QdS

Journal of Methodological and Applied Statistics

QUADERNI DI STATISTICA
VOLUME 15 - 2013

LIGUORI EDITORE

Volume 15, 2013

ISSN 1594-3739 (printed edition)
eISSN 2037-6901 (digital edition)

Registration n. 5264, 6 December 2001, Court of Justice at Naples

ISBN **978 - 88 - 207 - 5290-3** (p. e.)
eISBN **978 - 88 - 207 - 5291-0** (d. e.)

Director: Gennaro Piccolo

First italian edition, December 2013
Printed in Italy in December 2013 by Liguori Editore, Naples (IT)

This publication is financially supported by:

- Dipartimento di Scienze Economiche e Statistiche, Università di Salerno
- Dipartimento di Scienze Politiche, Università di Napoli Federico II

Our policy is to use permanent paper from mills that operate a sustainable foresty policy and which has been manufactured from pulp that is processed using acid-free and elementary chlorine free practice. Furthermore, we ensure that the materials used have met acceptable environmental accreditations standard.

Index

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

Francesco Bartolucci
Department of Economics, University of Perugia (IT)
E-mail: bart@stat.unipg.it

Silvia Bacci
Department of Economics, University of Perugia (IT)
E-mail: silvia.bacci@stat.unipg.it

Claudia Pigini
Department of Economics, University of Perugia (IT)
E-mail: pigini@stat.unipg.it

Summary: 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. In this paper, we focus on a class of item response theory models for ordinal responses, named graded response models, taking into account a constrained version with constant discriminating indices. In this class of models, we introduce a conditional likelihood estimator, which requires no assumptions on the latent distribution. Through a Monte Carlo study, we compare the behavior of the proposed estimator with that of two competitors, based on the maximization of the marginal likelihood, which is computed assuming the normality of the latent variable in one case, and the discreteness of the latent variable in the other case. The method also allows us to implement a Hausman test to compare the marginal and conditional likelihood estimators, which results in a test for the normality assumption on the latent distribution. We conclude with an application based on data coming from the administration of a questionnaire on the perception of science and technology.

Keywords: Graded response model; Hausman test; Latent class analysis; Rating scale parametrization.

1. Introduction

In the Item Response Theory (IRT) literature, several types of models have been proposed for items with ordered response categories; for a review see, among others, van der Linden and Hambleton (1997) and Nering and Ostini (2010). Among the most popular ones, there are the Partial Credit Model (PCM - Masters, 1982; see also Andrich, 1978, for the rating scale version of PCM) and the Graded Response Model (GRM - Samejima, 1969), which differ for the type of link function (Agresti, 2002) and for the item parametrization. PCM is based on logits for adjacent categories, whereas GRM rests on global logits. Moreover, PCM is characterized by items with the same discriminating power, whereas in GRM each item has a specific discrimination parameter. The main consequence of this difference is that PCM may be estimated through the Conditional Maximum Likelihood (CML) approach, which is a particularly appealing method, as it does not require any specific assumption about the latent trait distribution, and it is consistent. Since the CML approach cannot be directly applied to GRM, other estimation methods have to be adopted in this last case, being the Marginal Maximum Likelihood (MML - Bock and Lieberman, 1970; Bock and Aitkin, 1981) the most popular. The main drawback of MML is that it is a parametric approach, which is typically based on the hypothesis of normality of the latent trait distribution and it is, therefore, consistent only under the correct specification of the distributional assumption. Moreover, MML requires some computational effort, as quadrature methods are needed for computing the log-likelihood. Alternatively to the MML, a semi-parametric approach based on assuming a discrete latent trait can be also applied, which is more flexible and easier to implement than MML.

In this paper, we consider a constrained version of GRM consisting in a model with free item difficulty parameters, whereas all items are assumed to discriminate in the same way, analogously to PCM. For this restricted class of models, it is possible to specify a CML estimator, following an approach proposed in a different field by Baetschmann et al. (2011). The first aim of the paper is to compare, for the restricted class of IRT models of interest, the performance of the three maximum likelihood approaches at issue in terms of efficiency and robustness: the parametric MML approach that relies on the hypothesis of normality of the latent trait, the semi-parametric MML based on the discreteness of the latent trait, and the non-parametric one based on CML. Furthermore, we show that the MML and CML estimators can be compared by means of a Hausman test: MML is consistent and more efficient only when the latent distribution is normal, whereas CML is robust with respect to departures from normality of the latent trait but less efficient. Therefore, the Hausman test can be seen as a test for the normality assumption.

The remainder of the article is organized as follows. In the next section we illustrate the class of polytomous item response models of interest. Then, we provide a description of the above mentioned maximum likelihood approaches (CML and MML), the efficiency and robustness of which are compared through a Monte Carlo study. We also describe the Hausman test for the normality assumption, which compares the proposed CML estimator with the two MML estimators. The properties of this test are investigated through a Monte Carlo study. We conclude with an application, based on data about the perception of science and technology, and with some final remarks.

2. A class of item response models

The class of polytomous item response models on which we base our study relies on the Graded Response Model (GRM) introduced by Samejima (1969). Let θ indicate the level of the latent trait Θ and let X_j denote the response variable for the j-th item ($j = 1, \ldots, r$), which is assumed to have l_j categories, indexed from 0 to $l_j - 1$.

GRM is based on the following main assumptions:

- *unidimensionality*: all test items measure the same latent trait Θ;
- *local independence*: the response variables $X_1, \ldots, X_r$ are independently distributed given Θ, that is,

$$p(X_1, \ldots, X_r|\theta) = \prod_{j=1}^{r} p(X_j|\theta);$$

- *monotonicity*: $p(X_j \geq x|\theta)$ is nondecreasing in θ, that is,

$$p(X_j \geq x|\theta_1) \geq p(X_j \geq x|\theta_2), \quad j = 1, \ldots, r, \; x = 1, \ldots, l_j - 1.$$

for $\theta_1 > \theta_2$.

In particular, GRM is based on the following parametrization:

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

where γ_j measures the *discriminating power* of item j and β_{jx} corresponds to the *difficulty level* of item j and category x. These item difficulty levels satisfy the ordering $\beta_{j1} < \beta_{j2} < \ldots < \beta_{j,l_j-1}$. We also observe that the left side of (1) expresses a global logit, which corresponds to a cumulative logit with opposite sign (according to the definition of cumulative logit by Agresti, 2002) and compares the probability that the item response is in category x or higher with the probability that it is in a lower category. This parametrization distinguishes GRM from other types of item response models that adopt different link functions (i.e., adjacent logits in the case of partial credit models and continuation ratio logits in the case of sequential models).

In the following, we focus on a special case of GRM in which all the items discriminate in the same way (van der Ark, 2001), that is, we assume that

$$\gamma_j = 1, \quad j = 1, \ldots, r.$$

We refer to the resulting model as One Parameter GRM (1P-GRM). Moreover, we consider a further special case based on a rating scale parametrization in which the parameters β_{jx} are constrained so that the distance between difficulty levels from category to category is the same for every item. This constraint is expressed as

$$\beta_{jx} = \beta_j + \tau_x, \quad j = 1, \dots, r, \ x = 0, \dots, l-1,$$

where β_j represents the difficulty of item j and τ_x indicates the difficulty of response category x, which is the same for all items. The resulting model is denoted by 1P-RS-GRM. We note that the rating scale parametrization makes sense only when all items have the same number of response categories, that is, $l_j = l$, $j = 1, \dots, r$.

3. Maximum likelihood estimation approaches

Given a sample of observations x_{ij}, $i = 1, \dots, n$, $j = 1, \dots, r$, different maximum likelihood estimation methods may be used for the estimation of 1P-GRM and 1P-RS-GRM. In the following we first describe the standard methods based on a random-effects formulation, such as the parametric Marginal Maximum Likelihood (MML) method and its semi-parametric version. Then, we propose an alternative non-parametric approach, which rests on the Conditional Maximum Likelihood (CML) method.

3.1. Standard methods for maximum likelihood estimation

Among the most well-known estimation methods for IRT models, those based on a random-effects formulation conceive the ability Θ as a random variable having a certain distribution. More specifically, under the assumption that Θ is normally distributed, we can use the MML method (Bock and Lieberman, 1970; Bock and Aitkin, 1981) based on maximizing the *marginal log-likelihood*

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

with $\phi(\theta_i; 0, \sigma^2)$ denoting the density function of the distribution $N(0, \sigma^2)$ and where the parameter vector $\boldsymbol{\eta}$ contains the item parameters further to the variance of the latent distribution σ^2. The main advantage of the MML method is that the parameter estimates are consistent under the above assumption of normality. However, if this assumption does not hold, parameter estimates are typically biased; moreover, the method has a certain complexity due to the presence of the integral in expression (2). There exist several approaches to deal with this integral, such as Gaussian quadrature based methods (Abramowitz and Stegun, 1965), among which the adaptive Gaussian quadrature (Pinheiro and Bates, 1995), and Monte Carlo based integration methods. Each of these methods replaces the integral in (2) by a finite sum and the resulting expression is then maximized through direct algorithms, such as the Newton-Raphson or the Fisher-scoring, or indirect algorithms, such as the Expectation-Maximization (EM - Dempster et al., 1977); see Bacci et al. (2014) and Bartolucci et al. (2014) for an `R` implementation of this algorithm. In the case of Gaussian quadrature, the number of terms of this sum is typically moderate, such as 21.

In order to reduce the dependence of the parameter estimates on parametric assumptions about the latent distribution, we can rely on a semi-parametric approach based on the maximization of the marginal likelihood under the assumption that the latent trait has a discrete distribution. This distribution is based on k support points $\xi_1, \ldots, \xi_k$, which identify latent classes of individuals with homogeneous unobservable characteristics, and corresponding weights $\pi_1, \ldots, \pi_k$, where $\pi_c > 0$, $c = 1, \ldots, k$, and $\sum_{c=1}^{k} \pi_c = 1$. The method at issue is denoted as MML-LC and is based on the following marginal log-likelihood function:

$$\ell_{LC}(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log \sum_{c=1}^{k} \pi_c \prod_{j=1}^{r} p(x_{ij}|\theta_i = \xi_c), \tag{3}$$

with the parameter vector $\boldsymbol{\psi}$ including, further to the item parameters, the support points of the latent distribution and the corresponding weights. The maximization of function (3) is easier to perform than that of function (2) because it skips the problem of solving the integral involved in the MML approach. Typically, the maximum of MML-LC is found through an EM algorithm.

The main drawback of the MML-LC method is the need of choosing properly k. In this regard, we may use the information criteria, such as AIC (Akaike, 1973) based on the index

$$AIC = -2\ell_{LC}(\hat{\boldsymbol{\psi}}) + 2\#\text{par},$$

or BIC (Schwarz, 1978) based on the index

$$BIC = -2\ell_{LC}(\hat{\boldsymbol{\psi}}) + \log(n)\#\text{par},$$

where $\hat{\boldsymbol{\psi}}$ denotes the vector of estimated parameters and $\#\text{par}$ is the number of free parameters, seen as a measure of the model complexity. On the basis of these criteria, the selected number of classes is the one corresponding to the minimum value of AIC or BIC. It has to be noted that the selected number of classes strongly depends on the criterion adopted.

Relying on the main results reported in the literature about finite mixture models (see, among others, McLachlan and Peel (2000), Chapter 8, and the references therein), we prefer to use BIC rather than AIC. Indeed, BIC tends to select a more parsimonious model than AIC and so it allows us to limit the instability problems in the maximization process that may arise with large values of k. On the other hand, using more support points leads to a more flexible latent distribution. In any case, it may be verified that the estimates of the parameters involved in the model that are constant across the latent classes (i.e., β_{jx}) are substantially unaffected by the specific value of k that is selected.

3.2. Conditional maximum likelihood method

We suggest an alternative to the parametric and semi-parametric MML approaches for the maximum likelihood estimation of the models illustrated in Section 2. The proposed method is based on CML (Andersen, 1970, 1972; Chamberlain, 1980) and, as shown in Andersen (1970), it gives a consistent estimator of the β_{jx} parameters under mild regularity conditions and independently of the true distribution of Θ.

There exist several ways to implement CML with ordered response modalities; see Baetschmann et al. (2011) for a review in a different context. Here we rely on the idea of reducing the model of interest to a model for binary data by suitably dichotomizing the response variables and considering the contributions to the conditional likelihood as those resulting from all the possible dichotomizations of these variables.

For the case in which the response variables have the same number l of response categories, we consider the $l-1$ possible dichotomizations indexed by $d = 1,\ldots,l-1$. For each dichotomization d we transform the response variables X_j, for every individual, in the binary variables

$$Y_j^{(d)} = 1\{X_j \geq d\}, \quad j = 1,\ldots,r,$$

where $1\{\cdot\}$ is the indicator function. We then maximize the function given by the sum of the conditional log-likelihood functions (Andersen, 1970) corresponding to each dichotomization:

$$\ell_C(\boldsymbol{\beta}) = \sum_{d=1}^{l-1} \log p(y_{i1}^{(d)},\ldots,y_{ir}^{(d)}|y_{i+}^{(d)}), \quad y_{i+}^{(d)} = \sum_{j=1}^{r} y_{ij}^{(d)}. \tag{4}$$

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

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

In fact, it is well known that the total score $Y_+^{(d)} = \sum_{j=1}^r Y_j^{(d)}$ is a sufficient statistic for the ability parameter θ (seen as a fixed parameter) and the thresholds; see Andersen (1970, 1972) and Chamberlain (1980). The resulting conditional probability involved in $\ell_C(\boldsymbol{\beta})$ has expression

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

with $\sum_{\boldsymbol{z}:z_+=y_{i+}^{(d)}}$ extended to all binary vectors $\boldsymbol{z}$ of dimension r with elements summing up to $y_{i+}^{(d)}$. Note that since the joint probability of $y_{i1}^{(d)},\ldots,y_{ir}^{(d)}$ does not depend on the individual parameters θ_i, the likelihood function (4) does not depend on the distribution of the latent trait. In particular, $\ell_C(\boldsymbol{\beta})$ only depends on the item parameters (β_{jx} or β_j) collected in $\boldsymbol{\beta}$. In fact, different sets of parameters are identifiable under the conditional approach depending on the estimated model:

- under 1P-GRM, the identifiable parameters are β_{jx} for $j = 2, \ldots, r$ and $x = 1, \ldots, l-1$ (we use the constraint $\beta_{1x} = 0$, $x = 1, \ldots, l-1$);
- under 1P-RS-GRM, the identifiable parameters are β_j for $j = 2, \ldots, r$ (we use the constraint $\beta_1 = 0$), whereas the cut-off points τ_x are not identified.

Therefore, in the first case $\boldsymbol{\beta}$ has dimension $(r-1)(l-1)$, whereas in the second case it has dimension $r-1$.

Function (4) may be simply maximized by a Newton-Raphson algorithm using the *pseudo-score* vector

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

and the *pseudo-observed information* matrix

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

Moreover, the asymptotic variance-covariance matrix may be obtained by the sandwich formula

$$\begin{aligned} \hat{V}_C(\hat{\boldsymbol{\beta}}_C) &= \boldsymbol{H}_C(\hat{\boldsymbol{\beta}}_C)^{-1} \boldsymbol{S}(\hat{\boldsymbol{\beta}}_C) \boldsymbol{H}_C(\hat{\boldsymbol{\beta}}_C)^{-1}, \\ \boldsymbol{S}(\boldsymbol{\beta}) &= \sum_{i=1}^n \boldsymbol{s}_{C,i}(\boldsymbol{\beta}) [\boldsymbol{s}_{C,i}(\boldsymbol{\beta})]', \end{aligned}$$

and standard errors may be extracted in the usual way from $\hat{V}_C(\hat{\boldsymbol{\beta}}_C)$.

4. Simulation study

In this section, we rely on a Monte Carlo simulation study in order to compare the performance of the estimation approaches described above. We first describe the simulation setting and, then, we illustrate the main results. All the analyses have been implemented in the R software.

4.1. Structure of the Monte Carlo study

For the baseline setup of our experiment, we assumed a model for ordinal response variables with $l = 5$ categories which is based on assumption (1) with $\beta_{jx} = \beta_j + \tau_x$, where the β_j parameters are fixed as r equidistant points in the interval $[-2, 2]$ and the cut-off points are $\tau_1 = -2$, $\tau_2 = -0.5$, $\tau_3 = 0.5$, and $\tau_4 = 2$. The latent variable Θ is assumed to have distribution $N(0, 1)$. We run 1000 Monte Carlo replications considering a number of items $r = 5, 10$ and sample size $n = 1000, 2000$.

We repeated the Monte Carlo experiment under different distributions for Θ. In particular, we assumed a standardized gamma distribution with shape and scale parameters both equal to 2, denoted by $\Gamma(2, 2)$. We then considered a symmetric discrete distribution with probabilities 0.25, 0.5, and 0.25 (LC1) and a skewed discrete distribution with probabilities 0.4, 0.5, and 0.1 (LC2). In the discrete case, the support points are such that the resulting distributions are standardized. In this way, we considered 16 different scenarios overall. As for the basic scenario, for all the drawn samples we fitted models 1P-GRM and 1P-RS-GRM by the MML, MML-LC (where k is chosen by BIC), and CML methods.

4.2. Simulation results

In the following, a summary of the main simulation results is provided: Table 1 reports the results for 1P-GRM, whereas Table 2 reports those for 1P-RS-GRM. For each scenario, we considered the estimates of the item parameters β_{jx} for 1P-GRM and β_j for 1P-RS-GRM and we computed the corresponding average values of the absolute bias and of the Root Mean Squared Error (RMSE). In this way, the performance of each estimation approach may be evaluated in terms of bias and efficiency.

According to the results in Table 1, the RMSE decreases when the sample size and/or the number of items increase, independently of the true distribution of Θ and the type of estimation method. Also the average bias tends to decrease in the presence of higher values for the sample size and the number of items when CML or MML are adopted, whereas the behavior of the MML-LC approach significantly depends on the distribution of the latent trait: the average bias of the item parameter estimates increases when data come from $N(0, 1)$ or $\Gamma(2, 2)$ and it decreases when data come from a population having a discrete distribution.

Concerning the comparison between the three estimation approaches, we note very similar performances in terms of efficiency, even if some differences can be highlighted. The MML-LC approach tends to outperform the others in all cases, except with normally distributed data and $n = 2000$, when the MML approach is more efficient. Furthermore, the MML approach is more efficient than the CML approach in the case of latent variables having normal and symmetric multinomial distributions. As far as the bias of the estimators, the MML approach provides values of absolute bias that are the best ones for the normal distribution and, on the contrary, they are the worst ones for the other distributions, especially for $\Gamma(2, 2)$. Finally, for CML and MML-LC approaches the bias is negligible.

According to the results in Table 2, referred to 1P-RS-GRM, we draw similar conclu-

Table 1. Simulation results for 1P-GRM: average values of absolute bias and RMSE for the estimates of the 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

sions. Indeed, we observe that, as n and/or r increase, RMSE decreases independently of the estimation method, whereas the average bias decreases for CML and MML, and it decreases for MML-LC only for data coming from discrete distributions.

The comparison between the three estimation methods does not allow us to conclude so clearly as in the case of 1P-GRM. On the one hand, MML-LC and MML methods are more efficient than the CML method and, on the other hand, this last one tends to be less biased than the two competitors. However, as concerns the comparison between MML and MML-LC, neither of them has a clearly better behavior.

5. Hausman test for normality of the latent trait

The results of the simulation study illustrated above are in agreement with the fact that the MML estimator is consistent only when the latent trait is normally distributed. Under this assumption, it is also more efficient than the competing estimators. On the other hand, the CML estimator is robust to misspecifications of the latent trait distribution, although it is less efficient than the MML estimator under normality. These arguments imply that a Hausman (1978) test may be used for the hypothesis of normality of the latent trait. In the following, we first illustrate this test for the class of models at issue and, then, we perform a simulation study to investigate the size and power properties of the proposed test.

Table 2. Simulation results for 1P-RS-GRM: average values of absolute bias and RMSE for the estimates of the parameters β_j.

			CML		MML		MML-LC	
Distrib.	n	r	abs.bias	RMSE	abs.bias	RMSE	abs.bias	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	0.0013	0.0872	0.0168	0.0902
$N(0,1)$	2000	10	0.0013	0.0637	0.0009	0.0623	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	0.0055	0.0883	0.0109	0.0890
$\Gamma(2,2)$	2000	10	0.0046	0.0648	0.0067	0.0615	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	0.0064	0.0896	0.0019	0.0801
LC1	2000	10	0.0011	0.0646	0.0027	0.0602	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	0.0042	0.0886	0.0032	0.0819
LC2	2000	10	0.0007	0.0662	0.0031	0.0639	0.0011	0.0638

5.1. The proposed test

The proposed Hausman (1978) test is based on the comparison between two estimators: one is consistent and more efficient under the null hypothesis, such as the MML estimator under the hypothesis of normality; the other is always consistent but less efficient under the null hypothesis, such as the CML estimator. Therefore, in this context, this is a test for the assumption of normality of the latent trait.

The test statistic is defined as

$$T = (\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_C)'\widehat{\boldsymbol{W}}^{-1}(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_C), \tag{5}$$

with $\hat{\boldsymbol{\beta}}_M$ and $\hat{\boldsymbol{\beta}}_C$ being the estimators of $\boldsymbol{\beta}$ based on the MML method and the CML method, respectively. We recall that this parameter vector has different elements according to the adopted model (see Section 3.2); furthermore, $\hat{\boldsymbol{\beta}}_M$ is the MML estimator of this parameter vector obtained from $\hat{\boldsymbol{\eta}}_M$ taking into account the identifiability constraints adopted under each model specification. Besides, $\widehat{\boldsymbol{W}}$ is the estimator of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_C$ obtained starting from the following sandwich formula:

$$\widehat{\boldsymbol{V}}\begin{pmatrix}\hat{\boldsymbol{\eta}}_M\\ \hat{\boldsymbol{\beta}}_C\end{pmatrix} = \begin{pmatrix}\boldsymbol{H}_M(\hat{\boldsymbol{\eta}}_M) & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{H}_C(\hat{\boldsymbol{\beta}}_C)\end{pmatrix}^{-1}\boldsymbol{S}\begin{pmatrix}\hat{\boldsymbol{\eta}}_M\\ \hat{\boldsymbol{\beta}}_C\end{pmatrix}\begin{pmatrix}\boldsymbol{H}_M(\hat{\boldsymbol{\eta}}_M) & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{H}_C(\hat{\boldsymbol{\beta}}_C)\end{pmatrix}^{-1},$$

$$\boldsymbol{S}\begin{pmatrix}\hat{\boldsymbol{\eta}}_M\\ \hat{\boldsymbol{\beta}}_C\end{pmatrix} = \sum_{i=1}^{n}\begin{pmatrix}\boldsymbol{s}_{M,i}(\hat{\boldsymbol{\eta}}_M)\\ \boldsymbol{s}_{C,i}(\hat{\boldsymbol{\beta}}_C)\end{pmatrix}\left([\boldsymbol{s}_{M,i}(\hat{\boldsymbol{\eta}}_M)]' \quad [\boldsymbol{s}_{C,i}(\hat{\boldsymbol{\beta}}_C)]'\right).$$

The matrix $\widehat{\boldsymbol{W}}$ is obtained by the transformation

$$\widehat{\boldsymbol{W}} = \boldsymbol{D}\widehat{\boldsymbol{V}}\boldsymbol{D}',$$

with $\boldsymbol{D} = \begin{pmatrix}\boldsymbol{E} & -\boldsymbol{I}\end{pmatrix}$, where $\boldsymbol{I}$ is identity matrix of dimension $(r-1)(l-1)$ for 1P-GRM and of dimension $l-1$ for 1P-RS-GRM, and $\boldsymbol{E}$ is a matrix such that $\hat{\boldsymbol{\beta}}_M = \boldsymbol{E}\hat{\boldsymbol{\eta}}_M$.

Under the null hypothesis of normality of the latent trait, the test statistic T has asymptotic χ^2 distribution with a number of degrees of freedom equal to $(r-1)(l-1)$ for 1P-GRM and to $r-1$ for 1P-RS-GRM. Also note that the test statistic proposed above relies on a more general expression than the original formulation of Hausman (1978) because $\widehat{\boldsymbol{W}}$ is not simply the difference between two asymptotic variance-covariance matrices.

5.2. Simulation study about the Hausman test properties

Table 3 reports the results of a Monte Carlo simulation that investigates size and power properties of the Hausman test based on statistic (5) under 1P-GRM. We used the same design described in Section 4.1.

The results in Table 3 show that the test has overall good size properties: the rejection rate approaches the nominal value of 0.05 with both sample sizes ($n = 1000, 2000$) when $r = 5$; in contrast, increasing the number of items, while holding the sample size equal, leads the test statistic to slightly under-reject the null hypothesis of normality. Nevertheless, the nominal size 0.05 always lies within the 99% confidence interval based on the corresponding empirical size.

The test also exhibits good power against the Γ distribution for the latent trait. In particular, under this distribution, the rejection rate considerably increases with the sample size, while it seems to be invariant to the number of items. The same pattern arises when the distribution of the latent trait is discrete and symmetric: in this case, however, the rejection rate slowly increases with the sample size. Power is considerably higher when the discrete distribution is asymmetric and it increases with both sample size and number of items.

