than the maximum of last one ($= m\,2^{2(2-m)}$), uniformly over the range of ξ, when $\xi \neq \xi_0$.

Finally, given $\pi > 0$, we observe that: $\xi \neq \xi_0 \iff p_1 \neq 1/m$.

Lemma 2. *For any admissible CUB model and a given real number $\overline{\rho}$, the equation $\rho(\xi) = \overline{\rho}$ admits one or two real solutions ξ_1, ξ_2. Then, with respect to the singular point ξ_0, their location is characterized by:*

$$\begin{aligned}
\xi = \xi_2 > \xi_0 \ &\text{if and only if } -\infty < \overline{\rho} < -m;\\
\xi = \xi_1 < \xi_0\,;\ \xi = \xi_2 > \xi_0 \ &\text{if and only if } -m \le \overline{\rho} \le -m/(m-1);\\
\xi = \xi_1 > \xi_0 \ &\text{if and only if } -m/(m-1) < \overline{\rho} < +\infty\,.
\end{aligned}$$

The first part of Lemma is immediately proven if one considers definition and range of $\rho(\xi)$, as discussed in Lemma 1. Specifically, by continuity and monotonicity of $\rho(\xi)$, $\rho(\xi) < -m$ if and only if $\xi > \xi_0$; as a consequence, a value $\xi > \xi_0$ exists such that $\rho(\xi) = \overline{\rho}$ when $\overline{\rho} \in (-\infty, -m)$. Similarly, $\rho(\xi) > -m/(m-1)$ if and only if $\xi < \xi_0$ and there is just one solution to $\rho(\xi) = \overline{\rho}$. Instead, when $\overline{\rho} \in [-m, -m/(m-1)]$, both left and right branches of the function $\rho(\xi)$ support values in this interval, and there are two solutions $\xi_1 < \xi_0$ and $\xi_2 > \xi_0$, respectively, for the equation: $\rho(\xi) = \overline{\rho}$.

A visual support to Lemma 2 is offered by Figure 1, when $m = 9$.

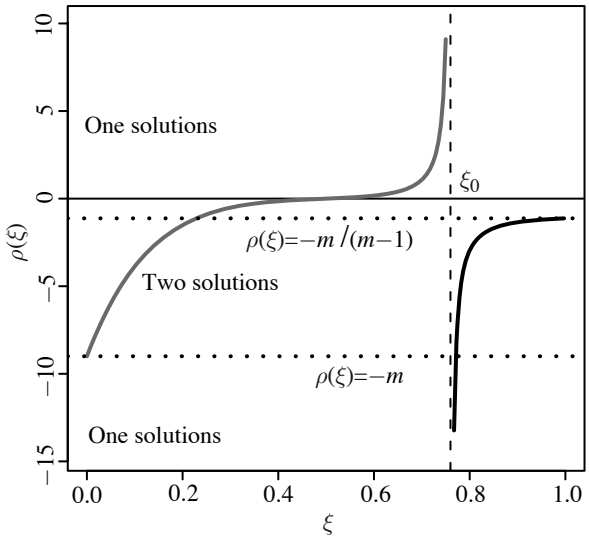

Figure 1. Solutions of the equation $\rho(\xi) = \overline{\rho}$, when $m = 9$.

4. Identifiability of CUB models

A CUB model is *identifiable* if and only if, for any given probability distribution $p_r = p_r(\boldsymbol{\theta})$, $r = 1, 2, \ldots, m$, generated by an admissible parameter vector $\boldsymbol{\theta} = (\pi, \xi)'$, it is not possible to find a different parameter vector $\boldsymbol{\theta}^* = (\pi^*, \xi^*)'$ such that:

$$p_r(\boldsymbol{\theta}) \equiv p_r(\boldsymbol{\theta}^*), \quad r = 1, 2, \ldots, m.$$

Theorem. *An admissible CUB model is identifiable.*

Let $p_r = p_r(\boldsymbol{\theta})$, $r = 1, 2, \ldots, m$ be a given probability distribution generated by an admissible parameter vector $\boldsymbol{\theta} = (\pi, \xi)'$. We have to prove that the

set $\{p_1, p_2, \ldots, p_m\}$ uniquely specifies $\boldsymbol{\theta}$. Of course, this statement is true if we will prove that $\{p_1, p_m\}$ will uniquely specify the parameters vector $(\pi, \xi)'$.