Table 3. Simulation results for the Hausman test under 1P-GRM: empirical size and power.

		Empirical size at 0.05 nominal value	
Distribution	n	$r = 5$	$r = 10$
$N(0, 1)$	1000	0.058	0.027
$N(0, 1)$	2000	0.059	0.034
		Power at 0.05 nominal value	
Distribution	n	$r = 5$	$r = 10$
$\Gamma(2, 2)$	1000	0.648	0.645
$\Gamma(2, 2)$	2000	0.981	0.992
LC1	1000	0.150	0.104
LC1	2000	0.254	0.238
LC2	1000	0.178	0.216
LC2	2000	0.378	0.505

6. Application

In order to illustrate the proposed approach, we consider the Science dataset (available in `R` package `ltm`; Rizopoulos, 2006) 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 (Karlheinz and Melich, 1992). The dataset concerns the responses to $r = 7$ items with $l = 4$ ordered response categories (0=strongly disagree, 1=disagree to some extent, 2=agree to some extent, and 3=strongly agree):

- **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.

Table 4. Estimates of the item parameters from the CML and MML methods under the constraint $\beta_{j1} = 0$, $j = 1, \ldots, r$ (in brackets the standard errors based on the observed information matrix).

	CML					
	1st cut-off		2nd cut-off		3rd cut-off	
Environment	4.290	(0.558)	3.372	(0.233)	1.657	(0.273)
Work	2.238	(0.543)	1.714	(0.218)	0.755	(0.211)
Future	1.226	(0.556)	1.075	(0.205)	-0.129	(0.194)
Technology	4.255	(0.549)	3.510	(0.245)	2.282	(0.315)
Industry	4.855	(0.557)	4.411	(0.267)	3.009	(0.420)
Benefit	1.696	(0.491)	1.584	(0.208)	0.217	(0.198)
	$\hat{\ell}_C = -1549.985$					
	MML					
	1st cut-off		2nd cut-off		3rd cut-off	
Environment	3.794	(0.467)	3.387	(0.214)	1.408	(0.235)
Work	1.996	(0.489)	1.697	(0.211)	0.658	(0.197)
Future	1.043	(0.530)	1.062	(0.218)	-0.081	(0.177)
Technology	3.771	(0.467)	3.536	(0.216)	1.951	(0.277)
Industry	4.223	(0.467)	4.407	(0.235)	2.532	(0.348)
Benefit	1.503	(0.507)	1.585	(0.212)	0.203	(0.183)
	$\hat{\ell}_M = -3033.693$					

- **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.

Categories of the 2nd, 5th, and 6th items were reversed before parameter estimation in order to have all item response categories ordered in the same way. Then, we estimated 1P-GRM on the resulting data adopting the CML and MML methods. Table 4 reports the estimates of the 18 parameters β_{jx}, taking into account the required identifiability constraints, that is, under the constraint $\beta_{1x} = 0$, $x = 1, \ldots, l - 1$.

On the basis of the results in Table 4 and using the proper standardization matrix $\widehat{\boldsymbol{W}}$, we computed the Hausman test statistic which is equal to $t = 54.107$. The corresponding p-value is equal to:

$$\mathrm{P}\left(\chi^2_{18} \geq t\right) < 0.001,$$

Table 5. Estimates of the item parameters from the MML-LC method under the constraint $\beta_{j1} = 0$, $j = 1, \ldots, r$ (in brackets the standard errors computed by a parametric bootstrap method based on 199 replications).

	MML-LC					
	1st cut-off		2nd cut-off		3rd cut-off	
Environment	3.798	(0.471)	3.340	(0.227)	1.502	(0.257)
Work	2.015	(0.485)	1.714	(0.211)	0.691	(0.211)
Future	1.047	(0.524)	1.075	(0.242)	-0.083	(0.188)
Technology	3.780	(0.455)	3.483	(0.222)	2.078	(0.300)
Industry	4.222	(0.458)	4.365	(0.251)	2.639	(0.364)
Benefit	1.510	(0.522)	1.594	(0.224)	0.215	(0.198)
	$\hat{\ell}_{LC} = -3016.362$					

which leads to rejecting the hypothesis of normality.

We then estimated the model by the MML-LC method with a number of latent classes $k = 3$, which was selected on the basis BIC (see Section 3.1); this model has a maximum log-likelihood equal to $\hat{\ell}_{LC} = -3016.362$, which is higher than that obtained under the MML method (see Table 4). The estimates of the item parameters obtained by the MML-LC method are a compromise between those obtained by the MML method and the CML method (see Table 5), but they seem in general closer to the MML estimates. However, we noticed that increasing the number of support points k leads to estimates closer to those obtained with the CML method; this is in agreement with well-known results about the coincidence between CML estimates and MML-LC estimates for the Rasch model (Lindsay et al. 1991) when k is large enough.

Based on the results of the MML-LC model, we can cluster individuals in 3 latent classes, which are associated with different levels of the latent trait (Table 6). The latent distribution is standardized and results to be skewed, with skewness index equal to 0.996. Class 2 is the most representative, with almost 90% of subjects belonging to this class. Class 1 gathers the 8.4$\%$ of individuals with the lowest level of the latent trait, whereas class 3 collects the remaining 4.3$\%$ of individuals, which have the highest level of accordance along the latent trait.

Table 6. Estimates of support points and weights from the MML-LC method with $k = 3$.

c	$\hat{\xi}_c$	$\hat{\pi}_c$
1	-2.288	0.084
2	0.044	0.873
3	3.607	0.043

7. Conclusions

In this paper, we propose a method for estimating the parameters of a constrained version of Graded Response Model (GRM - Samejima, 1969; van der Ark, 2001). Under the assumption of equal discriminating power of all items, we are able to introduce a Conditional Maximum Likelihood (CML) estimation approach, implemented in a similar way as in Baetschmann et al. (2011).

The CML approach provides an estimator that is easy to implement and is non-parametric, in the sense that it does not require distributional assumptions on the latent trait since sufficient statistics for the individual effects exist. On the other hand, the Marginal Maximum Likelihood (MML) estimator (Bock and Lieberman, 1970; Bock and Aitkin, 1981), which is commonly used for GRM, relies on the normality assumption of the latent trait for consistency and requires the use of quadrature techniques to compute the log-likelihood function. An interesting alternative to the MML estimator is represented by the estimator based on the discreteness distribution of the latent trait (MML-LC - Bacci et al., 2014; Bartolucci et al., 2014). The MML-LC estimator allows us to remove the dependence on the normality assumption and it is less computationally demanding.

We compared the performance of the CML estimator with that of the two competitors, MML and MML-LC, through a Monte Carlo study. The results of this experiment confirm that the CML estimator is robust to departures from normality of the latent trait and that there is not a significant loss of efficiency compared to the MML estimator under normality.

In this context, we also propose a Hausman (1978) test to compare the CML and the MML estimators. Given that their behavior depends on the distribution of the latent trait, the Hausman test can be seen as a normality test. We computed the test statistic using data from the Consumer Protection and Perceptions of Science and Technology section of the 1992 Euro-Barometer Survey. In this case, the hypothesis of normality is rejected and we find that the semi-parametric MML-LC method, with a suitable number of latent classes, is an interesting alternative to MML.

To conclude, we outline that, in the longitudinal setting, a recent work by Skrondal and Rabe-Hesketh (2013) has proved the robustness of the CML estimator in comparison to the MML one in presence of data missing not at random. For a further developments of our work, it is worth to study this aspect with reference to proposed CML estimator in the item response setting.

Acknowledgements: Authors acknowledge the financial support from the grant “Finite mixture and latent variable models for causal inference and analysis of socio-economic data” (FIRB - Futuro in ricerca) funded by the Italian Government (RBFR12SHVV).

References

Abramowitz, M., Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, New York: Dover.

Agresti, A. (2002). *Categorical data analysis*, New York: Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, in: B. Petrov, F. Csaki (eds.): *Second International Symposium on Information Theory*, Budapest: Akademinai Kiado, pp. 267–281.

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

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations, *Journal of the Royal Statistical Society, Series B*, **34**, 42–54.

Andrich, D. (1978). A rating formulation for ordered response categories, *Psychometrika*, **43**, 561–573.

Bacci S., Bartolucci F., Gnaldi M. (2014). A class of multidimensional IRT models for ordinal polytomous item responses, *Communications in Statistics - Theory and Methods*, 43, 787–800.

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

Bartolucci, F., Bacci, S. and Gnaldi, M. (2014). MultiLCIRT: An R package for multidimensional latent class item response models, *Computational Statistics & Data Analysis*, **71**, 971-985.

Bock, R. D., Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm, *Psychometrika*, **46**, 443–459.

Bock, R. D., Lieberman, M. (1970). Fitting a response model for n dichotomously scored items, *Psychometrika*, **35**, 179–197.

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

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.

Hausman, J. (1978). Specification Tests in Econometrics, *Econometrica*, **46**,1251–1271.

Karlheinz, R., Melich, A. (1992). *Euro-Barometer 38.1: Consumer Protection and Perceptions of Science and Technology*, Brussels: INRA (Europe).

Lindsay, B., Clogg, 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.

Masters, G. (1982). A Rasch model for partial credit scoring, *Psychometrika*, **47**, 149–174.

McLachlan, G., Peel, D. (2000). *Finite mixture models*. NewYork: Wiley.

Nering, M. L., Ostini, R. (2010). *Handbook of polytomous item response theory models*, New York: Taylor and Francis.

Pinheiro, J. C., Bates, D. M. (1995). Approximation to the log-likelihood function in the nonlinear mixed effects models, *Journal of Computational and Graphical Statistics*, **4**, 12–35.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology, *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, 321–333.

Rizopoulos, D. (2006). `ltm`: An `R` package for latent variable modeling and item response analysis, *Journal of Statistical Software*, **17**, 1–25.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores, *Psychometrika Monograph*, **17**.

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

Skrondal, A., Rabe-Hesketh, S. (2013). Protective estimation of mixed-effects logistic regression when data are not missing at random, *Biometrika*, doi: 10.1093/biomet/ast054.

van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models, *Applied Psychological Measurement*, **25**, 273–282.

van der Linden, W., Hambleton, R. K. (1997). *Handbook of modern item response theory*, Springer.

QdS – Vol. 15, 2013

Nested continuation logit models for ordinal variables

Roberto Colombi
Department of Engineering, University of Bergamo
E-mail: colombi@unibg.it

Sabrina Giordano
Department of Economics, Statistics and Finance, University of Calabria
E-mail:sabrina.giordano@unical.it

Abstract We introduce a new type of logits: the nested continuation logits, which are a generalization of the well-known continuation logits and we show that these logits reflect a sequential process that may be used by an individual to select a category of an ordinal variable. Logit models based on the nested continuation logits are also discussed to take into account the effect of categorical covariates.

Keywords: Contingency tables; Logit models; Continuation logits; Marginal models.

1. Introduction

In many applications, respondents are asked to select among a list of ordered categories to express their personal level of agreement or opinion about attitudes, lifestyles, services, items etc., and the choice among the ordered alternatives may stem from quite different mechanisms. We assume a selection process such that every respondent makes a decision by discarding consecutive categories, starting from the first in the list, until the category that reflects his/her status is reached.

Let A denote an ordinal variable with categories in $\mathcal{A} = \{a_1, a_2, ..., a_J\}$.

At the j-th step, $j = 1, 2, ..., J$, given that categories $a_1, a_2, ..., a_{j-1}$ have been discarded, the sequential selection process stops by selecting a_j with probability p_j ($p_J = 1$) or continues towards higher categories.

Thus, the probability of selecting a_j is given by $\pi_j = (1-p_1)(1-p_2).....(1-p_{j-1})p_j$

and, in terms of the continuation logits (Agresti, 2010) $\eta_j = log\frac{p_j}{1-p_j}$, is

$$\pi_j = \frac{\exp(\eta_j)}{\prod_{l=1}^{j}(1+\exp\{\eta_l\})}, \quad j=1,2,...,J-1, \quad \pi_J = \frac{1}{\prod_{l=1}^{J-1}(1+\exp\{\eta_l\})}. \quad (1)$$

In the special case of $p_j = p$, i.e. when the odds of attaining level a_j rather than a higher level are equal for all j, the probabilities of the categories become: $\pi_j = (1-p)^{j-1}p$, $j = 1,2,...,J-1$, and $\pi_J = (1-p)^{J-1}$, that are the probabilities of a censored Geometric (or Pascal) variable with parameter p.

The sequential mechanism of selection seems the most natural as it respects the way of reading from left to right commonly used. It is particularly motivated when the selecting mechanism is naturally sequential, that is when the categories can be reached only successively, for example, the level of learning reaches the advanced state only after the intermediate state. Otherwise, the sequential selection procedure can be forced by the design of the questionnaire, for example, when a customer is first invited to manifest his/her dissatisfaction or satisfaction, and successively, if satisfied, to express his/her (low, high) degree of satisfaction.

This paper extends the use of the mentioned choice method to a multistage selection process. In certain cases, in fact, the final choice can be made via subsequent stages where, at each stage, the selection moves through sequential discards from left to right among the alternatives. More specifically, the first stage begins with the sequential choice of an interval of categories among those that partition $\mathcal{A}$, then at each consecutive stage, an interval of categories is sequentially chosen among those that form a partition of the interval selected at the previous stage. This selection process iterates until one category is chosen.

Consider, for example, a respondent who is asked to declare his level of job satisfaction on a scale of ten ordered categories: *extremely dissatisfied* (ED), *very dissatisfied* (VD), *dissatisfied* (D), *moderately dissatisfied* (MD), *a little dissatisfied* (LD), *a little satisfied* (LS), *moderately satisfied* (MS), *satisfied* (S), *very satisfied* (VS), *extremely satisfied* (ES). Suppose he is happy with his job. At the first sight, he can restrict his selection to the last five categories representing the positive opinion, by refusing the first five categories on the negative side. At the second step, he can realize to be *a little satisfied*, for example, concluding the selection, or alternatively he can discard the first two mild positions: *a little satisfied*, *moderately satisfied* and orient his choice to the set of categories expressing a stronger feeling: *satisfied*, *very satisfied*, *extremely satisfied*, and finally he can decide to select the category *satisfied* which better represents his status.

We show that this multistage choice method corresponds to a particular type of logits, the nested continuation logits, which will be the focus of the next section. A two stage approach related to that proposed in this paper has been discussed by Fahrmeir and Tutz (2001, ch. 3) and by Tutz (2013, ch. 9), among others.

Other kinds of nested models have been described by McCullagh and Nelder (1989) to emphasize the nested or hierarchical structure of the response categories when these can be considered also as categories of different response variables at successive levels of a hierarchy.

We will consider also the possibility that covariates can affect the selection process at certain stages only. For example, the probability of being satisfied about own job can vary if the respondent is male or female, but when the set of categories describing the positive opinion is selected at the first stage, the probability to declare successively an *extreme satisfaction* instead of a *simple satisfaction* can be independent of the gender of the respondent. This further possibility strengthens the relevance of taking into account the sequence followed by an interviewee in responding to an item.

The rest of the paper is organized as follows. Section 2 introduces the nested continuation logits and Section 3 describes how to model the effects of qualitative covariates on the proposed logits. Section 4 is devoted to the estimation of parameters and an example is illustrated in Section 5.

2. *Nested continuation logits*

When A is an ordinal variable with categories in $\mathcal{A} = \{a_1, a_2, ..., a_J\}$ and $\pi_j = P(A = a_j)$, $j = 1, 2, ..., J$, the probability of a subset I of categories is denoted by $P(I) = \sum_{j \in I} \pi_j$.

A family $\mathcal{I}$ of sets of contiguous categories $I_{i,k} = \{a_i, a_{i+1}, ..., a_k\}$, $i < k$, associated to the ordinal variable A, is called hierarchical if $\mathcal{A} \in \mathcal{I}$ and contains intervals which are nested or disjoint, i.e. for every pair of sets $I_{i,k}, I_{l,m}$ belonging to $\mathcal{I}$, it holds that $I_{i,k} \cap I_{l,m} \in \{\emptyset, I_{i,k}, I_{m,l}\}$. Furthermore, $\mathcal{S} = \{\{a_j\} : j = 1, 2, ..., J\}$ denotes the family of singletons. Let $\mathcal{P}_{i,k} = \{S_h^{i,k}, h = 1, 2, ..., l_{i,k}\}$ be the family of the sets belonging to $\mathcal{I} \cup \mathcal{S}$ that are maximal subsets of $I_{i,k}$. The sets in $\mathcal{P}_{i,k}$ constitute a partition of $I_{i,k}$. Moreover, the sets $S_h^{i,k}, h = 1, 2, ..., l_{i,k}$ are ordered so that for $m > n$ the categories in $S_n^{i,k}$ precede those in the set $S_m^{i,k}$.

In the previous example of job satisfaction, $\mathcal{A} = I_{1,10} = \{ED, VD, D, MD, LD, LS, MS, S, VS, ES\}$ contains 10 categories. The set $\mathcal{A}$ can be partitioned in $I_{1,5} = \{ED, VD, D, MD, LD\}$ and $I_{6,10} = \{LS, MS, S, VS, ES\}$. Then, $I_{1,5}$ can be partitioned in $I_{1,4} = \{ED, VD, D, MD\}$ and $\{LD\}$; while $I_{6,10}$ in $\{LS\}$; $\{MS\}$; $I_{8,10} = \{S, VS, ES\}$. So, in this example, the family $\mathcal{I}$ is defined as $\mathcal{I} = \{I_{1,10}, I_{1,5}, I_{6,10}, I_{1,4}, I_{7,10}\}$.

The family $\mathcal{I}$ can be described by a tree where $\mathcal{A}$ is the root node, every interval $I_{i,k}$ is an intermediate node, the elements of $\mathcal{P}_{i,k}$ are the children of the node $I_{i,k}$ and the terminal nodes are the singletons belonging to $\mathcal{S}$. Therefore, the selection process can be seen as a path that leads to a terminal node of a tree starting from the root node. For the example of job satisfaction, the tree is displayed in Figure 1.

The definition of the hierarchical family $\mathcal{I}$ implies that for every category $a_j \in \mathcal{A}$, a sequence of sets I_{i_l,k_l}, $l = 1, 2, ..., r_j + 1$, with $r_j \geq 0$, of $\mathcal{I}$ exists such that

$$\{a_j\} \subset I_{i_1,k_1} \subset I_{i_2,k_2} \subset I_{i_3,k_3} \subset .... \subset I_{i_{r_j+1},k_{r_j+1}} = \mathcal{A}, \tag{2}$$

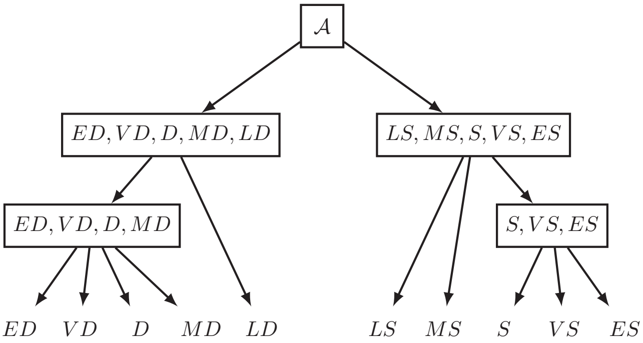


Figure 1. A tree representing a respondent's multistage selection process to express own opinion on the job satisfaction

and, moreover, I_{i_l,k_l} is the maximal subset of $I_{i_{l+1},k_{l+1}}$ containing a_j and belonging to $\mathcal{I}$. As a consequence of (2), it is immediate to get the probability

$$\pi_j = P(I_{i_{r_j},k_{r_j}}|\mathcal{A})P(I_{i_{r_j-1},k_{r_j-1}}|I_{i_{r_j},k_{r_j}})....P(I_{i_1,k_1}|I_{i_2,k_2})P(\{a_j\}|I_{i_1,k_1}), \quad (3)$$

$j = 1, ..., J$.

The hierarchy (2) corresponds to a path in a tree, as that illustrated in Figure 1, and there is a probability (3) associated to every path in a tree.

The factorization in (3) reflects a multistage process leading to the choice of a category on an ordered scale. The sequential mechanism selects: at the first stage, $I_{i_{r_j},k_{r_j}}$ among the sets in $\mathcal{I} \cup \mathcal{S}$ that partition $\mathcal{A}$; at the second stage, $I_{i_{r_j-1},k_{r_j-1}}$ among the sets of $\mathcal{P}_{i_{r_j},k_{r_j}}$, and so on. Finally, the last stage ends with the choice of a category a_j from the minimal set I_{i_1,k_1}. Note that the multistage selection process simplifies into a two stage approach if the subsets of $\mathcal{A}$, belonging to the hierarchical family $\mathcal{I}$, are disjoint. In this case, if at the first stage one interval is selected, at the second stage a category is chosen.

If the sequential mechanism described in Section 1 applies to the sets $S_h^{i,k}$, $S_h^{i,k} \in \mathcal{P}_{i,k}$, the logits associated to each set $I_{i,k} \in \mathcal{I}$ are the nested continuation logits so defined

$$\eta(i,k;h) = \log \frac{P(S_h^{i,k})}{P(\cup_{l=h+1}^{l_{i,k}} S_l^{i,k})}, \quad h = 1, 2, ..., l_{i,k} - 1, \quad (4)$$

from which it is easy to get the probabilities

$$P(S_h^{i,k}|I_{i,k}) = \frac{\exp\{\eta(i,k;h)\}}{\prod_{l=1}^{h}(1+\exp\{\eta(i,k;l)\})}, \quad h = 1, 2, ..., l_{i,k} - 1. \quad (5)$$

In particular, when $I_{i,k} = \{a_i, a_{i+1}, ..., a_k\}$ is a minimal set of $\mathcal{I}$, the family $\mathcal{P}_{i,k}$ involves only the singletons $\{a_j\}, j = i, i+1, ..., k$ and equations (4) and (5) simplify to

$$\eta(i,k;h) = \log \frac{\pi_{i-1+h}}{\sum_{l=h}^{k-i} \pi_{i+l}}, \quad h = 1,2,...,k-i,$$

and

$$P(\{a_{i-1+h}\}|I_{i,k}) = \frac{\exp\{\eta(i,k;h)\}}{\prod_{l=1}^{h}(1+\exp\{\eta(i,k;l)\})}, \quad h = 1,2,...,k-i.$$

A general expression of (3) in terms of nested continuation logits (4) is described in the Proposition below where the boolean function $\delta(\cdot)$ is equal to 1 when the argument is true and 0 otherwise.

Proposition 1. *The probabilities* $\pi_j, j = 1,2,...,J$, *can be computed from the nested continuation logits* $\eta(i,k;h)$, $I_{i,k} \in \mathcal{I}$, $h = 1,2,...,l_{i,k}-1$ *described in (4) as follows*

$$\pi_j = \prod_{I_{i,k} \in \mathcal{I}} \frac{\prod_{h=1}^{l_{i,k}-1} \exp\{\delta(a_j \in S_h^{i,k})\eta(i,k;h)\}}{\prod_{h=1}^{l_{i,k}-1}(1+\exp\{\eta(i,k;h)\})^{\delta(a_j \in \cup_{l=h}^{l_{i,k}} S_l^{i,k})}}. \tag{6}$$

Proof. A factor in the outer product (6) is equal to one if $a_j \notin I_{i,k}$ so the product can be restricted to the sets of $\mathcal{I}$ that contain a_j and can be ordered as shown in (2). For each of these sets, the probability $P(I_{i_{r_j-1},k_{r_j-1}}|I_{i_{r_j},k_{r_j}})$ is calculated according to (5) and the Proposition follows by (3).

The logits so far described are not limited to the selection scheme that starts from lower category or intervals in an ordered list and proceeds toward higher categories or intervals. For example, when the sets, belonging to the hierarchical family $\mathcal{I}$, are $I_{1,j} = \{a_1, a_2, ..., a_j\}$, $j = 2,3,...,J$, the underlying multistage selection mechanism leads to discard categories sequentially starting from the last one in the ordered scale. This shows that the reverse continuation logits can be seen as a special case of the nested continuation logits (4).

Moreover, by using the sets of categories below and above the median, the family $\mathcal{I}$ involving the following intervals can be defined

a) the interval $\mathcal{A}$,

b) the intervals on the left of the median: $I_{1,j} = \{a_1, a_2..., a_j\}$, $j = 2,3,...,m$, where $m = \frac{J}{2}$ if J is even and $m = \frac{J-1}{2}$ if J is odd,

c) the interval on the right of the median: $I_{m,J} = \{a_m, a_{m+1}..., a_J\}$, where $m = \frac{J+2}{2}$ if J is even and $m = \frac{J+3}{2}$ if J is odd.

According to this family, the sets of categories on the left or on the right of the median are selected at the first stage (if J is odd the category $a_{\frac{J+1}{2}}$ is selected or discarded at this stage). Then, if the categories below the median are chosen, the selection proceeds from the highest category to the lowest, otherwise it proceeds in the inverse order from the lowest to the highest one.

The next example illustrates how to construct the logits following the scheme of selection described.

Example 1. *With respect to the variable Political Orientation with 7 categories: extremely liberal (EL), liberal (L), slightly liberal (SL), moderate (M), slightly conservative (SC), conservative (C), extremely conservative (EC), we model the behavior of respondents who, when the median category moderate is discarded, continue the selection by moving on the left or on the right as described above. By this scheme, the sets in* $\mathcal{I}$ *are* $I_{1,7} = \mathcal{A} = \{EL, L, SL, M, SC, C, EC\}$, $I_{1,3} = \{EL, L, SL\}$, $I_{1,2} = \{EL, L\}$, *and* $I_{5,7} = \{SC, C, EC\}$.

The set $\mathcal{P}_{1,7}$ *contains* $S_1^{1,7} = I_{1,3}$, $S_2^{1,7} = \{M\}$ *and* $S_3^{1,7} = I_{5,7}$. *In this case, the nested continuation logits associated to the set* $I_{1,7}$ *are*

$$\eta(1,7;1) = \log \frac{\pi_1 + \pi_2 + \pi_3}{\pi_4 + \pi_5 + \pi_6 + \pi_7}, \quad \eta(1,7;2) = \log \frac{\pi_4}{\pi_5 + \pi_6 + \pi_7}. \tag{7}$$

The logits corresponding to the sets $I_{1,3}$, $I_{1,2}$ *are of reverse continuation type*

$$\eta(1,3;1) = \log \frac{\pi_3}{\pi_1 + \pi_2}, \quad \eta(1,2;1) = \log \frac{\pi_2}{\pi_1}, \tag{8}$$

whereas the nested continuation logits for the conservative set $I_{5,7}$ *are*

$$\eta(5,7;1) = \log \frac{\pi_5}{\pi_6 + \pi_7}, \quad \eta(5,7;2) = \log \frac{\pi_6}{\pi_7}. \tag{9}$$

3. Logit models

Let X_j, $j = 1, 2, ..., q$, be a set of categorical covariates. The categories of X_j are denoted by $x_{ji} : i = 1, 2, ..., s_j$, $j = 1, 2, ..., q$.

A configuration $x_{1i_1}, x_{2i_2}, ..., x_{qi_q}$ of the covariates is denoted by the vector $\boldsymbol{i} = (i_1, i_2, ..., i_q)'$. If $\mathcal{M} \subset \mathcal{V} = \{1, 2, ..., q\}$ then $\boldsymbol{i}_{\mathcal{M}}$ denotes the vector with components $i_j : j \in \mathcal{M}$. If $\boldsymbol{i}_{\mathcal{M}\cup\mathcal{N}}$ is a vector such that $\boldsymbol{i}_{\mathcal{M}} = \boldsymbol{h}_{\mathcal{M}}$, $\boldsymbol{i}_{\mathcal{N}} = \boldsymbol{k}_{\mathcal{N}}$, with disjoint sets $\mathcal{M}, \mathcal{N}$, we also write $\boldsymbol{i}_{\mathcal{M}\cup\mathcal{N}} = (\boldsymbol{h}_{\mathcal{M}}, \boldsymbol{k}_{\mathcal{N}})$. Every X_j has a baseline category, usually the first one, indicated as i_j^*. Thus, any configuration which includes categories x_{ji} for $j \notin S$, $S \subset \mathcal{V}$, at the baseline value is denoted by $(\boldsymbol{i}_S, \boldsymbol{i}^*_{\mathcal{V}\setminus S})$.

A nested continuation logit computed in the distribution of A conditioned on the configuration $\boldsymbol{i}$ of the covariates is denoted by $\eta(i, k; h|\boldsymbol{i})$.

In order to model the dependence of the probabilities on the conditioning categories $\boldsymbol{i}$, we adopt the usual factorial expansion

$$\eta(i, k; h|\boldsymbol{i}) = \sum_{Q \subseteq \mathcal{V}} \theta^Q(i, k; h|\boldsymbol{i}_Q). \tag{10}$$

The Möbius inversion theorem (Lauritzen, 1996) ensures that

$$\theta^{Q}(i,k;h|\boldsymbol{i}_Q) = \sum_{\mathcal{H}\subseteq Q} (-1)^{|Q\setminus\mathcal{H}|}\eta(i,k;h|\boldsymbol{i}_{\mathcal{H}},\boldsymbol{i}^{*}_{\mathcal{V}\setminus\mathcal{H}}). \tag{11}$$

A useful restriction that considerably simplifies the model is the hypothesis of *additivity* of the effects of the explanatory variables. This additive dependence allows the nested continuation logits to be expressed by a sum of main effects

$$\eta(i,k;h|\boldsymbol{i}) = \theta^{\emptyset}(i,k;h) + \sum_{j=1}^{q}\theta^{j}(i,k;h|i_j), \tag{12}$$

where $\theta^{\emptyset}(i,k;h) = \eta(i,k;h|\boldsymbol{i}^{*})$ and $\theta^{j}(i,k;h|i_j) = \eta(i,k;h|i_1,i_2,...,i^{*}_j,...,i_q) - \eta(i,k;h|\boldsymbol{i}^{*})$ are nested continuation baseline log-odds ratios according with the terminology adopted by Cazzaro and Colombi (2013).

Other interesting hypotheses concern the possibility that some or all the covariates influence the conditional probabilities in the product (3) only from (up to) a certain point in the chain. This means that the covariates can affect the choice in the sequential process of selection from the beginning until a stage or starting from a certain stage up to the final choice.

More specifically, two kinds of independence of the response from covariates will be considered.

Let $\mathcal{X}_{\mathcal{C}} = \{X_j,\ j\in\mathcal{C}\}$, $\mathcal{C}\subseteq\mathcal{V}$, be a subset of covariates and $\bar{\mathcal{C}}$ be the family of the non empty proper and improper subsets of $\mathcal{V}\setminus\mathcal{C}$.

Given an interval of categories $I_{i,k}\in\mathcal{I}$, let $\mathcal{I}^{-}_{i,k}$ be the subfamily of $\mathcal{I}$ containing the sets $I_{m,n}$ such that $I_{m,n}\subseteq I_{i,k}$. The constraints

$$\theta^{Q}(m,n;h|\boldsymbol{i}_Q) = 0, \quad I_{m,n}\in\mathcal{I}^{-}_{i,k}, \quad Q\notin\bar{\mathcal{C}}, \tag{13}$$

state that the covariates in $\mathcal{X}_C$ do not affect the probabilities of singletons and intervals $I_{m,n}$ that are subsets of $I_{i,k}$.

In the contingency tables where A assumes the categories involved in $I_{i,k}$ only, this hypothesis corresponds to the independence of the response from the covariates in $\mathcal{X}_{\mathcal{C}}$, conditionally on the remaining covariates. For instance, the covariate *Religion* (*Catholics*, *not Catholics*) may affect the *Opinion on teenage birth control* (*strongly agree*, *agree*, *disagree*, *strongly disagree*) in the sense that probabilities of being in agreement {*strongly agree*, *agree*} or to *disagree*, or to *strongly disagree* with the teenage birth control can vary if the respondent is *Catholic* or not, but given that the respondent is in agreement, the probabilities of having a strong or mild position do not depend on whether the respondent is *Catholic* or not.

Another kind of independence considers the family $\mathcal{I}^{+}_{i,k}$, $\mathcal{I}^{+}_{i,k}\subset\mathcal{I}$, of intervals $I_{m,n}$ that are not subsets of a given interval $I_{i,k}$, i.e. $I_{m,n}\nsubseteq I_{i,k}$.

According to the constraints

$$\theta^{Q}(m,n;h|\boldsymbol{i}_Q) = 0, \quad I_{m,n}\in\mathcal{I}^{+}_{i,k}, \quad Q\notin\bar{\mathcal{C}}, \tag{14}$$

the covariates in $\mathcal{X}_{\mathcal{C}}$ influence the choice inside the interval $I_{i,k}$, but not outside. So that, in the subtable where the categories belonging to $I_{i,k}$ are collapsed in one category, A is independent of the covariates in $\mathcal{X}_{\mathcal{C}}$, given the remaining covariates.

For example, the choice among the *liberal* $(\{EL, L, SL\})$, *moderate* (M), *slightly conservative* (SC), *conservative*, (C), and *extremely conservative* (EC) political attitudes may not depend on the *Opinion on teenage birth control* (*strongly agree, agree, disagree, strongly disagree*), but when *liberal* side is chosen, the probabilities of belonging to one of the 3 groups, from *slightly* (SL) to *extremely liberal* (EL) people, may vary according to the degree of agreement on birth control.

The previous hypothesis can be extended to two disjoint intervals I_{i_1,k_1} and I_{i_2,k_2}, as follows

$$\theta^Q(m, n; h|\boldsymbol{i}_Q) = 0, \quad I_{m,n} \in \mathcal{I}^+_{i_1,k_1} \cap \mathcal{I}^+_{i_2,k_2}, \quad Q \notin \bar{\mathcal{C}}. \tag{15}$$

In the subtable where the categories belonging to I_{i_1,k_1} are collapsed into one category and the categories of I_{i_2,k_2} are collapsed into one category, the previous constraints correspond to the independence of the response from the covariates in the set $\mathcal{X}_{\mathcal{C}}$, given the remaining covariates. The extension to more than two disjoint intervals of categories is straightforward.

Similar hypotheses formulated on variables with partially ordered categories have been also examined by Cazzaro and Colombi (2013), who introduced a parameterization of multi-way contingency tables based on nested baseline logits and higher order marginal interactions defined with respect to families of category sets with the same structure of $\mathcal{I}$.

The linear model (10) can be obviously extended to include continuous covariates and, as in the case of categorical covariates, they can affect the response inside and/or outside a given interval $I_{i,k}$.

4. Maximum likelihood estimation

The vector $\boldsymbol{\eta}$ of the nested continuation logits can be written as follows

$$\boldsymbol{\eta} = \boldsymbol{C} \log \boldsymbol{M}\boldsymbol{\pi} \tag{16}$$

coherently with Colombi and Forcina (2001), Bartolucci et al. (2007).

The rows of the matrix $\boldsymbol{C}$ are linearly independent contrasts and the matrix $\boldsymbol{M}$, with elements equal to 0 or 1, depends on the nested continuation logits that are used.

Let $\boldsymbol{\theta}$ be the vector of the factorial effects introduced in (10). The hypotheses illustrated in the previous section, that constrain the nested continuation logits, can be described as: $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\theta}$. The design matrix $\boldsymbol{X}$ depends on the interaction effects that are not set to zero.

Under the product-multinomial sampling, this model defined by linear constraints on the nested continuation logits, is a special case of the HLP model described by Lang

(2005) and the asymptotic results and the testing procedures of Lang (2004, 2005) apply also to this context (see also Cazzaro and Colombi, 2009).

With respect to the Lang's work we note that the link function $\boldsymbol{\eta} = \boldsymbol{C} \log \boldsymbol{M\pi}$ is invertible.

The parameters vector $\boldsymbol{\theta}$ can be estimated using the algorithm described by Colombi and Forcina (2001) and Lang (2004, 2005). Models specified by linear constraints on nested continuation logits can be easily estimated and tested by using the R-package *hmmm* by Colombi et al. (2012).

The method implemented in *hmmm* is based on the *constrained formulation*

$$\boldsymbol{UC} \log \boldsymbol{M\pi} = \boldsymbol{0} \quad \text{with} \quad \boldsymbol{UX} = \boldsymbol{0} \tag{17}$$

and has the advantage of being applicable to a large variety of marginal models even when an explicit expression of the probabilities in terms of marginal parameters is not available. For example, a multi-way contingency table of ordered variables can be parameterized by the Gloneck-McCullagh marginal parameters (Gloneck and McCullagh, 1995) where the logits defined in the univariate distributions are of the nested continuation type. However, in the univariate case, the expression of the probabilities in terms of nested continuation interactions (6) allows also an explicit parametrization of the probabilities in function of the non null interactions of $\boldsymbol{\theta}$ and the resort to an unconstrained maximization of the log-likelihood function. This approach is necessary when continuous covariates are involved in the linear predictor (10). In such a case, in fact, the number of restrictions on the parameters in the constrained formulation (17) would be extremely large.

5. Example

Consider the three-dimensional table reported in Bergsma et al. (2009) p. 30, also available in the R-package *hmmm*, where a sample of 911 U.S. citizens is classified according to their Political orientation *Pol*, Religion *Rel* and Opinion on teenage birth control *Birth*. In particular, the variable *Pol*, described in Example 1, has 7 categories from *extremely liberal* to *extremely conservative*, *Rel* has 3 categories *Catholics*, *Protestants* and *None*, *Birth* 4 categories ranging from *strongly agree* to *strongly disagree*.

The main question is to investigate whether the covariates *Birth* and *Rel* condition the respondents' political opinion, in the first choice among the three political positions *liberal* $= \{EL, L, SL\}$, *moderate* $= \{M\}$, *conservative* $= \{SC, C, EC\}$, and/or in the second step in declaring the intensity of their own orientation from a simple to an extreme belief. In this regard, to study the effects of the covariates *Birth* and *Rel*, separately on the position and on the intensity of the response *Pol*, it is convenient to use the nested continuation logits (7, 8, 9) and hypotheses corresponding to the constraints (13, 14, 15) will be tested.

Table 1 illustrates the tested hypotheses. Every row refers to one hypothesis, and for each hypothesis the columns report: the nested continuation *logits* (first column)

Table 1. Hypotheses testing results: the logits hypothesized to be independent of the covariates, likelihood ratio statistic test, degrees of freedom and p-values

Logits	*Covariates*	*LRT*	*df*	*p-value*
$\eta(1,7;1), \eta(1,7;2)$	*Birth*	17.6883	18	0.4763
	Rel	254.4908	16	0.0000
	Birth, Rel	261.5026	22	0.0000
$\eta(1,2;1), \eta(1,3;1)$	*Birth*	34.7838	18	0.0101
	Rel	267.8825	16	0.0000
	Birth, Rel	298.4026	22	0.0000
$\eta(5,7;1), \eta(5,7;2)$	*Birth*	25.1237	18	0.1252
	Rel	139.3793	16	0.0000
	Birth, Rel	150.4783	22	0.0000
$\eta(1,7;1), \eta(1,7;2), \eta(1,2;1), \eta(1,3;1)$	*Birth*	52.4722	36	0.0374
	Rel	522.3733	32	0.0000
	Birth, Rel	559.9052	44	0.0000
$\eta(1,7;1), \eta(1,7;2), \eta(5,7;1), \eta(5,7;2)$	*Birth*	42.8121	36	0.2019
	Rel	393.8701	32	0.0000
	Birth, Rel	411.9808	44	0.0000

hypothesized to be independent of the *covariates* indicated in the second column, the value of the likelihood ratio statistic test (*LRT*), the degrees of freedom (*df*) and the *p-value*.

Let us start by testing if respondents at first declare their political belief choosing one of the three sets: *liberal* $= \{EL, L, SL\}$, *moderate* $= \{M\}$, *conservative* $= \{SC, C, EC\}$, independently of their opinion on the teenage *Birth* control.

Constraints of type (15) are imposed under this hypothesis. Null values are assigned to the interactions of the factorial expansion (10) of the logits associated to sets belonging to $\mathcal{I}^+_{1,3} \cap \mathcal{I}^+_{5,7}$, that in this case are the nested continuation logits (7) defined on the set $I_{1,7}$. These interactions are: the 6 main effects, i.e. nested continuation-baseline log-odds ratios

$$\theta^1(1,7;1|i_1,i_2^*) \quad = \quad \eta(1,7;1|i_1,i_2^*) - \eta(1,7;1|i_1^*,i_2^*) \tag{18}$$

$$\theta^1(1,7;2|i_1,i_2^*) \quad = \quad \eta(1,7;2|i_1,i_2^*) - \eta(1,7;2|i_1^*,i_2^*) \tag{19}$$

for $i_1 = 2,3,4$, $i_1^* = 1$, $i_2^* = 1$, and the 12 interactions of second order which are contrasts of the logits (7)

$$\theta^{12}(1,7;1|i_1,i_2) = \eta(1,7;1|i_1,i_2) - \eta(1,7;1|i_1,i_2^*) - \eta(1,7;1|i_1^*,i_2) + \eta(1,7;1|i_1^*,i_2^*) \tag{20}$$

$$\theta^{12}(1,7;2|i_1,i_2) = \eta(1,7;2|i_1,i_2)-\eta(1,7;2|i_1,i_2^*)-\eta(1,7;2|i_1^*,i_2)+\eta(1,7;2|i_1^*,i_2^*) \tag{21}$$

for $i_1 = 2,3,4$, $i_1^* = 1$, and $i_2 = 2,3$, $i_2^* = 1$. Such interactions are defined according to (11).

The above constraints correspond to the independence of the response from the covariate *Birth*, given the Religion, in the table where the 7 categories of *Pol* are collapsed into 3 categories: *liberal* $= \{EL, L, SL\}$, *moderate* $= \{M\}$, *conservative* $= \{SC, C, EC\}$.

This hypothesis cannot be rejected as $LRT = 17.6883$, $p - value = 0.4763$, *df* $=$ 18 (row 1 in Table 1).

We continue by considering if, given the initial selection of the *conservative* set $\{SC, C, EC\}$, the successive choice among extreme, normal or slightly *conservative* ideology does not depend on the covariate *Birth*.

Such hypothesis holds if and only if the nested continuation logits (9) associated to the only interval $I_{5,7}$ of $\mathcal{I}_{5,7}^-$ do not depend on *Birth*. The parameters of the factorial expansion (10) that must be null are: the main effects $\theta^1(5,7;1|i_1,i_2^*)$, $\theta^1(5,7;2|i_1,i_2^*)$ for $i_1 = 2,3,4$, $i_2^* = 1$, which are 6 log-odds ratios of nested continuation-baseline type; the 12 interactions $\theta^{12}(5,7;1|i_1,i_2), \theta^{12}(5,7;2|i_1,i_2)$ for $i_1 = 2,3,4$ and $i_2 = 2,3$ which are contrasts of nested continuation logits (9). These interactions are defined as in equations (18, 19) and (20, 21).

The mentioned constraints are of type (13), and imply independence of *Pol* from *Birth*, conditionally on *Rel*, in the table where the response assumes only categories SC, C, EC.

Testing this hypothesis yields $LRT = 25.1237$, $p - value = 0.1215$, *df* $= 18$ (row 7 in Table 1), so that it cannot be rejected.

A similar hypothesis formulated for *liberal* citizens is instead rejected, $LRT = 34.7838$, $p - value = 0.01006$, *df* $= 18$ (row 4 in Table 1).

The intersection of the discussed hypotheses of rows 1 and 7 holds if and only if the nested continuation logits associated to the intervals in $\mathcal{I}_{1,3}^+$ are not affected by the covariate *Birth*. This imposes null values on the 36 interactions described above.

The constraints associated to this hypothesis are of the type (14), implying independence of the response *Pol* from the covariate *Birth*, given the variable *Rel*, in the subtable where the categories EL, L, SL are collapsed into one category.

Results in row 13 of Table 1 shows that this hypothesis is not rejected.

We have also tested the analogous hypotheses illustrated so far but with the role of *Birth* and *Rel* inverted, and the results in Table 1 confirm that these hypotheses are strongly rejected. It seems that the different attitudes of *Catholics*, *Protestants* and *non-religious* people actually affect either the initial choice among the three ideologies (first stage), and the intensity of the political orientation (second stage).

Finally, we further investigate whether, under the joint hypotheses of no role of *Birth* in the choice at the first stage (row 1) and for the intensity of the *conservative* attitude at the second stage (row 7), there is an additive effect of both the covariates for *liberal* people who indicate the strength of their political opinion. This additivity assumption comports the nullity of further 12 parameters $\theta^{12}(1,2;1|i_1,i_2)$ and $\theta^{12}(1,3;1|i_1,i_2)$, for $i_1 = 2,3,4$ and $i_2 = 2,3$ as well as the 36 interactions set at zero for the hypotheses of rows 1 and 7, but this additional restriction is rejected ($LRT = 68.2020$, $df = 48$, $p - value = 0.0291$).

To sum up, with regard to the main question on whether *Birth* and *Rel* affect the position and the intensity of the political opinion of the respondents, we summarize the main findings of the analysis as follows. In particular, the opinion on teenage *Birth* control does not influence the respondents when they manifest their general political position (*liberal, moderate, conservative*), while the religious belief does. Nevertheless, when the political position is declared, the intensity of the political opinion of *conservative* people seems not to depend on their view about the teenage *Birth* control, while *liberal* respondents behave differently and make their decision about the strength of their opinion to vary according to the agreement on *Birth*. Finally, as well as the position, also the intensity of the political orientation is related to the respondent's religious creed.

6. Concluding remarks

Every continuation logit η_j is the logarithm of the ratio between π_j and the probability $\sum_{i=j+1}^{k} \pi_i$ of selecting a higher category. An alternative approach uses the local logits $\lambda_j = log\frac{\pi_j}{\pi_{j+1}}$, $j = 1,2,...,J-1$, that have at the denominator only the probability of the next category. In this case, the following expression for the probabilities π_j holds

$$\pi_j = \frac{\exp\{-\sum_{h=1}^{j} \lambda_h\}}{1+\sum_{i=1}^{k-1} \exp\{-\sum_{h=1}^{i} \lambda_h\}}, \quad j = 1,2,...,J-1. \tag{22}$$

As an alternative to (4) and coherently with the choice of the local logits, it is possible to define the nested local logits, associated to the interval $I_{i,k}$, in the following way

$$\lambda(i,k;h) = \log \frac{P(S_h^{i,k})}{P(S_{h+1}^{i,k})}, \quad h = 1,2,...,l_{i,k}-1.$$

The models discussed in Section 3 also apply to the nested local logits. It is worthwhile noting that the normalizing factor in (22) involves all the local logits. This implies that a probability π_j is function of all the logits, so that a change in any local logit affects the probabilities of every category. On the contrary, the denominator of (1) depends only on the first j continuation logits. This means that a probability π_j is not affected by a change in the logits η_m, $m > j$. This feature, which generalizes to the multistage case, is essential to understand why the continuation logits are coherent with a sequential selection procedure that stops once a category is not discarded without considering the higher categories, while the local logits do not meet the same property.

Moreover, note that the local logits and the expression (22) capture an essential aspect of the partial credit model by Masters (1982), where the local logits have a special structure in terms of person and item parameters, while the continuation logits are linked to the sequential Rash models by Tutz (1990).

Finally, as shown by Dardanoni and Forcina (1998), Cazzaro and Colombi (2006), the choice of a logit type may also follow from a chosen monotone dependence criterion. In particular, the likelihood ratio criterion is equivalent to increasing local logits with respect to covariates, while increasing continuation logits (with respect to covariates) correspond to the uniform dependence criterion. As the uniform dependence is the weakest, continuation logits are more likely to order stochastically the response distributions conditioned on the covariates.

References

Agresti, A. (2010). Analysis of Ordinal Categorical Data, 2^{nd} edition, J. Wiley & Sons, Hoboken.

Bartolucci, F., Colombi, R., Forcina, A. (2007). An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica*, **17**, 691-711.

Bergsma, W., Croon, M., Hagenaars, J. A. (2009). Marginal Models for Dependent, Clustered, and Longitudinal Categorical Data. Springer-Verlag, New York.

Cazzaro, M., Colombi, R. (2006). Maximum likelihood inference for log-linear models subject to constraints of double monotone dependence. *Statistical Methods and Applications*, **15**, 177-190.

Cazzaro, M., Colombi, R. (2009). Multinomial-Poisson models subject to inequality constraints. *Statistical Modelling*, **9**, 215-233.

Cazzaro, M., Colombi, R. (2013). Marginal nested interactions for contingency tables. *Communications in Statistics - Theory and Methods*, to appear.

Colombi, R., Forcina, A. (2001). Marginal regression models for the analysis of positive association. *Biometrika*, **88**, 1007-1019.

Colombi, R., Giordano, S., Cazzaro, M. (2012). R-package *hmmm*: hierarchical multinomial marginal models. *http://CRAN.R-project.org/package=hmmm.*

Dardanoni, V., Forcina, A. (1998). A unified approach to likelihood inference on stochastic orderings in a nonparametric context. *Journal of the American Statistical Association*, **93**, 1112-1123.

Fahrmeir, L., Tutz, G. (2001). Multivariate Statistical Modelling Based on Generalized Linear Models. Springer, New York.

Glonek, G. F. V., McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, **57**, 533-546.

Lang, J. B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *The Annals of Statistics*, **32**, 340-383.

Lang, J. B. (2005). Homogeneous linear predictor models for contingency tables. *The Journal of the American Statistical Association*, **100**, 121-134.

Lauritzen, S. L. (1996). Graphical Models. Clarendon Press, Oxford.

Masters, G. N. (1982). A Rash model for partial credit scoring. *Psychometrika*, **47**, 149-174.

McCullagh, P., Nelder, J. A. (1989). Generalized Linear Models. Chapmann and Hall, London.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, **43**, 39-55.

Tutz, G. (2013). Regression for Categorical Data. Cambridge University Press, Cambridge.

QdS – Vol. 15, 2013

On a copula model with CUB margins

Federico Andreis
Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano
E-mail: federico.andreis@unimi.it

Pier Alda Ferrari
Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano
E-mail: pieralda.ferrari@unimi.it

Summary: Particular emphasis has been put, lately, on the analysis of categorical data and many proposals have appeared, ranging from pure methodological contributions to more applicative ones. Among such proposals, the CUB class of distributions, a mixture model for the analysis of ordinal data has been successfully employed in various fields, and seems of particular interest. CUB are univariate models and do not possess, at present, a multivariate version. In this work, moving in this direction, we investigate the use of CUB in the framework of copula models, with respect to a Plackett copula model specification with CUB margins and discuss its potential and limitations.