Then, we distinguish the proof according to $p_1 \neq 1/m$ or $p_1 = 1/m$, respectively.

► If $p_1 \neq 1/m$ (that is, $\xi \neq \xi_0$), from the values of the given distribution, we compute the real number:

$$\overline{\rho} = \frac{p_m - p_1}{p_1 - 1/m}\,.$$

Then, the solutions for ξ and π are determined, respectively, by:

$$\begin{cases} \rho(\xi) = \overline{\rho}; \\ \pi = \dfrac{p_1 - 1/m}{\xi^{m-1} - 1/m}\,. \end{cases}$$

If there is a unique $\xi = \xi^*$ that solves the first equation, then the second equation will specify a unique $\pi = \pi^*$. Thus, in order to prove identifiability it is necessary and sufficient to prove that the first equation, for given p_1 and p_m, admits one and only one admissible solution for both parameters.

Now, if $\overline{\rho} \notin [-m, -m/(m-1)]$, the solution fo ξ is unique according to Lemma 2, and the CUB model is identified by the given probability distribution. Moreover, the solution for π is admissible for it has been shown that there is at most one solution. Since p_1 and p_m are obtained by an admissible CUB model, there is at least one solution, and this implies uniqueness of solution.

Instead, if $\overline{\rho} \in [-m, -m/(m-1)]$ we obtain two different real solutions (ξ_1, ξ_2) for ξ and two corresponding solutions (π_1, π_2) for π. However, these solutions imply necessarily that:

$$\frac{\pi_1}{\pi_2} = \frac{\xi_2^{m-1} - 1/m}{\xi_1^{m-1} - 1/m} < 0$$

since, as shown by Lemma 2, the admissible solutions ξ_1 and ξ_2 are smaller and greater than ξ_0, respectively. However, if the ratio of π_1/π_2 is negative, one of the π solutions cannot be accepted. As a consequence, there is one and only one admissible solution for (π, ξ).

► If $p_1 = 1/m$, then $\xi = \xi_0$ and π is uniquely determined by:

$$\pi = \frac{p_m - 1/m}{(1 - \xi_0)^{m-1} - 1/m}\,.$$

Finally, if $p_1 = p_m$ then $\xi = 1/2$ (distribution is symmetric) and $\pi = \frac{p_m - 1/m}{(1/2)^{m-1} - 1/m}$ is always well defined since denominator could be 0 if and only if $m = 1, 2$; thus, admissible CUB models never could meet this case.

5. Concluding remarks

In this paper we focused on the main issue of the identifiability of CUB models without covariates. When there are information on the raters and the sample size is adequate, the introduction of significant relationships among parameters and subjects' covariates improves interpretation of data and fitting of the models.

In this regards, recent literature warns against assuming automatic identifiability of CUB models with covariates even when linear relationships are of full rank, as shown by Hennig (2000) (with counterexamples) and Frühwirth-Schnatter (2006) in regression contexts. As a consequence, more research is necessary in this field for extending the previous formal proof of identifiability of CUB models.

Acknowledgments

The work has been partially supported by PRIN-2006 research project: "Stima e verifica di modelli statistici per l'analisi della soddisfazione degli studenti universitari" and benefited from research structures of CFEPSR, Portici. Suggestions by referees significantly improved the paper and are gratefully acknowledged.

REFERENCES

Atienza N., Garcia-Heras J. and Muñoz-Pichardo J. M. (2006) A new condition for identifiability of finite mixture distributions, *Metrika*, 63, 215–221.