Keywords: multivariate ordinal data, CUB, copula models, plackett copula

1. Introduction

The analysis of ordinal data is nowadays a field of great interest for the vast majority of applied fields and poses interesting challenges to statisticians in the development of an adequate methodology. Diverse proposals have been introduced during the recent years for their treatment, leading to important theoretical contributions from the scholars worldwide. Among such proposals, the authors deem worth of particular consideration the CUB (Piccolo, 2003) models, a class of univariate mixture distributions that has been successfully applied in many fields such as semiotics, ability assessment, medical research and customer satisfaction; the parsimonious parameterization and the ease of estimation and interpretation make CUB models a very useful tool for ordinal data analyses. Unfortunately, to date, a general version of CUB for multivariate data (that are

common when analyzing, for example, survey results), is still unavailable.
Feeling that multidimensional data coming from the same source (such as responses to a questionnaire items) should be treated as an ensemble in order to account for existing dependence structures, we see the need for an extension of this class. We note that an interesting first attempt at defining a bivariate CUB distribution using a multivariate model with fixed margins (specifically, the Plackett distribution) is made by Corduas (2011); we show how this proposal constitutes a special case of a more general class of models which is embedded in the framework of copulas (see Nelsen, 2010) and discuss its properties, limitations and possibility of extension to more than two dimensions. Particular attention is put on estimation methods and parameters interpretation, and a simulation study is presented, in which we investigate the behaviour of this bivariate version of CUB under many different scenarios.

2. Background

This section is intended to briefly review the general framework of both CUB and copula models.

2.1. CUB models

The CUB is a class of mixture models, possibly involving covariates, developed as a new approach for modeling discrete choices processes. The most common situation in which such approach can be employed regards the analysis of questionnaire data, with items responses evaluated on Likert scales and, thus, in the presence of ordinal data. CUB models are characterized by two components, related to *uncertainty* and *feeling*. The inherent uncertainty in answering an item is modeled through a discrete uniform variable, whereas the latent process leading to the choice is governed by the subjective *feeling*, and modeled using a Shifted-Binomial distribution. The probability of observing a particular response $r = 1, 2, ..., m$, with m known ($m > 3$ to ensure identifiability), to an item is expressed as a mixture of two such components as follows:

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

with $\pi \in (0, 1]$ and $\xi \in [0, 1]$.
π define the mixture weights and as such is inversely related to the amount of *uncertainty* in the answers (the higher π, the less the uniform component contributes to the mixture); ξ, on the other hand, is related to personal preferences and measures the strength of *feeling* or *adherence*, *agreement* with the item (the interpretation of ξ also depends on the kind of ordering adopted for the item).

2.2. Copula models

Following Nelsen (2010), a (two-dimensional) copula is a function $C : [0,1]^2 \to [0,1]$ that satisfies:

1. $\forall u, v \in [0,1]$,
$$C(u,0) = 0 = C(0,v)$$
and
$$C(u,1) = u, C(1,v) = v$$
2. $\forall u_1, u_2, v_1, v_2 \in [0,1]$ such that $u_1 \leq u_2, v_1 \leq v_2$,
$$C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \geq 0.$$

The definition is easily extended to k dimensions to characterize functions $C : [0,1]^k \to [0,1]$ with the same properties. Copulas are, thus, k-place, grounded and k-increasing real functions with the unit hypercube as domain. Sklar's Theorem (Sklar,1959) is central to the theory of copulas; for the bivariate case, it can be stated as follows.
Let H be a joint distribution function with margins (i.e. marginal distribution functions) F and G. Then there exists a copula C such that for all $x, y \in \overline{\mathbf{R}} = [-\infty, +\infty]$

$$H(x,y) = C[F(x), G(y)] \tag{2}$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on $Ran(F) \times Ran(G)$. Conversely, if C is a copula and F, G are distribution functions, then the distribution function H defined in (2) is a joint distribution function with margins F and G.

This remarkable result shows that copulas can be used to:

1. express known joint distribution functions in such a way that the dependence structure is captured by the copula as a function of the marginal distribution functions
2. link univariate distribution functions to different multivariate distribution functions having different dependence structure (copulas).

Copulas are usually characterized by parameters that govern the dependence among the margins, and different choices for the function C lead to different dependence structures. The use of copula models in this work relates to point 2, i.e. we use C as building tools to form a multivariate distribution, given the marginal distributions.

Particular care is needed when working with non-continuous margins, as in the case of the CUB models, due to the non-uniqueness of the copula representation; non-uniqueness stems from the fact that marginal distribution functions are not strictly monotonically increasing, rather monotonically non-decreasing, and do not possess an inverse in the usual sense, rather a pseudo-inverse (see, for example, Nelsen, 2010). The most severe consequence of this is that distribution functions built with copulas might not inherit interesting properties of C. Specifically, the copula parameterization cannot, alone, define the dependence structure (as in the continuous case): dependence-related measures become, in this case, margin-dependent (Genest et al., 2007). Nonetheless, copulas still are an easy-to-implement and interesting tool to build multivariate models, and under certain circumstances it is still possible to make assessments about dependence among margins. For example, some copula families possess the property of being ordered by Positive Quadrant Dependence (PQD, see Joe, 1997) and it is possible to show that this minimal requirements for copula parameters to be interpretable as dependence measures is granted even in the non-continuous case (Genest et al., 2007). The Plackett copula, as will be discussed in 3.2, does possess such ordering.

A concept that has been given much attention in the field of copulas, thanks to its practical implications (especially in the financial field), is *tail dependence*. General dependence indices (e.g. Spearman's ρ and Kendall's τ) provide a measure of the overall strength of the association, but can't give any insight about how this varies across the distribution (Venter, 2003, Nelsen, 2010); it might be of interest to focus the attention on events on the tails of a distribution, and the aforementioned measures fail to provide insight on such aspects. Focusing on the bivariate case, *lower* and *upper tail dependence* coefficients are defined (see, e.g., Nelsen, 2010) in terms of a copula C and for continuous margins, and usually labelled λ_L and λ_U, respectively. We say that a copula has no lower (upper) tail dependence if $\lambda_L = 0$ ($\lambda_U = 0$). Intuitively, tail dependence coefficients measures the dependence between the variables in the lower-left and upper-right quadrants of the unit square; geometrically, they measure the departure of the slope of C from the slope of the copula that assumes independence between the margins, the so called independence copula $\Pi(u, v) = uv$, in the corners of the unit square.
Genest et al. (2007) show that, at least when the margins have the same distribution, a copula based model retain the tail dependence properties of the copula C used to construct it even in the presence of non-continuous variables. We will discuss this point in Section 3.2, when referring to the Plackett copula with CUB margins.

2.3. Estimation methods for copula parameters

Let C_θ be a copula with parameter $\theta \in \Theta$ and margins $X \sim F_\alpha(x), Y \sim G_\beta(y)$ parameterized by $\alpha \in A$ and $\beta \in B$, respectively; all parameters may be vectorial. Let $c_\theta, f_\alpha, g_\beta$ be, respectively, the copula density and the margins densities (with respect to an appropriate measure), and let $\gamma = (\alpha, \beta, \theta)' \in \Gamma$ be the vector collecting all the model parameters. Estimation can be performed either by Joint Maximum Likelihood (JML), i.e.

$$\hat{\gamma} = \underset{\gamma \in \Gamma}{\operatorname{argmax}} \ln c_\theta$$

or following the Inference From Margins (IFM, Joe et al., 1996) scheme, which requires to estimate the marginal parameters separately first, to plug them into the copula density, and then to maximise with respect to the copula parameter, i.e.

$$\hat{\alpha}^{IFM} = \underset{\alpha \in A}{\operatorname{argmax}} \ln f_\alpha$$

$$\hat{\beta}^{IFM} = \underset{\beta \in B}{\operatorname{argmax}} \ln g_\beta$$

and then

$$\hat{\theta}^{IFM} = \underset{\theta \in \Theta}{\operatorname{argmax}} \ln c_\theta(\hat{\alpha}^{IFM}, \hat{\beta}^{IFM})$$

thus yielding

$$\hat{\gamma}^{IFM} = (\hat{\alpha}^{IFM}, \hat{\beta}^{IFM}, \hat{\theta}^{IFM})'.$$

IFM estimation is much more computationally efficient than joint ML and grants numerical stability even with small sample sizes. Moreover, Joe et al. (1996) have proven that, under regularity conditions, the IFM estimator is consistent and asymptotically normal. An important remark has to be done, however, with respect to the parameters variance/covariance matrix: whereas it is possible to obtain an estimate of it directly from the Hessian matrix when working with JML estimation, the theory of Generalised Estimating Equations underlying IFM requires to evaluate the Godambe Information matrix, which can be quite difficult to compute. Joe et al. (1996) show how the jackknife method can be used to obtain standard error estimates in an easier way.

Rank-based inversion methods, very useful and common in the copulas practice, should not, as strongly pointed out in Genest et al. (2007) be used when dealing with non-continuous margins because of bias-related issues. Bayesian estimation by means of Monte Carlo Markov Chain (MCMC) algorithms is also possible, but is not considered in this work.

3. Multivariate approach to CUB models

As said, CUB models have been developed to describe univariate discrete phenomena, e.g. the distribution of answers to a single questionnare item. Since questionnaires are usually composed by many different questions (say k), a complete analysis with CUB would require to separately estimate the k couples $(\pi_i, \xi_i), i = 1, ..., k$, that characterize each item. This disjoint analysis approach does not take into account the dependence (possibly) existing among items, which could be exploited to better catch further information about the phenomenon and enrich its understanding. Drawing on this, we intend to evaluate the feasibility of a multivariate approach to CUB modeling, through the use of copula models.

3.1. The CO-CUB model

We define a multidimensional extension of CUB, called CO-CUB model, as a multivariate copula with discrete margins, each following a CUB distribution.

Definition A k-dimensional ($k \geq 2$) CO-CUB model with copula C is a multivariate discrete variable with margins $R_i \sim CUB(\pi_i, \xi_i)$, $i = 1, ..., k$, each with support $\{1, ..., m_i\}$, $m_i > 3$, and joint distribution function given by:

$$\begin{aligned}\Psi(r_1, ..., r_{;}\underline{\pi}, \underline{\xi}, \underline{\theta}) &= P(R_1 \leq r_1, ..., R_k \leq r_k; \underline{\pi}, \underline{\xi}, \underline{\theta}) = \\ &= C_{\underline{\theta}}[F_1(r_1; \pi_1, \xi_1), ..., F_k(r_k; \pi_k, \xi_k)]\end{aligned} \tag{3}$$

where $\underline{\pi} = (\pi_1, ..., \pi_k)'$, $\underline{\xi} = (\xi_1, ..., \xi_k)'$ and for a particular choice of copula C, characterized by a parameter $\underline{\theta} = (\theta_1, ..., \theta_d)'$ taking values in some real d-dimensional space Θ defining the dependence structure of its components. $F_i(r_i) = F_i(r_i; \pi_i, \xi_i)$ stands for the distribution function of the i-th margin, i.e. $F_i(r_i) = P(R_i \leq r_i)$, and the support of the CO-CUB variable is the grid $\{1, ..., m_1\} \times ... \times \{1, ..., m_k\}$.

The whole parameter set for a k-dimensional CO-CUB is, then, the ordered triplet $(\underline{\pi}, \underline{\xi}, \underline{\theta}) \in (0, 1]^k \times [0, 1]^k \times \Theta$, having the following interpretation: by definition of copula, being margins of (3) CUB, parameters $(\underline{\pi}, \underline{\xi})$ retain the same interpretation as in the unidimensional case, while for what concerns the copula parameter $\underline{\theta}$, its interpretation as a dependence measure is connected with the specific copula C adopted, and will be further discussed later.

3.2. Possible choices for C: the Plackett copula

In this work we adopted, for the reasons discussed in Section 1, the Plackett copula in the bivariate case. The Plackett copula family is defined, for $k = 2$, as:

$$C_\theta(u, v) = \begin{cases} \frac{A_\theta(u,v) - \sqrt{A_\theta^2(u,v) - 4\theta uv(\theta-1)}}{2(\theta-1)} & \theta \in (0, +\infty) \setminus \{1\} \\ uv & \theta = 1 \end{cases}$$

where $(u, v) \in [0, 1]^2$, $A_\theta(u, v) = 1 + (\theta - 1)(u + v)$ and $\theta > 0$ is a dependence parameter. Specifically, this multivariate model is able to describe different dependence structures between the margins u, v for different values of θ:

$$\begin{cases} \theta \in (0, 1) & \text{negative dependence} \\ \theta = 1 & \text{independence} \\ \theta \in (1, +\infty) & \text{positive dependence.} \end{cases}$$

This copula is *comprehensive* (Nelsen, 2010), meaning that it can attain the Fréchet-Hoeffding lower and upper bounds (as $\theta \to 0^+$ and $\theta \to \infty$, respectively) as well as the independence copula (for $\theta = 1$). This is a desirable property in that, at least when dealing with continuous margins, this model can cover a very wide range of dependence structures, ranging from countermonotonicity (lower bound) to comonotonicity (upper bound). It is interesting to note that there is a functional relationship between the Plackett copula parameter θ and Spearman's ρ when the margins are continuous and $\theta \neq 1$:

$$\rho(\theta) = \frac{\theta + 1}{\theta - 1} - \frac{2\theta}{(\theta - 1)^2} \ln \theta. \tag{4}$$

It is easy to verify that $\lim_{\theta \to 0^+} \rho_\theta = -1$, $\lim_{\theta \to 1} \rho_\theta = 0$ and $\lim_{\theta \to +\infty} \rho_\theta = 1$. Moreover, ρ_θ is a strictly increasing function of θ, which makes the Plackett family *positively ordered* by its parameter, meaning that $C_\theta \geq C_{\theta'} \iff \theta \geq \theta'$; this is an important feature since it constitutes a minimal requirement to be able to interpret θ as a dependence parameter (Joe, 1997). There seems to be no closed form expression for Kendall's τ for members of the Plackett family.

When dealing with discrete margins, however, relationships such as the one in Equation 4 become more involved. The range of values for usual dependence measures (e.g., Pearson's ρ) is usually narrower in the discrete case than it is in the continuous case (see, e.g., Ferrari and Barbiero, 2012); if we consider Equation 4, we could then expect (also from a merely analytical point of view) a reduction of the range of θ's admissible values and, hence, of the copula's spannable range of dependence. In reality, it can be shown (Genest et al., 2007) that every measure of dependence among discrete margins computed on the basis of a copula C is dependent not only on the copula parameter θ, but also on the particular margins themselves: this means its expression could be more complex than Equation 4, involving, for example, also the parameters of the marginal distributions. In light of these considerations, and in the hope of retaining some flexibility in modeling, the choice of a copula that can, at least in principle, span the whole range of dependence (that is, a *comprehensive* copula), seems to be a good starting point.

The Plackett copula exhibits no tail dependence, i.e. $\lambda_L = \lambda_U = 0$. This might be a limitation thinking, e.g., of assessment questionnaires, in which dependence among items is expected to behave differently along the Likert measurement scale. For example, respondents that indicate a high level for an item, might be more likely to do the same for a closely related item, or vice-versa. The Plackett copula cannot take into account such feature. This copula, moreover, satisfies the conditions for *radial symmetry* (Joe, 1997), but whether this property is preserved even in the presence of CUB margins is still under study.

A more general k-dimensional version of the Plackett copula can be derived (see Molenberghs, 1992 and Tibaldi et al., 2004) but appears quite complex. We reckon that alternative interesting tools to go beyond the two dimensions, while still retaining this copula good properties and allowing, moreover, for more flexibility can be Pair Copula

Construction and Vines (not discussed here).

3.3. The CO-CUB model with Plackett copula - without covariates

Focusing on the bivariate case, the CO-CUB model with Plackett copula is then:

$$C_\theta(u_{\pi_1,\xi_1}, v_{\pi_2,\xi_2}) = \frac{A_\theta(u_{\pi_1,\xi_1}, v_{\pi_2,\xi_2}) - \sqrt{A_\theta^2(u_{\pi_1,\xi_1}, v_{\pi_2,\xi_2}) - 4\theta u_{\pi_1,\xi_1} v_{\pi_2,\xi_2}(\theta-1)}}{2(\theta-1)}$$

for $\theta \in (0,1) \cup (1,+\infty)$, and

$$C_1(u_{\pi_1,\xi_1}, v_{\pi_2,\xi_2}) = u_{\pi_1,\xi_1} v_{\pi_2,\xi_2}.$$

Here we have that

$$u_{\pi_1,\xi_1} = F_1(r_1) = \sum_{i=1}^{r_1} \left[\pi_1 \binom{m_1-1}{i-1} \xi^{m_1-i}(1-\xi_1)^{i-1} + (1-\pi_1)\frac{1}{m_1} \right]$$

where m_1 is the (fixed) number of categories of the first CUB model; similarly for v_{π_2,ξ_2} and we will from now on assume $m_1 = m_2 = m > 3$. The copula probability mass function can be obtained as:

$$c_\theta(r_1, r_2) = C_\theta(b_1, b_2) - C_\theta(a_1, b_2) - C_\theta(b_1, a_2) + C_\theta(a_1, a_2)$$

where $a_i = F_i(r_i), b_i = F_i(r_i - 1), i = 1, 2$.

The model parameters are then $\pi_1, \xi_1, \pi_2, \xi_2$ and θ. Estimation can be performed either by JML or IFM. Bayesian techniques have also been considered: drawing on an existing work on MCMC estimation for univariate CUB models (Deldossi et al., 2013), a routine in JAGS language has been written that can handle CO-CUB model parameters estimation.

3.4. The CO-CUB model with Plackett copula - with covariates

One of the most appealing features of CUB models is that they allow for a straightforward inclusion of covariates; the same can apply to the CO-CUB model, with a few remarks. Given a set of covariates $\mathbf{X}$, it is necessary to decide how to let them enter the model. Three possibilities exist:

1. covariates enter the model at the marginal level only, i.e. each margin is a CUB(p, q), as defined in Iannario (2008)
2. covariates enter the model at the dependence level only, i.e. $\theta = f(\mathbf{X})$, for a suitable link function f

3. covariates enter the model at both the marginal level and at the dependence level.

Method 1 would allow to describe different uncertainty and feeling patterns based on levels of $\mathbf{X}$, but this differences wouldn't be reflected by the dependence parameter, which would be a single value θ. For example, it would be possible to describe for each item whether males (M) tend to answer with higher (or lower) feeling and uncertainty as compared to females (F), but items dependence would be assumed to be independent of sex. In terms of model parameters, $\theta_M = \theta_F$, while $(\pi_{i,M}, \xi_{i,M})$ and $(\pi_{i,F}, \xi_{i,F})$ are allowed to be different; this might not be a realistic assumption. It could be more adequate, in this case, to decide to employ the copula to model marginal residuals, rather then the CUB(p, q) directly. A similar approach has been applied in a different context, see Disegna et al. (2013).

Method 2 would allow to describe different dependence patterns based on covariates, assuming implicitely that they do not affect the marginal distributions. For example, it would be possible to describe the effect of gender on how the responses to two items are related, but not how it affects the answers themselves. In terms of CO-CUB parameters, this would imply that θ_M might be different from θ_F, while $\pi_{i,M} = \pi_{i,F}$ and $\xi_{i,M} = \xi_{i,F}, i = 1, 2$. This might not be a realistic assumption and, moreover, in the modeling setting we are proposing relevant information might be neglected by ignoring the (possible) impact of covariates on the marginal behaviour.

Finally, method 3, being a combination of 1 and 2, would allow to describe both different dependence patterns and marginal behaviours, based on $\mathbf{X}$. This approach, known as conditional copula, has been widely applied (see, e.g., Patton, 2006 and Acar et al., 2011), especially in the financial context. The definition of a suitable link function to be able to describe θ as a function of the covariates without altering its range is required, and the margins can be modeled as CUB(p, q) distributions.

Again, both JML and IFM, as well as MCMC techniques, can be considered for estimation of the model parameters.

4. Simulation study

Due to the difficulty to obtain analytical results with regards to the CO-CUB model, a simulation study is conducted in the bivariate case and in the absence of covariates, aimed at investigating the following points:

1. consistency of CO-CUB results with those obtained from univariate CUB models
2. behaviour of the copula parameter θ with respect to the strenght of dependence between items
3. behaviour of the copula parameter θ with respect to the marginal parameters
4. goodness of fit of the CO-CUB model.

Specifically, we are interested in points 2-4, because they are directly connected to the characteristics of the bivariate (multivariate) version of CUB models we are studying. All computations have been carried out using *R 3.0.2* (R Core Team, 2013).

4.1. Simulation design

A number of combinations of marginal parameters levels and strength of dependence between items have been investigated, for total of 27 scenarios. Two data generating mechanisms are considered: *GenOrd* (Barbiero et al., 2012, implemented in the package *GenOrd*, function textitordsample) and Iman-Conover method (Iman et al., 1982, implemented in the package *mc2d*, function *cornode*)

- *GenOrd*: generates from every possible distribution with fixed cdf and given Pearson's ρ or Spearman's ρ, by means of a Normal copula
- Iman-Conover: generates from given CUB margins and Spearman's ρ, but is joint distribution free.

2000 runs with 1000 observations are generated with both methods for different scenarios; each scenario is characterized by the parameters vector $\gamma = (\pi_1, \pi_2, \xi_1, \xi_2, \rho)'$, each component of which takes values in $\{0.25, 0.50, 0.75\}$, where ρ is Spearman's correlation coefficient and $\pi_1 = \pi_2, \xi_1 = \xi_2, m_1 = m_2 = 7$ (this means that the marginal distributions we consider are identical).

No sensible differences between *GenOrd* and Iman-Conover in terms of results pertaining the four main targets of the simulation study have been observed, so we decide to discuss only those obtained with the Iman-Conover generating mechanism.

Both JML and IFM have been considered; the following table summarizes some considerations about the two estimation methods:

JML	IFM
• works fine with Plackett copula over all the investigated parameter space	• works fine with many copulas over all the investigated parameter space
• no effect of ρ detected on the joint distributions of $(\hat{\pi}_i, \hat{\xi}_i), i = 1, 2$ and $(\hat{\pi}_1, \hat{\pi}_2), (\hat{\xi}_1, \hat{\xi}_2)$	• no effect of ρ detected on the joint distributions of $(\hat{\pi}_i^{IFM}, \hat{\xi}_i^{IFM})$, *i=1,2*, and $(\hat{\pi}_1^{IFM}, \hat{\pi}_2^{IFM})$, $(\hat{\xi}_1^{IFM}, \hat{\xi}_2^{IFM})$
• computationally demanding and weak in the nearings of parameter space boundaries	• computationally efficient and somewhat numerically robust in the nearings of parameter space boundaries
• $\hat{\gamma}$ variance/covariance matrix obtainable by inversion of the Hessian matrix	• $\hat{\gamma}^{IFM}$ variance/covariance matrix obtainable from Godambe information matrix → jackknife method advocated
• ML theory for distribution of $\hat{\gamma}$	• IFM theory for asymptotic distribution of $\hat{\gamma}^{IFM}$, it's CAN

We present only the results obtained via IFM estimation: theoretical results and exploratory analyses, together with a sufficiently large sample size endorse the equivalence of the two methods, and the computational efficiency gain is really impressive (the time for the simulation is reduced by a factor of about 20).

The model fit is then assessed, following Corduas (2011), by means of the normalized dissimilarity index Δ:

$$\Delta = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}|\hat{c}_\theta - \frac{n_{ij}}{n}| \tag{5}$$

where n_{ij} is the observed frequency of the couple (i, j) and n the the total sample size.

4.2. Simulation results

We discuss the results regarding points 1-4 in the following plots that provide interesting insights into the properties of the CO-CUB model with Plackett copula.

Figure 1 shows the scatter plots of marginal CUB parameters estimates for one of the two generated margins along all the combinations of true parameters considered. The estimates distribution confirms itself unbiased for (π, ξ), exhibits correlation decreasing with π and incorrelation for $\xi = 0.5$ (graphically represented by segments of local regression lines of ξ on π). Given the estimation method (IFM) this results were, of course, expected on the basis of the theory for univariate CUB models; it is still interest-

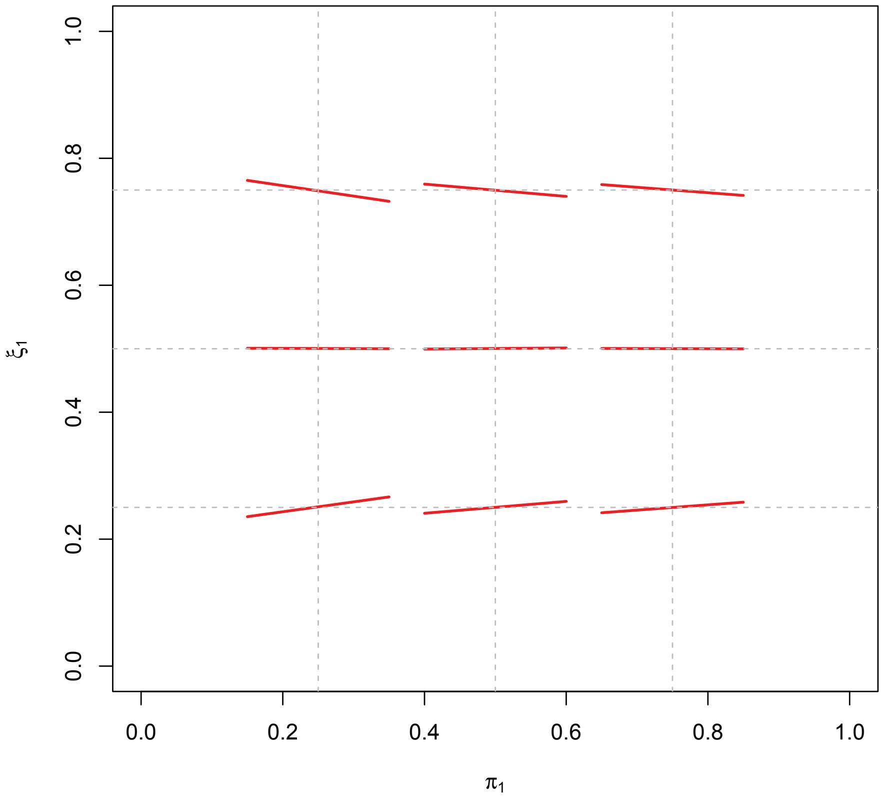


Figure 1. CUB estimates scatter plot

ing, nonetheless, to observe the correlation pattern between π and ξ, strongly dependent on the levels of both uncertainty and feeling.

Figure 2 highlights the relationships existing between θ and ρ and θ and the marginal parameters. The Plackett copula parameter θ appears to be in a monotonically increasing relationship with ρ, that is supported by the theory: Plackett copula family is ordered by PQD, thus higher values of θ correspond to higher levels of dependence. The relationship appears clearly non-linear, involving also θ variability, which seems to be increasing with correlation (note that θ is not a relative measure).

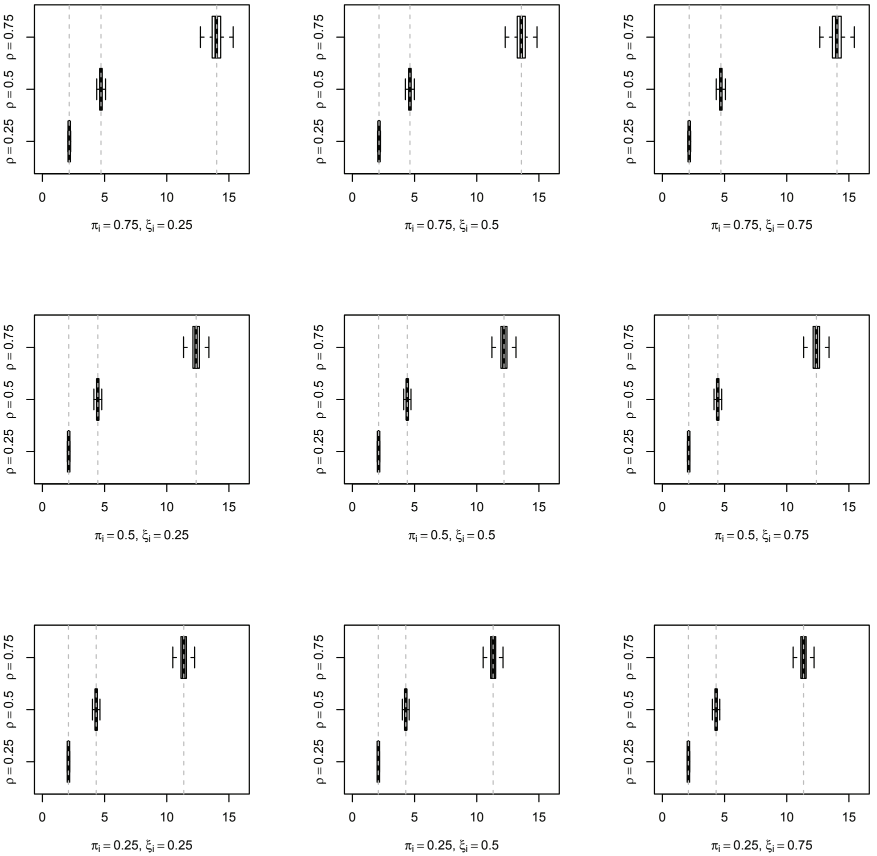


Figure 2. θ-plot: $\rho = 0.25$ (solid line), $\rho = 0.50$ (dashed), $\rho = 0.75$ (dotted)

As for what concerns the behaviour of the copula parameer as related to π and ξ, θ appears to be affected by marginal parameters values, even if slightly. It would seem that θ tends to increase in both location and variability monotonically with π, whereas the effect of ξ seems negligible. Derivation of a closed form relationship linking θ, ρ and the marginal parameters appears difficult.

Figure 3 summarizes the goodness of fit in terms of Δ (equation 5) of the CO-CUB with Plackett copula as compared to the independence assumption (copula Π).

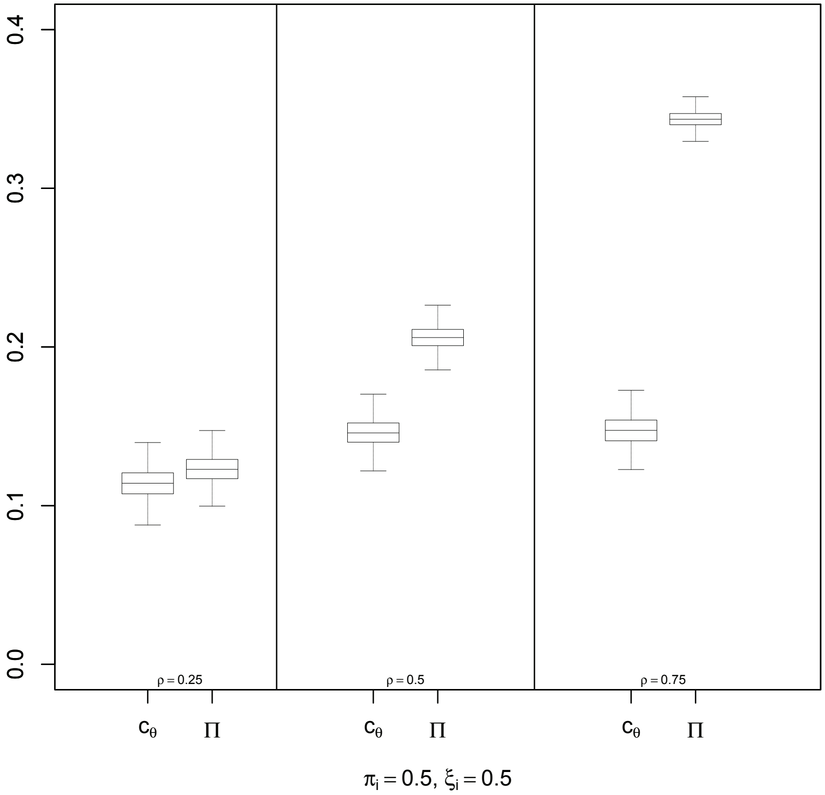


Figure 3. Δ-plot: Plackett copula c_θ vs Independence copula Π, $\pi = \xi = 0.5$

Inspection of the above plot shows how the fit is quite similar for $\rho = 0.25$ (low correlation). As correlation increases, though, the Plackett copula maintains a low (albeit slightly increasing) Δ, while Π yields an increasingly larger dissimilarity, reaching values as high as three times the Δ of the Plackett Copula when $\rho = 0.75$. Δ-plots for all the combinations of levels of true parameters are collected in Figure 4, showing the advantage of employing the copula model and its additional parameter θ in terms of fit when dependence exist, regardless of the specific value of (π, ξ).

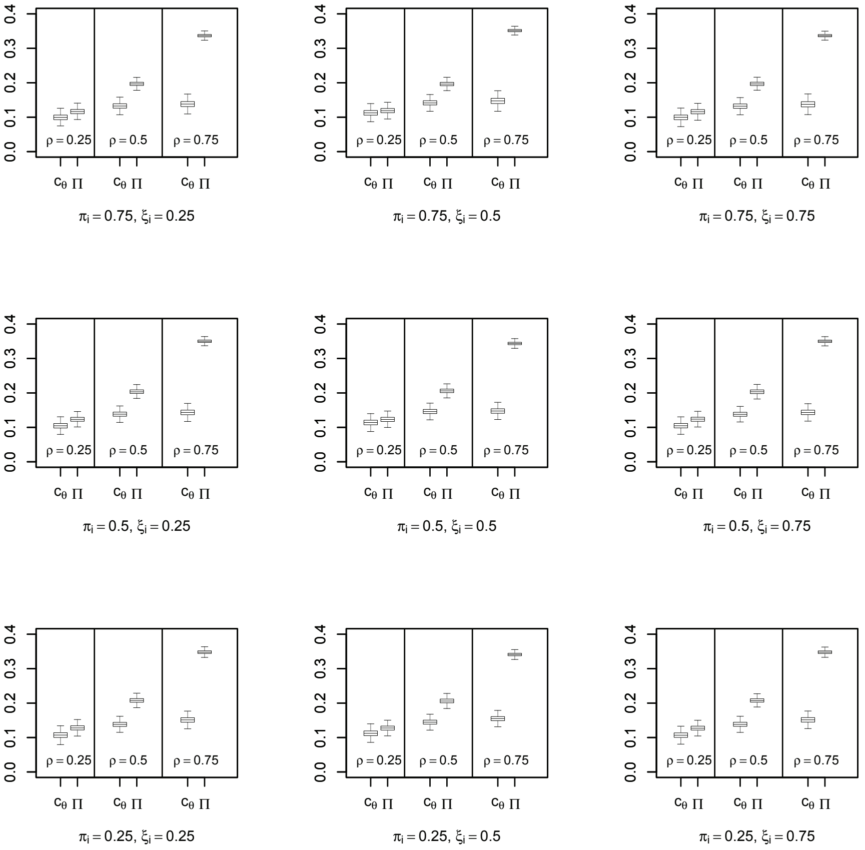


Figure 4. Δ-plot: Plackett copula c_θ vs Independence copula Π

Figure 5 depicts some examples CO-CUB models with Plackett copula varying dependence strength and marginal parameters. R_1 and R_2 axes show the ranks of the univariate margins, whereas the vertical axis shows the corresponding values of the CO-CUB density function; overimposed, on the margins, the univariate CUB distributions.

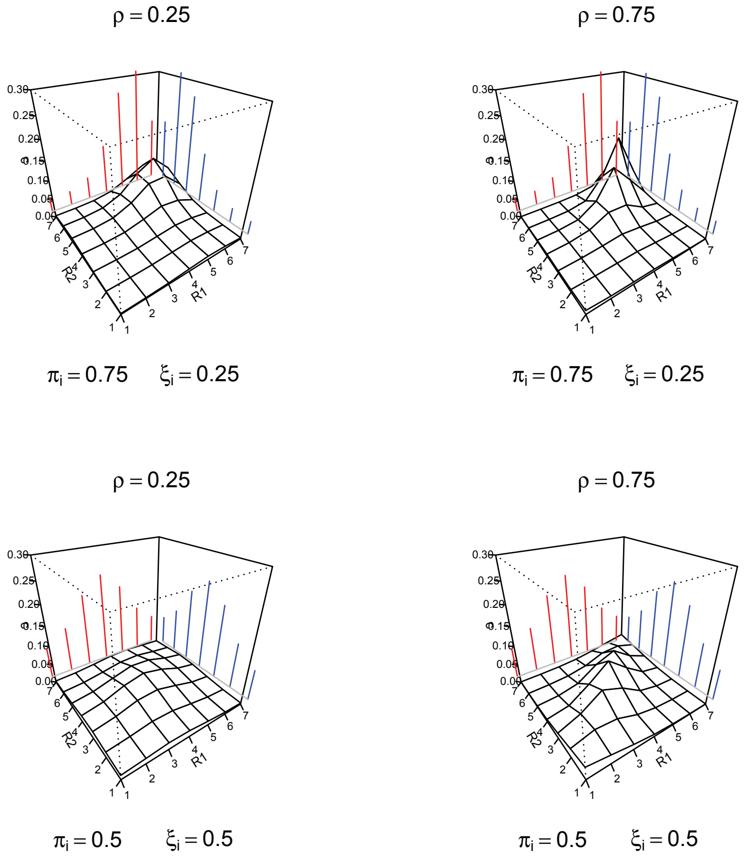


Figure 5. Plots of Plackett CO-CUB densities for some combinations of (π, ξ) *and* ρ

5. Conclusive remarks

In this work we investigate the properties of a copula model with CUB margins. A flexible k-dimensional approach to analyze ordinal data, the CO-CUB model, is discussed and advantages and limitations of the choice of a Plackett copula as its core structure are investigated with respect to the two-dimensional case; the possibility of including covariates in the CO-CUB is presented, showing how such extension is straightforward, yet requires a careful evaluation of the modeling assumptions. Attention is focused on the Plackett family because of preliminary studies on various copula choices

(Andreis et al., 2013) that appointed it as a good candidate thanks to its nice properties, and drawing on a first interesting work on the topic by Corduas (2011). Our main contribution is to contextualize the proposal of a bivariate Plackett distribution with CUB margins within the framework of copula theory, thus providing an alternative point of view that leads to a greater insight on this topic, also through a simulation study, and to lay the foundation for extension to higher multidimensional modeling using CUB variables.
The simulation study shows the consistency of the CO-CUB model with Plackett copula with respect to the univariate CUB and highlights the relationships linking the copula parameter θ to the strength of dependence between the margins and to the marginal parameters. The model's goodness of fit is compared to that of the naïf assumption of independence (Π copula), showing the advantage of including an additional parameter (θ). Copulas are a well known and widely used tool with a strong theoretical background when dealing with continuous margins, while their application in the presence of non-continuous variables is at present a very new and promising field. The discrete nature of the CUB variable we consider gives rise to some issues related, in particular, to the interpretation of the copula parameter θ. The simulation study reveals the dependence of θ on the marginal parameters, which should be subject to further investigation. On the basis of the results of the simulation study and of theoretical considerations, the Plackett copula would seem, overall, a good choice for modeling multivariate data with CUB margins. There is, nonetheless, a relevant limitation: when dealing with assessment data (e.g. responses to a questionnaire), heavy-tailed distributions are very likely to arise, and the absence of tail dependence in this copula ($\lambda_L = \lambda_U = 0$) prevents a proper modeling of such situations.

Further research on the topic should address a deeper understanding of the relationships between θ and marginal CUB parameters, as well as their interpretation. Modeling more than 2 dimensions in a flexible and informative way, possibly investigating the use of Pair Copula Construction and Vines, is also an important open question. Estimation methods such as MCMC algorithms can be considered as an alternative to JML and IFM and thus investigated, as well as selection among competing models.

Acknowledgements: The authors acknowledge financial support from the European Social Fund Grant (FSE), Lombardy Region.

References

Acar, E.F., Craiu, R.V., and Yao, F.: Dependence Calibration in Conditional Copulas: A Nonparametric Approach. Biometrics, 67, 2, 445-453 (2011).

Andreis, F. and Ferrari, P.A.: A proposal for the multidimensional extension of CUB models. Cladag 2013. 9th Meeting of the Classification and Data Analysis Group. Book of Abstracts (2013) available at:
http://www.cladag2013.it/images/file/CLADAG2013_Abstract.pdf.

Barbiero, A. and Ferrari, P.A.: Simulating Ordinal Data. Journal of Multivariate Behavioral Research, 47, 4, 566-589 (2012).

Carley, H.: Maximum and minimum extensions of finite subcopulas. Communications in Statistics: Theory and Methods, 31, 2151-2166 (2002).

Corduas, M.: Modelling correlated bivariate ordinal data with CUB marginals. Quaderni di Statistica **13**, 109–119 (2011).

Disegna, M., Durante, F. and Foscolo, E.: A Multivariate Nonlinear Analysis of Tourism Expenditures - PREPRINT. BEMPS - Bozen Economics & management Paper Series, 10 (2013) available at: **http://ideas.repec.org/p/bzn/wpaper/bemps10.html**.

Ferrari, P.A. and Barbiero, A.: Simulating ordinal data. Multivariate Behavioral Research, 47:4, 566-589 (2012).

Genest, C. and Neslehova, J.: A Primer on Copulas for Count Data. ASTIN Bulletin vol.37 no.2 (2007) available at:
http://www.actuaries.org/LIBRARY/ASTIN/vol37no2/475.pdf.

Iannario M.: A class of models for ordinal variables with covariates effects. Quaderni di Statistica, 10, 5372 (2008).

Iannario, M. and Piccolo, D.: A program in R for CUB models inference (Version 2.0), (2009) available at: **http://www.dipstat.unina.it**.

Iman, R.L. and Conover, W.J.: A distribution-free approach to inducing rank correlation among input variables. Communications in Statistics, B11, 311-334 (1982).

Joe, H.: Multivariate Models and Dependence Concepts. Chapman & Hall, London (1997).

Joe, H. and Xu, J.J.: The estimation method of inference functions for margins for multivariate models. Technical Report n.166, Department of Statistics, University of British Columbia (1996).

Molenberghs, G. (1992): A Full Maximum Likelihood Method for the Analysis of Multivariate Ordered Categorical Data. Unpublished Ph.D. dissertation, University of Antwerp.

Nelsen, R.B.: An Introduction to Copulas. Springer (2010).

Deldossi, L. and Paroli, R.: Inference on the CUB model: an MCMC approach. In Classification and Data Mining, 8, 19-26, Springer Berlin (2013).

Patton, A.J.: Modelling asymmetric exchange rate dependence. International Economic Review, 47, 527-556 (2006).

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

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2013) available at: **http://www.R-project.org**.

Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris, **8**, 229–231 (1959).

Tibaldi, F., Barbosa, F.T. and Molenberghs, G.: Modelling associations between time-to-event responses in pilot cancer clinical trials using a Plackett-Dale model. Statistics in Medicine, 23(14), 2173-2186 (2004).

Venter, G.: Tails of copulas. Proceedings of the Casualty Actuarial Society, **LXXXIX**, 170&171, 68-113, (2002) available at:
http://www.casact.org/pubs/proceed/proceed02/02068.pdf.

QdS – Vol. 15, 2013

Two competing models for ordinal longitudinal data with time-varying latent effects: an application to evaluate hospital efficiency

Fulvia Pennoni
Department of Statistics and Quantitative Methods, University of Milano-Bicocca
E-mail: fulvia.pennoni@unimib.it

Giorgio Vittadini
Department of Statistics and Quantitative Methods, University of Milano-Bicocca
E-mail: giorgio.vittadini@unimib.it

Summary: We propose a model for longitudinal data with a suitable parameterization based on global logits to account for the ordinal response variable which incorporates observed covariates and time-varying latent unit specific effects. As an example we consider a derived ordinal variable by using the total revenues and discharges of the hospitals. For example, the hospital can vary on the response variable because of the unobserved covariates such as general manager ability (unobserved heterogeneity). The distribution of the latter may be discrete-valued or continuous-valued. In the first case it is based on a first order homogeneous Markov chain with a fixed number of states. In the second case it is a mixture of auto-regressive AR(1) processes with specific mean values and correlation coefficients and common variances. Maximum likelihood estimation of the model parameters is performed by using the Expectation-Maximization algorithm and the Newton-Raphson algorithm. Standard errors are obtained by using the observed information matrix. The results of the application to data referred to some hospitals in Lombardy are illustrated.

Keywords: Diagnosis-related groups, Latent Markov model, Mixture latent auto-regressive model, Unobserved heterogeneity, Panel data, Path prediction

1. Introduction

In the analysis of longitudinal data, the interest is often focused on the evolution of a latent characteristic related to the units under investigation over time, which may be measured by one or more occasion specific ordinal response variables. The model proposed by Bartolucci, Bacci and Pennoni (2014) may be seen a promising tool to be used in many practical applications. In fact, with its flexible structure it can be applied when the ordinal response variable has a limited number of categories and both when few or many time occasions are available. The proposed model is based on two different formulations of the distribution of the latent process in order to properly model the unobserved heterogeneity. When the model is formulated according to a discrete distributive assumption on the unobserved heterogeneity a latent Markov model results (for a review see Bartolucci, Farcomeni, Pennoni, 2013). When it is formulated according to continuous distributive assumption on the unobserved heterogeneity a mixture of autoregressive processes results. The adopted parameterization accounts for a model which is parsimonious, easy to interpret and with hypotheses which may be interesting to test.

In this paper we illustrate the model formulation by focusing on a derived ordinal variable which can be considered as a measure of the efficiency of the hospital spending policy which is observed at four occasions. The application we propose concerns some hospitals in Lombardy to examine if they have efficiency gains during the period 2008-2011. In such a context the latent variables have the role to account for hospital unobserved heterogeneity by introducing a unit specific random intercept. This aspect is in connection with the inclusion of observed covariates which may also be time-varying and do not fully explain the heterogeneity between the level of efficiency gained by different hospitals. Considering the proposed application the observed covariates are measured each year and they are related with the hospital characteristics.

The paper focus on a suitable parameterization for the measurement model according to the statistical literature on the ordinal response variables (see among others Colombi and Forcina, 2001) and which makes the latent components of the model interpretable. We admit that the effect of the unobserved covariates has its own dynamics. In the context of study this is important for example to account for those factors such as the general manager ability which affect the budgetary of the hospital. In the applied case of study, the hospital general manager is indeed lawfully responsible for all the activities performed in her/his hospital. In such a context, experience, ability and skills which may increase over time are important features to be considered.

Among many models developed to address the related aspect of technical efficiency we mention the stochastic frontier model proposed by Aigner, Lovell and Schmidt (1977) which has been extended in several directions (see among others Green, 2005). The model proposed by Battese and Coeli (1995) has the advantage to allow for time-varying efficiency components. Recent reviews of such models may be found in Green (2009) and Kumbhakar, Lien, Brian (2014). The main goal of the stochastic frontier models is that to evaluate the performance of the hospitals considered as firms in terms of technical or cost inefficiency which is considered as a failure to attain the production frontier. The models are estimated econometrically by attaining for a random noise and a technical inefficiency component which is assumed as a non-negative random variable with a half-normal, exponential, truncated normal or gamma distribution. In such a context Colombi et al. (2014) suggest the use of a four random components stochastic frontier model to allow for inefficiencies of different nature.

The paper is organized as follows. In the next section, we introduce the data used to illustrate the proposed model and we outline the main research questions. In Section 3 we introduce the notation and the proposed longitudinal model for the ordinal response with the two types of formulation for the latent component. In Section 4 we focus on some details related to the log-likelihood and its maximization procedure first for the latent Markov formulation and then for the mixed latent auto-regressive formulation. In Section 5 we show the results of the application based on data of the ward of general medicine and in the last section we outline some conclusions.

2. The data

The data derive from a large administrative database provided by the health care department of Lombardy region regarding hospital's features. It is worth mentioning that among the Italian regions, Lombardy is the most populated one with the highest GDP per capita among the European countries. The data cover the full population of patients for the general medicine ward which is the one with the highest discharges and number of beds compared to the other wards in the region. They are related to 120 hospitals and cover the years 2008, 2009, 2010 and 2011.

The variables of interest are the yearly revenues from discharges and the number of outpatient discharges in the ward. It is important to mention that a feature of the healthcare system is the recent introduction of a new perspective hospital reimbursement regulating the hospital compensations for the different treatments provided. It is based on the diagnosis-related group tariff which accounts for the treatment's complexity. According to this system hospitals receive a fixed rate for each admission depending on a patient's diagnosis. Even if this system provides a more efficient way to administrate the hospital it also may contribute to develop some opportunistic behaviours known in the literature as upcoding, cream skimming and readmissions. The latter includes the monetary incentives coming from admitting patients. For a more detailed description on these features see among others Berta et al. (2010) and Herwartz and Strumann (2014). As a consequence, hospitals face an increased pressure on their financial performance and a risk of insolvency.

The yearly revenues are related also to the diagnosis-related group tariff which accounts for the treatment's complexity and therefore they account for the severity of health care procedure provided to the patient. In the proposed application we suggest to consider the ratio between the yearly revenues and the yearly number of discharges. It accounts for the more complex case mix of the patients and it can be interpreted as an efficiency monetary measure of the hospital and named pre capita revenue. For every hospital the following time-varying covariates are also available: the total number of beds, the yearly hours of activity of physicians, nurses and other employees of the hospital and the hours of activity of the surgery rooms. Table 1 shows some descriptive statistics of the pre capita revenue and of the available time-varying covariates over the time occasions. It is important to mention that of 120 hospitals two of them have been deleted from the analysis due to the fact that they showed a very high or very low value of the pre capita revenue compared to the other hospitals. Then, we consider as response variable for each year that based on four ordered categories corresponding to the four quarters of its distribution. The response variable is then measured on a scale based on four ordered categories: 'low', 'medium', 'high' and 'very high'.

Table 2 shows the empirical transition matrix of the response variable. Each row of this matrix shows the percentage frequencies of the four response categories at occasion t given the response at occasion $t-1$, with $t = 2,\dots,T$. The transition matrix shows a high degree of persistence on the same category of efficiency level since many hospitals at time t are in the same efficiency category in which they are at time $t-1$ and the percentages included between 20% and 31% lie in an adjacent category.

Table 1. Distribution of variables over the time occasions.