Balirano G. and Corduas M. (2008) Detecting semiotically expressed humor in diasporic TV productions, *HUMOR: International Journal of Humor Research*, 3, 227–251.

Bock R. D. and Moustaki I. (2007) Item response theory in a general framework, In: *Psychometrics*, (eds. C. R. Rao and S. Sinharay), Handbook of Statistics 26, 469–513.

Cappelli C. and D'Elia A. (2004) La percezione della sinonimia: un'analisi statistica mediante modelli per ranghi, In: *Le poids des mots - Actes de JADT2004*, (eds. Prunelle G., Fairon C. and Dister A.), Presses Universitaires de Louvain, Belgium, 229–240.

Corduas M. (2008) A testing procedure for clustering ordinal data by cub models, *Proceeding of the Joint SFC-CLADAG Meeting*, ESI, Napoli, 245–248.

Corduas M., Iannario M. and Piccolo D. (2009) A class of statistical models for evaluating services and performances, In: *Statistical Methods for the Evaluation of Educational Services and Quality of Products*, (eds. M. Bini, P. Monari, D. Piccolo, L. Salmaso), Contribution to Statistics, Springer, 99–117.

D'Elia A. (2008) A statistical modelling approach for the analysis of TMD chronic pain data, *Statistical Methods in Medical Research*, 17, 389–403.

D'Elia A. and Piccolo D. (2005) A mixture model for preference data analysis, *Computational Statistics Data Analysis*, 49, 917–934.

Frühwirth-Schnatter S. (2006) *Finite Mixture and Markov Switching Models*, Springer Series in Statistics, Springer, New York.

HENNIG C. (2000) Identifiability of models for clusterwise linear regression, *Journal of Classification*, 17, 273–296.

IANNARIO M. (2007) A statistical approach for modelling Urban Audit Perception Surveys, *Quaderni di Statistica*, 9, 149–172.

IANNARIO M. (2010) *Modelling shelter choices in ordinal surveys*, submitted for publication.

IANNARIO M. and PICCOLO D. (2009) *A program in R for CUB models inference*, Version 2.0, available at *www.dipstat.unina.it*

IANNARIO M. and PICCOLO D. (2010) A new statistical model for the analysis of customer satisfaction, *Quality Technology & Quantitative Management*, 7, 149–168.

MCCULLAGH P. (1980) Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society*, Series B, 42, 109–142.

MCCULLAGH P. and NELDER J. A. (1989) *Generalized Linear Models*, 2^{nd} edition, Chapman and Hall, London.

MCLACHLAN G. and PEEL G. J. (2000) *Finite Mixture Models*, J. Wiley & Sons, New York.

PICCOLO D. (2003) On the moments of a mixture of uniform and shifted binomial random variables, *Quaderni di Statistica*, 5, 85–104.

PICCOLO D. (2006) Observed information matrix for MUB models, *Quaderni di Statistica*, 8, 33–78.

PICCOLO D. (2008) Modelling University students' final grades by ordinal variables, *Quaderni di Statistica*, 10, 205–226.

PICCOLO D. and D'ELIA A. (2008) A new approach for modelling consumers' preferences, *Food Quality and Preference*, 19, 247–259.

PICCOLO D. and IANNARIO M. (2008) Qualitative and quantitative models for ordinal data analysis, *Proceedings of MTISD 2008, Methods, Models and Information Technologies for Decision Support Systems*, Università del Salento, Lecce, 140–143.

TEICHER H. (1963) Identifiability of finite mixtures, *The Annals of Mathematical Statistics*, 34, 1265–1269.

TITTERINGTON D.M., SMITH A.F.M. and MAKOV U.E. (1985) *Statistical Analysis of Finite Mixture Distributions*, J. Wiley & Sons, New York.

YAKOWITZ S. J. and SPRAGINS J. D. (1968) On the identifiability of finite mixtures, *The Annals of Mathematical Statistics*, 39, 209–214.

MARIA IANNARIO
Dipartimento di Scienze Statistiche
Università di Napoli Federico II
maria.iannario@unina.it