	Year			
$Variable$	2008	2009	2010	2011
pre capita revenue	2813.86	2886.57	3011.58	3074.65
beds (number)	45.51	45.39	44.78	44.10
physicans (hours)	245,597.44	247,104.11	214,122.28	206,485.25
nurses (hours)	481,504.42	485,475.39	398,980.55	345,871.21
others (hours)	460,843.22	459,612.36	309,272.25	156,393.88
surgey rooms (hours)	7,691.40	7,675.94	8,144.78	7,940.05

Table 2. Conditional empirical distribution of the response variable at time t given the response at time $t-1$, with $t = 2,\dots,T$ (percentage frequencies).

	Ratio at t				
Ratio at $t-1$	low	medium	high	very high	Total
low	76.7	20.0	3.3	0.0	100.0
medium	20.7	48.3	24.1	6.9	100.0
high	3.4	31.0	44.8	20.7	100.0
very high	0.0	0.0	26.7	73.3	100.0

3. The proposed model

With reference to a sample of n hospitals observed at T time occasions, let y_{it} be the ordinal response variable for hospital i at occasion t with a number of categories denoted by J, and let $\boldsymbol{x}_{it}$ be a corresponding column vector of covariates, with $i = 1, \dots, n$ and $t = 1, \dots, T$. We also denote by $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})$ the vector of response variables and by $\boldsymbol{X}_i = (\boldsymbol{x}_{i1} \cdots \boldsymbol{x}_{iT})$ the matrix of time-varying and time-constant covariates for hospital i.

The model we formulate is based on the assumption that $y_{it} = G(y^*_{it})$, where y^*_{it} follows the model

$$y^*_{it} = \alpha_{it} + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \eta_{it}, \quad i = 1, \dots, n, \; t = 1, \dots, T,$$

with η_{it} being independent error terms with a standard logistic distribution, and $G(\cdot)$ is a link function which models the relationship between each response variable y_{it} and the corresponding latent variable α_{it} and the vector of covariates $\boldsymbol{x}_{it}$. In such a case it is a function of cut-points $\mu_1 \geq \cdots \geq \mu_{J-1}$ and it can be formulated as

$$G(y^*) = \begin{cases} 1 & y^* \leq -\mu_1, \\ 2 & -\mu_1 < y^* \leq -\mu_2, \\ \vdots & \vdots \\ J & y^* > -\mu_{J-1}. \end{cases}$$

The basic assumptions of the model are that for every sample unit i, $y^*_{it}, \dots, y^*_{iT}$ are conditionally independent given $(\alpha_{i1}, \dots, \alpha_{iT})$ and $\boldsymbol{X}_i$, and that due to the ordinal nature of the response variable we have

$$\log \frac{p(y_{it} \geq j | \alpha_{it}, \boldsymbol{x}_{it})}{p(y_{it} < j | \alpha_{it}, \boldsymbol{x}_{it})} = \mu_j + \alpha_{it} + \boldsymbol{x}'_{it}\boldsymbol{\beta}, \tag{1}$$

with $i = 1, \dots, n$, $t = 1, \dots, T$, $j = 2, \dots, J$. The parameterization adopted which is based on global logits for the distribution of each response variable is particularly suitable as we deal with an underling continuous outcome which is suitable discretized (McCullagh, 1980). For parsimony and easiness of interpretation we are assuming that the effect of covariates and of the unobserved individual parameters do not depend on the specific response category. This parametrization is based on one parameter for each latent state, which is an aggregation of hospitals sharing the same propensity towards efficiency gains and one cut-point for each response category. Then, the latent states may be ordered according to the highest and lowest level of efficiency. Also note that the cut-points are common to all the response variables, since these variables correspond to repeated measurements of the same phenomenon.

The distribution of the latent variable may be based on a discrete or on a continuous latent process. The discrete latent process formulation is more natural in some contexts and it typically assumes that, for all i, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{iT})$ follows a first-order homogenous Markov chain with k states denoted by $\xi_1, \dots, \xi_k$. This chain has initial probabilities π_h and transition probabilities $\pi_{h_1 h_2}$, with

$$\begin{aligned} \pi_h &= p(\alpha_{i1} = \xi_h), \quad h = 1, \dots, k, \\ \pi_{h_1 h_2} &= p(\alpha_{i,t-1} = \xi_{h_1}, \alpha_{it} = \xi_{h_2}), \quad h_1, h_2 = 1, \dots, k, \; t = 2, \dots, T. \end{aligned}$$

It is assumed that every α_{it} is conditionally independent of $\alpha_{i1}, \dots, \alpha_{i,t-2}$ given $\alpha_{i,t-1}$, but apart from this assumption, the distribution of $\boldsymbol{\alpha}_i$ is unconstrained except that to ensure identifiability we require that $\sum_h \pi_h = 1$ and $\sum h_2 \pi_{h_1 h_2} = 1$ and one component of the support point is constrained to be zero.

In such case a latent Markov (LM) model (Wiggins, 1973) with covariates results where the covariates affect the measurement model; see Bartolucci, Farcomeni and Pennoni (2013) for a review. The continuous latent process formulation as proposed by Bartolucci, Bacci and Pennoni (2014) assumes that the hidden response variables in $y_{i1}^*, \dots, y_{iT}^*$ are conditionally independent given $\boldsymbol{X}_i$ and the latent process $\boldsymbol{\alpha}_i = (\alpha_{it}, \dots, \alpha_{it})$. Another hypothesis is that every hidden variable and then every response variable, only depends on α_{it} and $\boldsymbol{x}_{it}$ and that the latent process $\boldsymbol{\alpha}_i$ has distribution given by a mixture of k AR(1) processes with common variance σ^2. According to the latter we assume the existence of a discrete latent variable u_i, for $i = 1, \dots, n$, having a distribution with k support points and mass probabilities $\pi_1, \dots, \pi_k$ such that when $u_i = h$ we assume that

$$\alpha_{i1} = \xi_h + \eta_{i1}, \quad i = 1, \dots, n,$$

and that

$$\alpha_{it} = \xi_h + (\alpha_{i,t-1} - \xi_h)\rho_h + \eta_{it}\sqrt{1 - \rho_h^2}, \quad i = 1, \dots, n, \; t = 2, \dots, T,$$

where $\eta_{it} \sim N(0, \sigma^2)$ for all i and t and (ξ_h, ρ_h) are parameters which for $h = 1, \dots, k$ are estimated jointly with the common variance. To ensure identifiability of the model, we require that $\xi_1 = 0$ or, $\sum_h \xi_h \pi_h = 0$. We observe that when $h = 1$, the model is the latent auto-regressive model proposed by Chi and Reinsel (1989) and Heiss (2008), when $h > 1$ it is the mixture latent auto-regressive model proposed by Bartolucci, Bacci and Pennoni (2014). In Table 3 we provide a summary of the parameters of the two proposed formulations.

4. Estimation details

In the following we briefly illustrate the estimation methods we employed for the two different model formulations. Given a sample of n independent units, the model log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i | \boldsymbol{X}_i)$$

Table 3. Description of the model parameters when the distribution of the latent process is discrete (top panel) or continuos (bottom panel).

Parameter	Description	Range
μ_j	Cut-point	$j = 2, \dots, J$
β_l	Regression coefficient in the model (1)	$l = 2, \dots, L$
π_h	Initial probability for state h	$h = 1, \dots, k$
$\pi_{h_1 h_2}$	Transition probability from state h_1 to state h_2	$h_1, h_2 = 1, \dots, k$
μ_j	Cut-point	$j = 2, \dots, J$
β_l	Regression coefficient in the model (1)	$l = 2, \dots, L$
ρ_h	Correlation coefficient of the mixture component	$h = 1, \dots, k$
ξ_h	Parameter for the latent structure	$h = 1, \dots, k$
σ^2	Common variance of the mixture components	

where $\boldsymbol{\theta}$ is the vector of all free parameters affecting $p(\boldsymbol{y}_i|\boldsymbol{X}_i)$. The latter is the manifest distribution of the response vector $\boldsymbol{y}_i$ given all the observable covariates $\boldsymbol{X}_i$. The model estimation is performed by the Expectation-Maximisation (EM) algorithm which is based on the complete data log-likelihood. When we are assuming the discrete latent formulation the complete data log-likelihood has expression

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Bigg\{ \sum_{t=1}^{T} \sum_{h=1}^{q} \sum_{y=0}^{J-1} a_{ih\boldsymbol{x}y}^{(t)} \log p(y_{it}|h, \boldsymbol{x}_{it}) + \sum_{h=1}^{k} b_{ih}^{(1)} \pi_{ih}, \\ + \sum_{t=2}^{T} \sum_{h_1=1}^{k} \sum_{h_2=1}^{k} b_{ih_1h_2}^{(t)} \pi_{ih_1h_2} \Bigg\},$$

where $b_{ih}^{(1)}$ is a dummy variable for unit i in component h at occasion t, with reference to the same occasion and the same unit, $b_{ih_1h_2}^{(t)}$ is a dummy variable equal to 1 if this unit moves from state h_1 to state h_2, whereas $a_{ih\boldsymbol{x}y}^{(t)}$ is equal to 1 if the unit is in state h and provide response y and covariate configuration $\boldsymbol{x}$. The conditional response probabilities $p(y_{it}|h, \boldsymbol{x}_{it})$ are computed efficiently by using some recursions known in this literature. For more details see Ch. 3 and Ch. 5 of Bartolucci, Farcomeni, Pennoni (2013).

Whereas when we are assuming the continuous latent formulation the manifest distribution of the response vector $\boldsymbol{y}_i$ given all the observable covariates $\boldsymbol{X}_i$ is expressed through a T-dimensional integral which is approximately computed by a quadrature method based on a series of q nodes properly chosen. The expression for $p(\boldsymbol{y}_i|\boldsymbol{X}_i)$ based on the quadrature is an approximation which depends on the number of integration points used. It is important to mention that it has the same expression of the manifest distribution of a latent Markov model based on q states. The nodes are taken on an equispaced grid of points, to which we refer to as v_m, $m = 1, \dots, q$ in the following. On the basis of this choice the complete data log-likelihood may be expressed as

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{h=1}^{k} w_{ih}\Bigg\{ \sum_{m=1}^{q}\sum_{t=1}^{T} z_{imt}\log p(y_{it}|\nu_m, \boldsymbol{x}_{it}) + \log\pi_h,$$

$$+ \sum_{m_1=1}^{q}\sum_{m_2=1}^{q}\sum_{t=2}^{T} z^*_{im_1m_2t}\log\omega^{(h)}_{m_1m_2}\Bigg\},$$

where $w_{ih} = I\{u_i = h\}$ is a dummy variable for unit i in component h, $z_{imt} = I\{\alpha_{it} = v_m\}$ is a dummy variable for unit i given the mth quadrature point for the integral with respect to α_{ij}, $z^*_{im_1m_2t} = z_{im_1,t-1}z_{im_2t}$ is a dummy variable for unit i given the m_2th quadrature point for the integral with respect to α_{it} given the m_1th quadrature point used for the integral with respect to α_{it-1} and $\omega^{(h)}_{m_1m_2}$ denotes the m_2th weight for the integral with respect to α_{it} given the m_1th quadrature point for the integral with respect to $\alpha_{i,t-1}$. The latter is computed as

$$\omega^{(h)}_{m_1m_2} = \frac{f^{(h)v_{m_2}|v_{m_1}}}{\sum_l f^{(h)v_{m_2}|v_{m_1}}}, \qquad m_1, m_2 = 1,\ldots,q \quad h = 1,\ldots,k.$$

The EM algorithm alternates the following steps until convergence: the E-step of the algorithm computes the conditional expected values of dummy variables given the observed data and the current parameter vector $\bar{\boldsymbol{\theta}}$; the M-step of the algorithm updates the model parameters by maximising the posterior probabilities. Since the EM algorithm is rather slow to converge, after a certain number of EM steps we switch to a full Newton-Raphson algorithm to maximise the model log-likelihood $\ell(\boldsymbol{\theta})$. From the EM algorithm we obtain the score vector as

$$\boldsymbol{s}(\boldsymbol{\theta}) = \boldsymbol{E}_{\bar{\boldsymbol{\theta}}=\boldsymbol{\theta}}[\boldsymbol{s}^*(\boldsymbol{\theta})|obs.data],$$

where the expected value is at the parameter value $\bar{\boldsymbol{\theta}}$ and $\boldsymbol{s}^*(\boldsymbol{\theta}) = \partial\ell^*(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is the score vector of the complete-data log-likelihood whose expected value is computed at the beginning of each M-step. We then compute the observed information matrix $\boldsymbol{J}(\boldsymbol{\theta})$ as the minus the numerical derivative of $\boldsymbol{s}(\boldsymbol{\theta})$ obtained as above. The standard errors for the parameter estimates are obtained from $\boldsymbol{J}(\boldsymbol{\theta})^{-1}$ in the usual way.

The selection of the appropriate number of latent states in the latent Markov model formulation is made relying on the BIC criterion (Schwarz, 1978), which is based on the index

$$BIC = -2\ell(\hat{\boldsymbol{\theta}}) + g\log(n) \tag{2}$$

where $\ell(\hat{\boldsymbol{\theta}})$ denotes the maximum log-likelihood of the model of interest and g is the number of parameters.

For the mixture latent auto-regressive model the selection of the number components is made first selecting the number of quadrature points and then selecting the number of mixtures according to the following strategy:

- for a given k, we try to increase values of q until the maximum of $\ell(\boldsymbol{\theta})$ does not significantly change with respect to the previous value of q. We start from q = 21 and increase it by 10 until the maximum of $\ell(\boldsymbol{\theta})$ is close to the previous value. The difference is taken less than 0.001;
- we select the number of states relying on the BIC criterion in (2) which takes into account the goodness-of-fit and the parsimony of the model.

It is important to mention that for both models we adopted a proper estimation procedure to prevent problems due to the multimodality of the likelihood function by applying a multi-start strategy combining a deterministic rule with a random starting rule. The numerical work was undertaken on the R (R Core Team, 2013) package `LMest` (Bartolucci, 2012) in an improved version which will be available from http://cran.r-project.org/. It provides maximum likelihood estimates of the model parameters and their corresponding standard errors in a reasonable amount of time.

Moreover on the basis of the final parameters estimates ($\hat{\boldsymbol{\theta}}$) we can compute the predictions of the entire sequence of latent states α_{it} for hospital i which corresponds to the maximum with respect to $h_1, \ldots, h_k$ of the posterior probabilities. For the case of the latent Markov model these are computed on the basis of the following expression:

$$\tilde{\alpha}_{it} = \sum_{h=1}^{k} = \hat{\pi}_h^* \hat{f}^{(h)} \quad i = 1, \ldots, n,\ t = 1, \ldots, T.$$

where $\hat{\pi}_h^*$ denotes the estimate of the stationary probability for the h latent state which depends on the transition probabilities $\hat{\pi}_{h_1,h_2}$ and $\hat{f}^{(h)}$ denotes the estimated posterior conditional distribution of the latent variables.

For the case of the mixture latent auto-regressive model these are computed as:

$$\tilde{\alpha}_{it} = \sum_{h=1}^{k} \sum_{m=1}^{q} (\widehat{w_{ih} z_{imt}})(\widehat{\xi}_h + \nu_m \hat{\sigma}), \quad i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{3}$$

where $\widehat{w_{ih} z_{imt}}$ is the posterior density that subject i moves from state m_1 to state m_2 at occasion t given that $\mu_j = h$.

5. Results

We applied the proposed model to the available data illustrated in Section 2. First we specify the discrete latent variable formulation and we estimate the latent Markov model with parameterization (1) on the measurement model for an increasing number of latent states. In Table 4 we show the results in terms of maximum log-likelihood and of BIC index on the basis of which we select $k = 4$ latent states.

Then we specify the continuous latent variable formulation and we estimate the mixture latent auto-regressive model for an increasing number of quadrature points for each number of mixture components. In Table 5 we report the results of this model selection procedure for the optimal number of quadrature points and the number of mixture components according with the selection strategy illustrated in Section 4.

Table 4. Maximum log-likelihood of the latent Markov model, values of the BIC index and number of parameters for $k = 1, \ldots, 5$.

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
log-likel.	-522.254	-418.018	-377.330	-355.128	-348.327
BIC	1082.674	888.514	830.991	**819.981**	849.315
g	8	11	16	23	32

Table 5. Log-likelihoods and differences between consecutive values for the mixture latent auto-regressive models with $k = 1, 2, 3$ and q from 21 to 111 with step 10; in boldface are the differences between maximum values of consecutive log-likelihoods which are smaller than 0.001 for the first time.

	$k = 1$		$k = 2$		$k = 3$	
q	log-likel.	diff.	log-likel.	diff.	log-likel.	diff.
21	-354.126	–	-351.142	–	-345.690	–
31	-356.444	-2.317	-356.115	-4.972	-347.193	-1.503
41	-360.820	-4.376	-355.536	0.579	-349.373	-2.180
51	-359.271	-1.549	-356.831	-1.295	-349.978	-0.609
61	-360.820	-1.549	-356.853	0.023	-349.982	-0.074
71	-360.820	-1.549	-356.853	0.023	-350.056	-0.003
81	-360.823	-0.005	**-356.854**	**-0.000**	-350.079	0.074
91	**-360.824**	**-0.001**	–	–	-350.077	-0.022
101	360.824	0.000	–	–	-350.080	0.002
111	–	–	–	–	**-350.080**	**0.000**

Table 6. Maximum log-likelihood of the mixture latent auto-regressive model for $k = 1, 2, 3$ and values of $q = 91, 81, 111$ respectively, values of BIC index and number of parameters.

k	$k = 1$	$k = 2$	$k = 3$
log-likel.	-360.824	-356.854	-350.080
BIC	**769.354**	775.726	776.484
g	10	13	16

The selected number of quadrature points is equal to 91 for $k = 1$, 81 for $k = 2$ and 111 for $k = 3$. In Table 6 with respect to choice made in Table 5 we show the values of

Table 7. Estimates of the latent Markov model with k = 4 (LM(4)) together with standard errors for regression coefficients and of the mixture latent auto-regressive model with k = 1 (MLAR(1)).

	LM(4)	MLAR(1)
$\hat{\mu}_1$	-47.073	-125.283
$\hat{\mu}_2$	-52.617	-139.245
$\hat{\mu}_3$	-58.841	-152.552
$\hat{\beta}_1$ beds	1.714	7.268
	(0.547)	(8.754)
$\hat{\beta}_2$ physicians	2.478	6.153
	(0.590)	(7.151)
$\hat{\beta}_3$ nurses	2.548	4.307
	(0.721)	(5.399)
$\hat{\beta}_4$ others	-1.259	-1.908
	(0.410)	(2.401)
$\hat{\beta}_5$ surgery rooms	0.137	1.079
	(0.283)	(1.727)

BIC for each value of k according to the chosen value of q. On the basis of the BIC value we select one mixture component. In Table 7 we show the estimates of the parameters referred to the cut-points and the regression coefficients in equation (1), together with the corresponding standard errors for both types of models to which we refer as LM and MLAR. We use the translog function for the covariates (Christensen, Jorgenson, and Lau, 1973). As noted by Bauer (2009) we have implicitly changed the scale of the latent response variable and therefore we get that the estimates of the thresholds and of the regression coefficients of the mixed latent auto-regressive model are on another scale with respect to that of the other model. On the basis of the t-statistics that may be computed for the regression coefficients, we conclude that the first four covariates are significant on the latent Markov model with four latent states. On the other hand while retaining the same sign all the covariates are not significant under the mixture latent auto-regressive model. The effect of the number of beds and of the working hours of physicians and nurses is positive, while the effect of working hours of the other staff of the hospital is negative, indicating that in the wards considered the main important features are the first three. We conclude that the dimension of the hospital has a positive effect on the efficiency and that the hospital staff which is not directly related with the treatment of the patient may contribute to inefficiency for the hospital. With the latent Markov model the estimated initial probabilities in each state are $\hat{\pi}_1 = 0.22$, $\hat{\pi}_2 = 0.31$, $\hat{\pi}_3 = 0.28$, $\hat{\pi}_4 = 0.19$. Under the mixture latent auto-regressive model with one component all the hospitals are in one class with high auto-correlation coefficient ($\hat{\rho}_1 = 0.911$) which is statistically significant (s.e. 0.361) and $\sigma^2 = 14.556$.

Table 8. Estimates of the transition probabilities $\pi_{h_1 h_2}$ under the LM(4) model.

	$\hat{\pi}_{h_1 h_2}$			
h_2	$h_1 = 1$	$h_1 = 2$	$h_1 = 3$	$h_1 = 4$
1	0.911	0.042	0.047	0.000
2	0.064	0.936	0.000	0.000
3	0.000	0.037	0.907	0.056
4	0.000	0.000	0.089	0.911

Due to the adopted parameterization the cut-points correspond to different levels of the propensity of the hospital to have high levels of efficiency. The value of the first cut-point is higher then the others therefore the first latent state corresponds to those hospitals with the highest propensity towards efficiency. The fourth latent state corresponds to those hospitals with the lowest propensity to be efficiency. Regarding the distribution of the latent process for the LM(4) model in Table 8 we report the estimates of the transition probabilities. Looking at the estimates of the parameters of the transition matrix we can see the evolution of the probabilities of each state and therefore we can dispose of a characterization of the pathways of each of the four groups. The matrix is not symmetric and the persistence in the same latent state for the entire period is high. The hospitals which have a medium/high level of efficiency i.e which are in latent state 2, in the previous year tend to become more efficient in the next year and those less efficient i.e. which are in latent state 4, in the previous year tend to be more efficient as time goes.

The predicted values of α_{it} according to (3) and (4) with respect to the time occasion t are showed in Figure 1 for both selected models. The single predicted profile trajectories are less regular under the LM(4) model rather than under the MLAR(1) model. This means that we can detect in a more appropriate way the changes observed in the hospital which are due to events which are not observed through the covariates. Such prediction may be also used in a way to correct in advance some opportunistic behaviours of the hospitals in a cost effective strategy. The predicted values can also be used to rank the hospitals according to best and worst performer in terms of potential efficiency gains.

6. Conclusions

We have shown a model specially tailored for longitudinal data having an ordinal structure with time-varying latent effects and covariates. The model accounts for two types of formulation of the latent effect. The first one assuming a discrete distribution gives rise to a latent Markov model which is not very complex to fit. It is more natural in many contexts and very suitable for classification even if the number of parame-

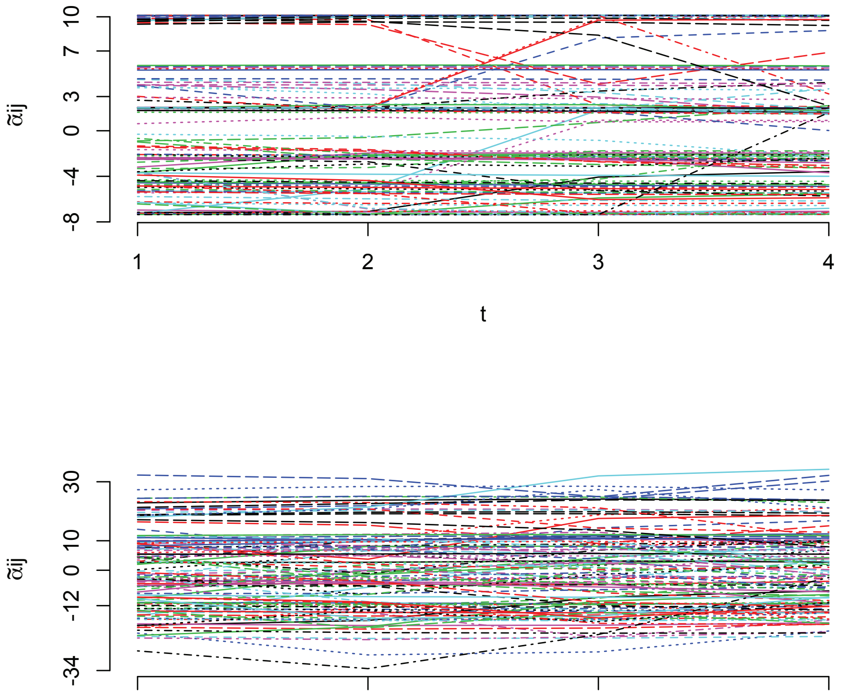


Figure 1. Predicted values of α_{it}, $t = 2, ..., T$ under the latent Markov model with $k = 4$ (top), the auto-regressive mixture model with $k = 1$ $(bottom)$.

ters increases with the number of latent states. The second model formulation relies on a continuous distribution for the unobserved heterogeneity. It gives rise to a mixture latent auto-regressive model which is more complex to fit as the distribution of the observed data given the covariates is obtained by solving a T dimensional integral. Maximum likelihood estimation of the model parameters is performed by a joint use of the Expectation-Maximization algorithm and of the Newton-Raphson algorithm. Standard errors for the parameter estimates are obtained by exploiting the observed information matrix. The number of latent states are selected by considering the BIC index for the latent Markov model and by an appropriate strategy for the mixture components.

To illustrate the proposal we applied the models to data related to the hospitals by considering a derived ordinal variable which accounts for the efficiency of the hospital spending policy observed over a four year period. The adopted parameterization on the response variable is based on global logits for each cut-point and it allows to measure the direct effect of each available covariate. When we estimate the model by relaying on a discrete distribution of the unobserved heterogeneity we select a latent Markov model with four latent states. The latter are clusters of hospitals sharing the same propensity towards efficiency gains. The covariates number of beds, yearly hours of activity of physicians and nurses are significant and positive, indicating the main important features of the hospital to get efficiency gains. The estimated transition matrix is useful to characterize pathways of the hospitals. When we estimate the model by relaying on a continuos distribution of the unobserved heterogeneity we select a mixture latent auto-regressive model with one component and none of the covariates are significant.

Finally, we show the prediction of the individual effect for every hospital at each time occasion on the basis of the parameter estimates. We notice that the predicted profile trajectories are less regular under the latent Markov model then under the mixed latent auto-regressive model. Therefore we conclude that the latent Markov model with four latent states allows us for less erratic trends of the hospital effects across time with respect to the other model formulation. To our knowledge the mixture latent auto-regressive formulation may be more promising when we dispose of many time occasions. The predicted values can also be used to rank the hospitals according to the best and the worst performer in term of potential efficiency gains as well as to correct in advance some opportunistic behaviours of the hospital.

Acknowledgements: We acknowledge “Finite mixture and latent variable models for causal inference and analysis of socio-economic data” (FIRB - Futuro in ricerca) funded by the Italian Government (RBFR12SHVV). F. Pennoni also thanks the financial support of the STAR project “Statistical models for human perception and evaluation”, University of Naples Federico II.

References

Aigner, D., Lovell, C. A. K. and Schmidt, P. (1977), Formulation and estimation of stochastic frontier production function models, *Journal of Econometrics*, 36, 21-37.

Bartolucci, F. (2012), University of Perugia, http://www.stat.unipg.it/ bart/, LMest: Fit latent Markov models in basic versions, R package version 1.0.

Bartolucci, F., Bacci, S., Pennoni, F. (2014), Longitudinal analysis of self-reported health status by mixture latent auto-regressive models, *Journal of the Royal Statistical Society: Series C*, 63, 267-288.

Bartolucci, F., Farcomeni, A., Pennoni, F. (2013), *Latent Markov Models for Longitudinal Data*, Chapman and Hall/CRC press.

Battese, G.E., Coelli, T.J. (1995), A model for technical inefficiency effects in a stochastic frontier production function for panel data, *Empirical Economics*, 20, 325–332.

Berta, P., Callea, G., Martini, G. , Vittadini, G. (2010), The effects of upcoding, cream skimming and readmissions on the Italian hospitals efficiency: a population-based investigation, *Economic Modelling*, 27, 789-890.

Bauer, D. J. (2009), A note on comparing the estimates of models for cluster-correlated or longitudinal data with binary or ordinal outcomes, *Psychometrika*, 74, 97-105.

Chi, E. and Reinsel, G. (1989), Models for longitudinal data with random effects and AR(1) errors, *Journal of the American Statistical Association*, 84, 452-459.

Christensen, L., Jorgenson, D., Lau, L. (1973), Transcendental logarithmic production frontiers, *Review of Economics and Statistics*, 55, 28–45.

Colombi, R., Forcina, A. (2001), Marginal regression models for the analysis of positive association of ordinal response variables, *Biometrika*, 88, 1007–1019.

Colombi, R., Kumbhakar, S., Martini, G., Vittadini G. (2014), Closed-skew normality in stochastic frontiers with individual effects and long/short run efficiency, *Journal of Productivity Analysis*, 1-14, doi: 10.1007/s11123-014-0386-y.

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

Herwartz, H., Strumann, C. (2014), Hospital efficiency under prospective reimbursement schemes: an empirical assessment for the case of Germany, *European Journal of Health Economics*, 15, 175-186.

Green, W. (2005), Reconsidering heterogeneity in panel data estimators of the stochastic frontier model, *Journal of Econometrics*, 126, 269–303.

Greene, W (2009), The econometric approach to efficiency analysis, in H. O. Fried, C. A. K. Lovell and S. S. Schmidt (Eds), *The measurement of productive efficiency techniques and applications*, Oxford University Press: Oxford, 92-251.

Kumbhakar, S. C., Lien, G. and Brian J. (2014), Technical efficiency in competing panel data models: a study of Norwegian grain farming, *Journal of Productivity Analysis*, 41, 321-337.

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

R Development Core Team (2013), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing.

Wiggins, L. M. (1973), *Panel Analysis: Latent Probability Models for Attitude and Behaviour Processes*, Elsevier, Amsterdam.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics*, 6, 461-464.

Modelling individual heterogeneity in ordered choice models: Anchoring Vignettes and the *Chopit* Model

Omar Paccagnella
Department of Statistical Sciences, University of Padua
E-mail: omar.paccagnella@unipd.it

Summary: This paper provides an overview of the approach of the Anchoring Vignettes, introduced ten years ago to analyse self-reported ordinal survey responses taking into account individual heterogeneity in the interpretation of the questions. The main characteristics and assumptions of this approach are reviewed, as well as two different statistical solutions introduced for dealing with vignette data. Special attention will be paid to the current discussion on some critical points and hints for future researches on this approach. An empirical application using information collected by some longitudinal vignettes within the Survey of Health, Ageing and Retirement in Europe (SHARE) will show the potentialities of this methodology.

Keywords: Anchoring vignettes; *Chopit* model; Individual heterogeneity; Ordered variables; Response scales.

1. Introduction

Vignettes have a long history in investigating social phenomena (Nosanchuck, 1972) and can be defined as "short descriptions of a person or a social situation which contain precise references to what are thought to be the most important factors in the decision-making or judgement-making process of respondents" (Alexander and Becker, 1978).

However, only ten years ago, King et al. (2004) introduced in the literature some statistical solutions exploiting the vignettes as an additional tool to identify and correct the systematic differences in the use of response scales within countries or socio-economic groups. This approach aims at making comparable, across respondents, self-evaluations affected by individual unobserved heterogeneity. Since the ratings of the vignette persons provide an *anchor* (a gold standard) for adjusting self-ratings, these instruments have been called *anchoring* vignettes.

The presence of individual heterogeneity leads respondents to interpret, understand or use the response categories for the same questions differently (Holland and Wainer, 1993). This evidence of response scale differences is known as *Differential Item Functioning* (DIF) in the educational testing literature or *Response Style* (RS) in the psychological setting (Paulhus, 1991).

In social and economic sciences there is a large use of subjective evaluations reported by survey respondents, from personal health to life satisfaction, from work disability to job satisfaction. Avoiding to take into account the (not-random) heterogeneity in reporting styles across different respondents may systematically bias the measurement of the variable of interest, obtaining misleading assessments of the relative performances in cross-cultural comparisons.

Despite the youthfulness of the King et al. (2004) contribution, the interest on the anchoring vignette approach has rapidly increased, both in terms of data collection and in terms of empirical applications. At the same time, a correct application of this instrument requires the fulfilment of some assumptions, which have not been adequately considered in many cases.

The aim of this paper is to provide a general overview of the anchoring vignette approach and its applicability, pointing out advantages and limitations, showing an empirical application and leaving some reflections and open questions for further debates. This contribution does not aim at providing evidence in favour or against the vignette assumptions. However, since this issue has now become a key research argument on this literature, we report a discussion of the state of art on the validity of the vignette assumptions.

The paper is organised as follows. In the next section the idea of the anchoring vignettes and how they work in practice to correct for DIF is introduced. Section 3 will briefly show the non-parametric and parametric solutions working with vignettes, while in Section 4 a detailed discussion on the main vignette assumptions is reported. An empirical application showing how the parametric solution and some extensions can help to correct for individual heterogeneity in reporting behaviour is presented in Section 5. Section 6 concludes the paper, summing up the main points of the current debate on vignettes and suggesting some directions for future researches.

2. *How do anchoring vignettes work?*

A growing number of socio-economic surveys is now designed to forcefully support cross-national or cross-population group researches. One the of main challenges conducting such studies in practice is the harmonisation of concepts, measures and survey designs across different social and cultural settings, in order to enhance data comparability.

However, in many areas (for instance, economics, health, psychology, etc.) methodologies and instruments for obtaining common measurements are not well understood; at the same time, the definition and the collection of quantitative objective measures are often hard to attain. For this reason, self-reported measures on the domain/aspect under investigation are widely adopted tools in cross-cultural studies. However, self-evaluation judgements are usually misled by respondents because they convey the real (but unobserved) value on the concept of interest plus DIF. For instance, self-reported data on life satisfaction are often used in the analysis of population well-being, even though there is wide evidence of large cross-country differences in the individual reported level of life satisfaction (Kahneman et al., 2004; Angelini et al., 2013). This happens because respondents use different benchmarks or scales in evaluating themselves, even when they are similar according to economic and non-economic conditions. To this aim, investigating self-evaluations of life satisfaction, Clark et al. (2005, page C118) note that these measures ”... are ordinal. A life satisfaction score of 6, on a scale of 1 to 7, does not correspond to twice as satisfied as a score of 3 . In this ordinal world, 6 only means more than 5 and less than 7... One worry regarding statistical analysis of subjective variables is that some people look at life pessimistically or optimistically, even though there is *really* no difference in their level of well-being. This *anchoring effect* or intercept heterogeneity is a source of potential bias...”.

Scale differences across respondents may be due to objective differences in the individual characteristics, as well as to different interpretations of numerical scales or cultural differences in the norms for what is called ”Very good”, ”Good”, etc. or ”Some say”, ”Little say”, etc.

Figures 1 provides an example on this heterogeneity of scale definitions across individuals. The domain under investigation is the measurement of political efficacy. In Figure 1, two types of individuals (i and j) interpret the same answer category differently in the upper and lower scales: respondent i turns the own (unobserved) level of the domain of interest $\tilde{y}'$ into the category ”Little say”, while respondent j turns the same latent location into the category ”Some say”. Alternatively, for the same unobserved level $\tilde{y}''$ respondent i would declare a self-assessment of ”Some say”, while ”Unlimited say” would be declared by respondent j. The notion of having ”Some say” in government is completely different between the two respondents. According to the self-ratings, respondent j has much more say than respondent i in getting the government to address issues that interest him/her, while according to the actual values there are no differences between them. Since the actual self-assessments provided by the two respondents are identical, the differences in the reporting evaluations must be due to DIF.

Anchoring vignettes are conceived as additional questions to answer by respondents, to adjust the self-evaluation in order to provide a DIF-free measurement of the concept of interest. This allows enhancing comparability (across countries or socio-economic groups) of the subjective assessments, because all of these evaluations are now reported to a common and DIF-free scale. As a rule, respondents rate themselves and many anchoring vignettes, which represent various levels of the trait of interest (i.e. life satisfaction or work disability). Each vignette is a brief text where an hypothetical individual, manifesting just the trait of interest, is depicted to a lower or higher level of severity.

Three examples of vignette questions (whose aim is to analyse work disability reporting) are listed below:

Kevin suffers from back pain that causes stiffness in his back especially at work but is relieved with low doses of medication. He does not have any pains other than this generalized discomfort.

Anthony generally enjoys his work. He gets depressed every 3 weeks for a day or two and loses interest in what he usually enjoys but is able to carry on with his day-to-day activities on the job.

Eve has had heart problems in the past and she has been told to watch her cholesterol level. Sometimes if she feels stressed at work she feels pain in her chest and occasionally in her arms.

How much is [Kevin/Anthony/Eve] limited in the kind or amount of work [he/she] could do?

These vignettes follow the self-evaluation question: *Do you have any impairment or health problem that limits the amount or kind of work you can do?* The same answer categories as for the self-rating are available in all vignette questions (1=None; 2=Mild; 3=Moderate; 4=Severe; 5=Extreme).

The basic idea on the vignette adjustment can be summarised by the example of Figure 2 (based on the measurement of political efficacy as in Figure 1), since it shows how the variation in the answers to three vignette questions provided by two different respondents may be used to construct a common scale of measurement across these individuals. The idea is to rescale the individual own self-rating on that common scale (the respondent 2's reported scale is deformed into one scale that is comparable to respondent 1's scale), taking into account the ordering of the individual responses.

The statistical solutions developed for using vignette data require two basic assumptions, *response consistency* and *vignette equivalence*. The first assumption is needed for connecting the individual ratings of the anchoring scenarios to the individual self-ratings: each respondent applies the same cutpoints for the self-evaluation as for the evaluation of the person depicted in the vignette question. The second assumption is essential for obtaining a DIF-free measurement of the variable of interest to be used as an anchor: each respondent perceives the (unobserved) level of the variable represented in any vignette in the same way (in other words, at same location on the latent scale). A detailed discussion on the meaning and the validity of each assumption is provided in Section 4.

The anchoring vignettes have been so far applied in several health and socio-economic domains, where subjective self-ratings reported by survey respondents are frequently used. Among them: political efficacy and visual activity (King et al., 2004); work disability (Kapteyn et al., 2007); health (Bago d'Uva et al., 2008; Grol-Prokopczyk et al., 2011a); health system responsiveness (Rice et al., 2012); job satisfaction (Kristensen and Johansson, 2008); life satisfaction (Angelini et al., 2013); satisfaction with social contacts (Bonsang and van Soest, 2012); marketing (Paccagnella, 2014).

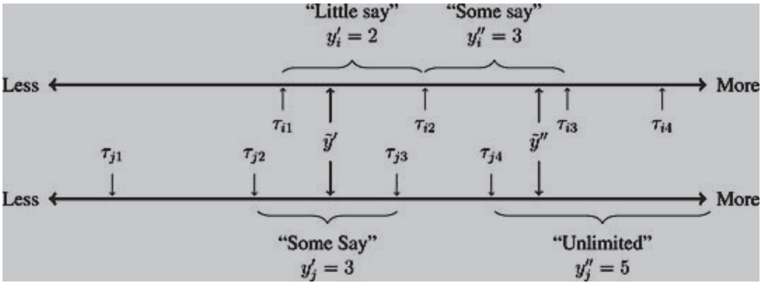


Figure 1. Turning the continuous unobserved level of self-ratings into reported ordinal categories for respondents i and j, subject to DIF. Source: Wand (2013).

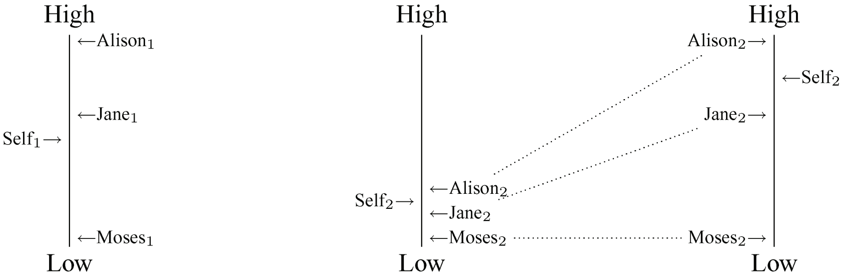


Figure 2. Rescaling self-ratings through the anchoring vignettes. Source: King et al. (2004).

3. The parametric and non-parametric solutions

The survey instruments and the measurement assumptions are combined together in order to introduce two different statistical solutions for this approach, a non-parametric and a parametric methodology (King et al., 2004; King and Wand, 2007).

3.1. The non-parametric solution

The non-parametric approach shows some important advantages: i) it is easy to implement; ii) it does not need any other assumptions than response consistency and vignette equivalence; iii) it does not need any explanatory variables. However, it also suffers from two main not-trivial disadvantages: i) each respondent has to answer to all vignette and self-reported questions (sometimes this can be difficult to administer); ii) it is statistically inefficient in some circumstances (even though this is typical of many non-parametric solutions).

At the beginning, these limitations have prevented the application of this solution in the empirical analysis. Since a few years, researchers have paid much more attention to the potentialities of the non-parametric solutions (Wand, 2013). On the one hand, a branch of the literature aims at comparing the performances of both parametric and non-parametric estimators to adjust for reporting heterogeneity (Jones et al., 2012). In general, both approaches lead to similar conclusions. On the other hand, the non-parametric solution could be exploited for testing the validity of the vignette assumptions (van Soest and Voňková, 2014). However, the literature on the non-parametric modelling is still lacking.

The idea of the non-parametric solution is to recode self-ratings relative to the set of vignettes. Self-evaluations are compared mapping them according to the scale fixed by the vignette evaluations in each country or socio-economic group.

Given the content of the vignettes (i.e. the severity of the problem depicted in each vignette scenario), the ranking of the vignette evaluations used by the majority of the individuals belonging to each country or group defines a *natural ranking* of vignettes for that country or group. For each respondent i (i=1,...,N), let Y_i be the categorical self-evaluation and Z_{ij} be the categorical evaluation of vignette j (j=1,...,J). Assuming that the natural ranking is $Z_{i,j-1} < Z_{ij}$, for all i, j, the adjusted (DIF-corrected) variable is given by:

$$C_i = \begin{cases} 1 & if \quad Y_i < Z_{i1} \\ 2 & if \quad Y_i = Z_{i1} \\ 3 & if \quad Z_{i1} < Y_i < Z_{i2} \\ \vdots & \vdots \\ 2J & if \quad Y_i = Z_{iJ} \\ 2J+1 & if \quad Y_i > Z_{iJ} \end{cases} \qquad (1)$$

The C-scale provides a new but DIF-free ordinal variable C_i, which can be studied by standard models for the analysis of ordinal data as ordered probit, contingency tables, etc.

In each data collection there could be cases where a respondent either evaluates the vignettes in a way different from the natural ranking, or provides the same rating to more vignettes. These inconsistencies are grouped and treated as ties and the loss of information due to them leads to the inefficiencies of this non-parametric solution. In such cases, C_i assumes a set of values rather than a single one.

Dealing with these ties represents a strong limitation to the implementation of this approach. King et al. (2004) suggest allocating C_i by assuming a uniform distribution for the values across the specified range; King and Wand (2007) extend the approach by developing a censored ordered probit modelling; Wand (2013) introduces the B-scale, which basically locates each self-rating relative to the average perceived location of the anchoring vignettes.

The B-scale provides an alternative solution to the C-scale, with the advantage of producing credible comparisons, asking for weaker assumptions. Indeed, the B-scale approach relaxes the vignette equivalence assumption. It requires the weaker condition that the perception of the vignettes are on the same side as the true location of the vignette relative to the respondent's own location (as an example, if a certain vignette depicts a hypothetical individual whose health is poorer than the respondent's health, the respondent cannot perceive this hypothetical individual to be healthier). Wand (2013) defines it as the *Order Preserving Imperfect Anchors* (OPIA) assumption. He also shows that inferences based on interpersonal comparisons may change according to the (stronger or weaker) assumptions that are invoked.

3.2. The parametric solution

The parametric approach has some important advantages: i) there is no need to collect the answers of all proposing vignettes for each respondent[1]; ii) it avoids the statistically inefficiencies of the non-parametric approach, recognizing that the variable of interest is perceived with a measurement error (i.e. idiosyncratic errors can explain violations of the natural ordering of the vignette evaluations provided by respondents); iii) thresholds are allowed to vary across respondents as a function of a set of observed variables. However, the parametric solution also suffers from some important disadvantages: i) it needs some additional assumptions other than response consistency and vignette equivalence, such as the specification of the function (only linear so far) relating the observed characteristics and the unobserved components, the functional shape of the thresholds and the distributional form of the error terms; ii) the response consistency

[1] Analysing longitudinal data or repeated cross-sections, researchers could include the vignettes on only some of the waves. However, investigating self-reported work disability among old people, Angelini et al. (2011) show that individual reporting styles are not stable over time.

assumption is crucial for identifying the parametric model using vignettes.

The parametric model is referred to as the Compound Hierarchical Ordered Probit (*chopit*) model, even though many researchers use the term *hopit* model equivalently. In a broad sense the *chopit* model can be seen as a generalisation of the standard ordered probit approach (even if the ordered probit is not formally nested into the *chopit* model), where DIF is modelled through variations in the thresholds.

The *chopit* model is basically composed by a self-assessment equation and as many vignette equations as the number of collected vignette questions.

In the original specification (King et al., 2004), the unobserved continuous response Y_i^*, which measures the perceived own level of the variable of interest (health status, life satisfaction, etc.) of respondent i, is modelled as follows:

$$\begin{array}{rcl} Y_i^* & = & X_i\beta + \varepsilon_i \\ \varepsilon_i & \sim & N(0,1) \end{array} \tag{2}$$

$$Y_i = k \qquad \textit{if} \qquad \tau_i^{k-1} \leq Y_i^* < \tau_i^k$$

$$-\infty = \tau_i^0 < \tau_i^1 < \ldots < \tau_i^K = \infty$$

$$\begin{array}{rcl} \tau_i^1 & = & \gamma^1 V_i \\ \tau_i^k & = & \tau_i^{k-1} + exp\left(\gamma^k V_i\right) \qquad k = 2, \cdots, K-1 \end{array} \tag{3}$$

It is assumed a linear combination of some individual observed variables X_i and an unobserved term component to model Y_i^*. With this specification, the error term ε_i reflects both individual unobserved heterogeneity and reporting error. The reported categories (Y_i) are obtained by means of a threshold model with individual-specific thresholds, that are modelled as a function of individual variables V_i (which may overlap the set of X_i variables). The exponential assumption in (3) guarantees that these thresholds increase with k.

The evaluations of J ($j = 1, ..., J$) vignettes by the same respondent i are modelled as follows:

$$\begin{array}{rcl} Z_{ij}^* & = & \theta_j + u_{ij} \\ u_{ij} & \sim & N(0, \sigma_u^2) \\ u_{ij} & \perp & (\varepsilon_i, X_i, V_i) \end{array} \tag{4}$$

$$Z_{ij} = k \qquad \textit{if} \qquad \tau_i^{k-1} \leq Z_{ij}^* < \tau_i^k$$

Here, Z_{ij}^* represents the unobserved perceived level provided by respondent i of the variable of interest described in vignette j. According to the vignette equivalence assumption, θ_js do not vary across respondents. According to the response consistency assumption, the thresholds τ_i^k are the same as the self-assessment equation (in other words, respondent i evaluates each vignette on the same scale as is used for the self-evaluation).

Some extensions of the standard *chopit* model have been already introduced in the literature.

Kapteyn et al. (2007) allow that threshold equations can vary both with a set of observed individual characteristics and with an individual unobserved heterogeneity term ξ_i. Equation (3) is then replaced by:

$$\begin{array}{rcll} \tau_i^1 & = & \gamma^1 V_i + \xi_i & \\ \tau_i^k & = & \tau_i^{k-1} + exp\left(\gamma^k V_i\right) & \quad k = 2, \cdots, K-1 \\ \xi_i & \sim & N(0, \sigma_\xi^2) & \\ \xi_i & \perp & (\varepsilon_i, u_{ij}, X_i, V_i) & \end{array} \tag{5}$$

The model specification with this unobserved heterogeneity term added in the threshold equations reduces substantially some misspecification problems of the original *chopit* model (van Soest and Voňková, 2014). The original King's et al. model is derived imposing the condition $\sigma_\xi^2 = 0$.

Kapteyn et al. (2007) also argue that respondents can perceive differently the vignette when the same hypothetical scenario describes a female instead of a male person. Consequently, respondents may use different thresholds when evaluating a vignette with a female name instead of a male name. This behaviour could potentially violate the response consistency assumption and for this reason Kapteyn et al. (2007) suggest to specify in the vignette equation a gender dummy variable of the scenario description. However, this model extension can be applied only when the survey plans to collect vignette data with a gender randomization of the individual depicted in each vignette.

Paccagnella (2011) extends the standard *chopit* model in order to account for sample selection bias. In cross-country comparisons of self-reported measures this problem may lead to inconsistent results, in particular when the country non-response patterns (due for instance to the lack of standardized fieldwork procedures) are not missing at random. This model adds a third component (the selection equation) to the self-assessment and vignette equations: the reported categories Y_i and Z_{ij} are observed only if a selection condition (S_i) applies. The selection rule is defined as:

$$\begin{array}{rcl} S_i^* & = & W_i \delta + \eta_i \\ S_i & = & \left\{ \begin{array}{ll} 1 & S_i^* > 0 \\ 0 & \textit{otherwise} \end{array} \right. \end{array}$$

where W_i is a set of exogenous variables (W_i may overlap X_i, even if it is a good practice to specify at least one exclusion restriction), δ is a vector of parameters to be estimated and η_i is an error term. A shared random effect ω_i is specified to introduce dependence between the three error terms:

$$\begin{array}{rcl} \varepsilon_i & = & \phi\omega_i + \psi_i \\ u_{ij} & = & \lambda_j \omega_i + \varsigma_{ij} \\ \eta_i & = & \omega_i + \nu_i \end{array}$$

where ω_i, ψ_i and ν_i are independently normally distributed with zero mean and unit variance and ς_{ij} is independently normally distributed with zero mean and variance σ_j^2. The reported categories and thresholds are defined as before.

Angelini et al. (2011) introduce a *longitudinal chopit* model, in order to investigate to what extent individual reporting styles are stable over time. This model specification allows to introduce time varying covariates. At the same time, unobserved terms are split in individual-specific time invariant components and idiosyncratic time varying errors. Formally, let t=1,2,...,T the time periods, Y_{it}^* be the perceived own level of the variable of interest by respondent i at time t and Z_{ijt}^* be the unobserved level of the variable of interest described in vignette j as perceived by individual i at time t. The *longitudinal chopit* model is defined as:

$$\begin{aligned} Y_{it}^* &= X_{it}\beta_t + \varepsilon_{it} \\ Z^*{}_{ijt} &= \theta_{jt} + u_{ijt} \end{aligned}$$

where X_{it} are time-variant observed variables, β_t is the vector of (time-variant) parameters to be estimated (without constant for identification) and θ_{jt} are vignette- and time-specific dummies. The error term for the self-assessment equation is defined as:

$$\begin{aligned} \varepsilon_{it} &= \eta_i + \omega_{it} \\ \eta_i &\sim N(0, \sigma^2) \\ \omega_{it} &\sim N(0, 1) \\ \eta_i &\perp \omega_{it} \\ \omega_{it} &\perp \omega_{is}, \quad t \neq s \end{aligned}$$

while the error term for the vignette equation is defined as

$$\begin{aligned} u_{ijt} &= \varsigma_{ij} + \nu_{ijt} \\ \varsigma_{ij} &\sim N(0, \rho^2) \\ \nu_{ijt} &\sim N(0, \tau_t^2) \\ \varsigma_{ij} &\perp \nu_{ijt} \\ \nu_{ijt} &\perp \nu_{ijs}, \quad t \neq s \\ \eta_i &\perp \varsigma_{ij} \end{aligned}$$

The reported categories are obtained through:

$$\begin{aligned} Y_{it} = k \quad if \quad \tau_{it}^{k-1} \leq Y_{it}^* \leq \tau_{it}^k, \quad k = 1, \cdots, K \\ Z_{ijt} = k \quad if \quad \tau_{it}^{k-1} \leq Z_{ijt}^* \leq \tau_{it}^k, \quad k = 1, \cdots, K \end{aligned}$$

where $-\infty = \tau_{it}^0 < \tau_{it}^1 < \ldots < \tau_{it}^K = \infty$ and, accordingly with Equation (3), the thresholds are modelled as a function of exogenous variables V_{it} and a vector of parameters γ_t:

$$\begin{aligned} \tau_{it}^1 &= \gamma_t^1 V_{it} \\ \tau_{it}^k &= \tau_{it}^{k-1} + exp\left(\gamma_t^k V_{it}\right) \qquad k = 2, \cdots, K-1 \end{aligned}$$

Standard *chopit* model or its extensions are estimated maximizing the log-likelihood function and integrating out the random terms.

4. Assumptions

4.1. Response Consistency

The response consistency assumption states that each respondent adopts the same thresholds τ_i^k for the self-ratings and the vignette evaluations. In other words, the way in which people evaluate themselves is equal to the way in which they evaluate other individuals.

The interpretation and the adoption of individual response scales are widely discussed in the scientific literature and deviations from the use of the same response scales in many domains is well documented. Justification bias (Bound, 1991) is an interesting example of this occurrence: for a given level of the “true health”, individuals who are not working understate their health in order to justify their (un)employment status. Investigating the role of the pain in some self-reported work disability comparisons, Banks et al. (2007) point out how pain can be both a subjective and an objective manifestation and, for this reason, respondents can react very differently in front of the same amount of pain. Moreover, in the psychological literature it is well know the self-enhancement bias, defined as “the tendency to describe oneself more positively than a normative criterion would predict” (Krueger, 1998).

Testing validity of this assumption is the topic of a growing empirical literature.

A branch of the literature follows the solution suggested by van Soest et al. (2011). On the one hand, they relax the response consistency assumption, allowing that response scales used to evaluate themselves could differ from the ones adopted for the vignette evaluation. On the other hand, the parameter identification relies on the availability of an objective indicator of the construct of interest, that are supposed to capture all variation in this construct associated with observed individual features. The indicator is defined as “objective” because it has to be unaffected by reporting heterogeneity, but driven by the same underlying latent process that produces self-ratings (the so-called one-factor assumption). This objective measure is modelled by means of a standard ordered probit model (in other words, in this model specification the thresholds between the categories are unknown constants). If the one-factor assumption holds, a formal test of response consistency can be obtained comparing the model imposing both the one-factor and the response consistency assumption with a model imposing the one-factor assumption only. van Soest et al. (2011) implement this approach to test the response consistency assumption investigating drinking behaviour in Ireland. They exploit the information collected on the number of drinks consumed by the respondents and their findings support the vignette assumption.

The same methodological solution is used by Datta Gupta et al. (2010) to test response consistency in the context of work disability vignettes among old Europeans. The grip strength measure is adopted as the objective indicator of work disability and their findings do not support the vignette assumption. However, while the number of drinks consumed by the respondents is arguably a good proxy of the self-reported drinking behaviour, the choice of a grip strength measure as an objective indicator is questionable in a multidimensional context like work disability.

Deriving a meaningful proxy which takes into account a variety of individual aspects in the domain of interest could be demanding.

To this aim, Bago d'Uva et al. (2011) extend the van Soest et al. (2011) approach, proposing tests of response consistency that rely on the assumption that a battery of objective indicators is available. They perform a *strong* and a *weak* test in the health domains of cognitive functioning and mobility, applied to a dataset of old English respondents. For mobility, both tests reject response consistency, while for the cognitive domain only the *strong* test rejects the assumption validity.

Grol-Prokopczyk et al. (2011b) analyse data coming from the WHO Study on Global AGEing and Adult Health (SAGE) on six non-European countries in the mobility and vision health domains. They support response consistency comparing cutpoints from vignette ratings and cutpoints generated from some objective measures of health.

van Soest and Voňková (2014) construct specification tests comparing non parametric rankings that come directly from the raw data with rankings implied by some parametric solutions. They apply them on six health domains (breathing, concentration, depression, mobility, bodily pains and sleep) of an aged population coming from eight European countries: while the standard *chopit* model always rejects response consistency, this is no longer valid for several health domains and socio-economic characteristics adopting the *chopit* model that incorporates unobserved heterogeneity in the threshold equation.

Using some qualitative analysis of the interview responses, Au and Lorgelly (2014) find that response consistency might not hold for a part of the sample respondents. However, they also provide some suggestions for making more consistent the validity of this assumption.

Kapteyn et al. (2011) introduce a new and interesting way to test this assumption. By means of an experiment conducted in the internet RAND American Life Panel, respondents are first asked to describe and rate their health and, in a subsequent interview, to evaluate some vignettes that are - in fact - descriptions of their health. Under some auxiliary assumptions related to the validity of the experiment, they analyse five health domains (sleep, mobility, concentration, breathing and affect) obtaining mixed evidence: according to a non-parametric approach, response consistency holds in the domain of sleep, while the other domains reject either the auxiliary assumptions or the response consistency assumption. The validity of the auxiliary assumptions creates much more problems adopting a parametric solution for this testing.

4.2. Vignette Equivalence

Vignette equivalence is the most controversial assumption on the vignette topic, so that it now represents the strongest criticism to this approach. This assumption states that the situations depicted in each vignette are perceived in the same way by all respondents, apart from a random measurement error. In other words, any differences across respondents in the perceived level of the variable represented in each vignette must be random and independent of the characteristic being measured.

Violation of the vignette equivalence assumption may occur more often than expected. For instance, in different countries the same scenario may be interpreted less or more problematic according to religion (i.e. mentioning suicide in Catholic vs non-Catholic populations), socio-economic status (i.e. considering a certain amount of earnings or unemployment benefits in countries with a more developed welfare state vs countries with less advanced states), health status (i.e. quoting obesity in malnourished vs well-nourished people) and so on.

Hence, it is not surprising there is a general agreement on that, before any analysis or validation testing, vignette wording may be a key for improving vignette equivalence: cultural- or linguistic-specific references have to be avoided in order to guarantee a logically coherent and consistent meaning in different cultures. On the other hand, the validity of the vignette equivalence is also strictly related to the domain under examination.

Voňková and Hullegie (2011) investigate how the approach is sensitive to the domain and the choice of the vignette for three domains of health (cognitive functioning, breathing and mobility) among old European respondents. The vignette approach indicates an important sensitivity to the choice of the vignette for cognition and a weak sensitivity for breathing, while the approach is basically not sensitive for mobility. Jürges and Winter (2013) use data from a survey experiment conducted on a large sample of older US individuals to study the effect of vignette names and their connotations on ratings of some vignettes in the domain of mobility. They find that vignette ratings may be sensitive to the gender of the person described in the vignette (in line with the reasoning of Kapteyn et al., 2007), as well as to the age implied by the first name (but for older respondents only). Cognitive functioning and mobility domains for old English respondents have been investigated also by Bago d'Uva et al. (2011) in their testing of the vignette equivalence validity. They test a necessary condition for this assumption, to be applied when at least two vignettes are available and under the hypothesis that response consistency assumption holds. Choosing a certain vignette as reference and specifying in the other vignette equation(s) that a set of individual characteristics impacts on the vignette evaluation, the necessary condition for vignette equivalence states there is no systematic variation in the perceived difference between the levels of the variable of interest represented by any two vignettes. Finding statistically significant variables provides evidence of the presence of systematic differences in the perception of a vignette *relative to the reference*, therefore violating the assumption. Their results are against the vignette equivalence validity. It is worth noting the findings from this set of studies: within the same health domain (mobility), there is evidence of a strong, mild or no rejection of the vignette equivalence assumption with respect to an old population of, respectively, English, US or European (excluding English) respondents.

There is a growing empirical literature devoted to the administration of the vignette questionnaire, for understanding effects due to priming, the ordering of vignettes and self-assessments (Buckley, 2008; Hopkins and King, 2010). Reversing the order of the questions (vignettes first and self-assessment question then, instead of the traditional order where self-ratings are asked at the beginning) may allow for a better correction of DIF. However, randomisation strategies in the order of the questions are suggested as means to reduce question order biases.

Rice et al. (2011) adopt different strategies and employ both non-parametric and parametric methods to assess the validity of the vignette equivalence assumption. Their results do not contradict this assumption. A test based on the global ordering of vignettes is also used by Angelini et. al (2013) in supporting vignette equivalence.

Using a large dataset coming from the WHO Study on Global AGEing and Adult Health (SAGE) and the World Health Survey (WHS) on ten international countries and eight health domains (mobility, affect, pain, social relationships, vision, sleep, cognition, and self-care), Grol-Prokopczyk et al. (2011b) support vignette equivalence assumption according to a test based on the global ordering of vignettes, while reject its validity according to the test, as described above, introduced by Bago d'Uva et al. (2011).

Ferrer-i-Carbonell et al. (2011) and Peracchi and Rossetti (2013) provide other evidence against vignette equivalence. However, in both cases, the model specification diverges from the usual *chopit* modelling, particularly for the former contribution. Differently from the other contributions, Peracchi and Rossetti (2013) examine the validity of the vignette equivalence by means of a joint test of the overidentifying restrictions implied by both vignette assumptions. They also find that these overidentifying restrictions are less likely to be rejected using only one vignette (among all available) or performing them separately by subgroups of respondents.

In the end, vignette equivalence assumption can be weakened analysing panel data on vignettes (Angelini et al., 2011). Indeed, when longitudinal information is available, the unobserved component in the vignette equation can be split in an *individual* specific random effect and an idiosyncratic error. This allows to relax the standard vignette equivalence assumption and therefore asking that the situation depicted in each vignette is *on average* perceived in the same way by each respondent.

5. Using vignettes to correct for DIF: an empirical application

This section presents a small application of the parametric solution to correct for the differences in the reporting scales. We do not test the validity of the vignette assumptions. Indeed, on the one hand, the goal of this empirical application is to show just the potentialities of this approach by means of a counterfactual example: a benchmark (e.g. the scale of a particular country) may be defined and the adjusted distributions of the variable of interest, based on the benchmark scale instead of the respondent own scale, are computed for all respondents. The adjusted measures of the self-evaluation reporting are on a common scale, purged from DIF, hence easily to be compared. On the other hand, we aim at presenting how some extensions of the *chopit* model may change the final results, once we take into account additional information related to the survey design and the data collection.

To this aim, we use vignette data drawn from the first and second waves of SHARE (Survey of Health, Ageing and Retirement in Europe), collected after a CAPI survey by means of paper and pencil questionnaires in 2004/05 and in 2006/07, respectively. SHARE is a panel survey that collects detailed data on health, socio-economic status and social and family networks of citizens aged 50 and over from 19 European countries.

In particular, we focus on the three longitudinal vignette questions that involve work disability measures. The same respondents are asked to rate twice the presence and severity of problems reducing their working abilities, as well as those of the hypothetical individuals described in the vignettes. The self-assessment and the vignette questions asked in both waves are exactly the ones reported in Section 2.

For our purpose, we keep only those respondents who have answered to both the self-reported and *all* vignette questions in both waves. As a consequence, the final sample is composed by 1265 individuals, coming from 8 countries (Sweden, Germany, the Netherlands, Belgium, France, Spain, Italy and Greece). Respondents are mainly females (56.6%) and 56 years old on average.

An individual is defined as work disabled if he/she has a ”moderate”, ”severe” or ”extreme” impairment or health problem that limits the kind or amount of work he/she can do.

According to the above definition, the percentage of individuals who rate themselves in the whole sample as work disabled (*unadjusted measures*) is about 16%. *Adjusted measures* are carried out after the *chopit* model (or its extensions) estimates, by means of some counterfactual exercises, that is calculating the percentage of all respondents, regardless of the country of belonging, who report work disability if they used the wave1 (2004/05) Italian thresholds instead of their own thresholds.

Self-ratings are therefore compared with the adjusted (DIF corrected) measures obtained from: i) the baseline solution (i.e. a *chopit* model not allowing for any threshold variation. This is a model specification very close to the standard ordered probit solution); ii) the standard *chopit* model (King et al., 2004); the *chopit* model accounting for individual heterogeneity in the thresholds (Kapteyn et al., 2007); the *chopit* model accounting for sample selection (Paccagnella, 2011); the longitudinal *chopit* model (Angelini et al., 2011).

Parameter estimations of each aforementioned model are reported from Tables 1 to 5 in the Appendix. In all models we control for the same set of variables, including:

Demographic characteristics: Gender (male as reference), age in classes (younger than 55 years as reference), education (no or low, middle and high, with "No or low education" as reference category) and household size.

Cognitive abilities: Word list learning test result - immediate recall (in this test, respondents are required to learn a list of ten common words and then will be asked to recall as many words as possible).

Depression and risk behaviour: Dummy variables equal to one when respondent is depressed, according to the EURO-D depression scale, overweighted and obese.

Physical health status: Maximum grip strength result and dummy variables equal to one when respondent has two or more chronic diseases, two or more symptoms, two or more limitations with mobility and arm function, one or more limitations with activities of daily living (ADL), one or more limitations with instrumental activities of daily living (IADL).

Country dummies: Italy as reference category.

Exclusion restrictions: (only for the model accounting for sample selection) Interviewer gender (male as reference) and interviewer age.

Cross-country differences clearly appear in each *chopit* model estimate. Moreover, results show the presence of some individual-specific thresholds: country dummies, physical health status, cognitive abilities and being obese are in general the most important determinants of these individual thresholds.

Figure 3 compares across the different model specifications the proportion of work disabled respondents, according to the previous definition of individual work disability.

Without controlling for any individual characteristics (in other words, the unadjusted measures), the rate of work disability is larger than 15%, overall. As soon as we control for some individual characteristics, the rates of work disabled respondents among European elderly substantially reduce, so that they become lower than 10% considering any model that allows for threshold variations (that is, correcting for DIF), as if all respondents used the wave1 (2004/05) Italian thresholds. In the analysed sample, there are no differences adopting a standard *chopit* model or the one that accounts for individual heterogeneity in the thresholds. Moreover, there are some interesting differences on the rates of work disability limitations taking into account sample selection and longitudinal answers[2].

[2] While in Paccagnella (2011) sample selection originates when respondents complete the CAPI interview, but do not fill in the paper and pencil vignette questionnaire, in this dataset sample selection also involves attrition from one wave to another. At the same time, differently from Angelini et al. (2011), vignette data collected in Greece and Sweden in wave2 (2006/07) are available and hence included in the dataset.

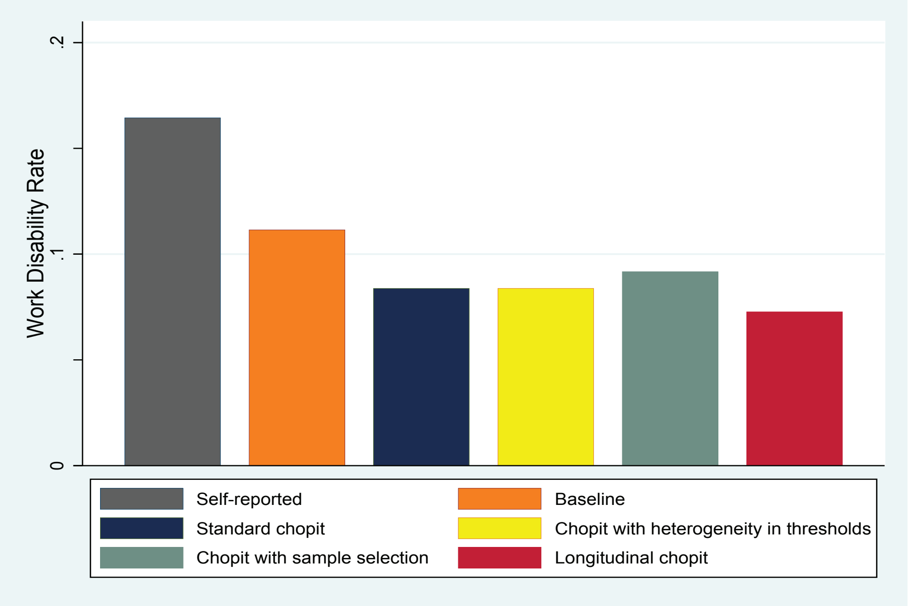


Figure 3. Comparing work disability rates among some European countries, using com-mon wave1 Italian thresholds.

6. Final discussion

Since its introduction ten years ago, the approach of the anchoring vignettes has shown a steadily increase of interest in the literature. Vignettes have been collected in several surveys and according to many domains. They have been applied successfully in many contexts to correct for differential item functioning and investigate the role played by response scale heterogeneity across respondents in different countries or socio-economic groups. The empirical application in this paper supports these findings.

Nevertheless, several authors suggest using vignettes with caution and, at the same time, testing validity of their assumptions first, rather than simply assumed them. However, so far vignette assumptions can be tested only using additional assumptions or restrictions to the *chopit* approach. Therefore, it cannot be fully known whether the vignette assumptions are rejected because they do not really hold or because the additional test assumptions are not indeed valid.

What is the future of this approach?

The first line of research undoubtedly comprises vignette assumptions testing. The research community needs some other instruments or methodologies to test their validity and/or still more evidence against or in favour them, according to the parametric or non-parametric adopted solutions or according to the domains of interest. As an example, health domains like mobility and cognitive functioning have been widely investigated (providing mixed evidence), while socio-economic domains are checked to a less extent. Moreover, many open questions require some answers yet. Among them, I would like to prompt: *Are vignette assumptions sensitive to the...*

i) number of the available response categories?

ii) number of proposed vignette questions?

iii) order of the vignette questions (random or according to the severity of the problems depicted in each vignette)?

iv) order of all questions (posing first self-evaluation and then vignette questions or vice-versa)?

Strictly connected to the previous one, a second line of research should investigate the issues related to the vignette writing. There is a general agreement that the process of writing vignettes has important effects for measuring the concept of interest and preventing from the rejection of the vignette assumptions. In his website on vignettes[3], Gary King even likens the process of writing vignettes to the process of testing a theory.

A third line of research has to be devoted to the improvement of the statistical solutions able to work with vignette data, as suggested by van Soest and Voňková (2014): the development of new extensions of the *chopit* model, allowing for more flexible parametric and semiparametric solutions (i.e. relaxing the linearity functional form in the self-assessment equation); new ways to introduce individual unobserved heterogeneity in the model specification; other hypotheses on the distributional form of the errors (i.e. error terms could be heteroscedastic).

In the end, anchoring vignettes are sometimes seen as a tool to enhance cross-national comparability only: differences in national cultures and/or in welfare, labour market or health-care systems across countries can violate the validity of the vignette assumptions. Actually, anchoring vignettes could be applied fruitfully to correct for differences in reporting behaviour in two other contexts at least: i) in longitudinal analysis (as in Angelini et al., 2011), where the within-person information can be exploited to weaken the vignette assumptions; ii) in comparisons of socio-economic strata of some populations, for a better adjustment of reporting heterogeneity with respect to socio-demographic characteristics, like gender, age classes or education.

Acknowledgements: This research was partially supported by the University of Padua (Italy) with grant CPDA121180 “Statistical and econometric approach to marketing: applications and developments to customer satisfaction and market segmentation”.

[3] http://gking.harvard.edu/vign

References

Alexander, C.S., and Becker, H.J. (1978). The use of vignettes in survey research, *Public Opinion Quarterly*, **42**, 93–104.

Angelini, V., Cavapozzi, D., Corazzini, L. and Paccagnella, O. (2013). Do Danes and Italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases, *Oxford Bulletin of Economics and Statistics*, forthcoming, on-line available at "http://onlinelibrary.wiley.com/doi/10.1111/obes.12039/abstract".

Angelini, V., Cavapozzi, D. and Paccagnella, O. (2011). Dynamics of reporting work disability in Europe, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**, 621–638.

Au, N. and Lorgelly, P.K. (2014). Anchoring vignettes for health comparisons: an analysis of response consistency, *Quality of Life Research*, forthcoming, on-line available at "http://link.springer.com/article/10.1007/s11136-013-0615-2".

Bago d'Uva, T., Lindeboom, M., O'Donnell, O., and van Doorslaer, E. (2011). Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity, *Journal of Human Resources*, **46**, 875–906.

Bago d'Uva, T., van Doorslaer, E., Lindeboom, M. and O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities?, *Health Economics*, **17**, 351–375.

Banks, J., Kapteyn, A., Smith, J. and van Soest, A. (2007). Work disability is a pain in the *****, especially in England, the Netherlands, and the United States, in: D. Cutler and D. Wise (eds.): *Health in Older Ages: The Causes and Consequences of Declining Disability Among the Elderly*, Chicago: University of Chicago Press.

Bonsang, E. and van Soest, A. (2012) Satisfaction with social contacts of older Europeans, *Social Indicators Research*, **105**, 273–292.

Bound, J. (1991). Self-reported versus objective measures of health in retirement models, *The Journal of Human Resources*, **26**, 106–138.

Buckley, J. (2008). Survey context effects in anchoring vignettes, *Working Paper*, available from http:// polmeth.wustl.edu/media/Paper/surveyartifacts.pdf.

Clark, A.E., Etil, F., Postel-Vinay, F., Senek, C. and Van der Straeten, K. (2005). Heterogeneity in Reported Well-Being: Evidence from Twelve European Countries, *Economic Journal*, **115**, C118–C132.

Datta Gupta, N., Kristensen, N. and Pozzoli, D. (2010). External validation of the use of vignettes in cross-country health studies, *Economic Modelling*, **27**, 854–865.

Ferrer-i-Carbonell, A., Van Praag, B.M.S. and Theodossiou, I. (2011). Vignette equivalence and response consistency: The case of job satisfaction, Discussion Paper No. 6174, *IZA Discusssion Paper Series*.

Grol-Prokopczyk, H., Freese, T. and Hauser R.M. (2011a). Using Anchoring Vignettes to Assess Group Differences in General Self-Rated Health, *Journal of Health and Social Behavior*, **52**, 246–261

Grol-Prokopczyk, H., McEniry, M. and Verdes, E. (2011b). Categorical borders across borders: Can anchoring vignettes identify cross-national differences in health-rating style?, *Proceedings of 2011 Population Association of America Meeting*, Washington, DC, March 31 - April 2, 2011.

Holland, P. and Wainer H. (1993). *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale, NJ.

Hopkins, D. and King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability, *Public Opinion Quarterly*, **74**, 201–222.

Jones, A.M., Rice, N. and Robone, S. (2012). A comparison of parametric and non-parametric adjustments using vignettes for self-reported data, WP12/10, *HEDG Working Paper Series*, University of York.

Jürges, H. and Winter, J. (2013). Are Anchoring Vignette Ratings Sensitive to Vignette Age and Sex?, *Health Economics*, **22**, 1–13.

Kahneman, D., Krueger, A.B., Schkade, D.A., Schwarz, N. and Stone, A.A. (2004). Towards National Well-Being Accounts, *American Economic Review*, **94**, 429–434.

Kapteyn, A., Smith, J.P. and van Soest, A. (2007). Vignettes and self-reports of work disability in the United States and the Netherlands, *American Economic Review*, **97**, 461–473.

Kapteyn, A., Smith, J.P., van Soest, A. and Voňková, H. (2011). Anchoring Vignettes and Response Consistency, WR-840, *RAND Labour and Population Working Paper Series*.

King, G., Murray, C., Salomon, J. and Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research, *American Political Science Review*, **98**, 191–207.

King, G. and Wand, J. (2007). Comparing incomparable survey responses: New tools for anchoring vignettes, *Political Analysis*, **15**, 46–66.

Krueger, J. (1998). Enhancement bias in descriptions of self and others, *Society for Personality and Social Psychology Bulletin*, **24**, 505–516.

Kristensen, N. and Johansson, E. (2008). New evidence on cross-country differences in job satisfaction using anchoring vignettes, *Labour Economics*, **15**, 96–117.

Nosanchuck, T.A. (1972). The vignette as an experimental approach to the study of social status: an exploratory study, *Social Science Research*, **1**, 107–120.

Paccagnella, O. (2011). Anchoring vignettes with sample selection due to nonresponse, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**, 665–687.

Paccagnella, O. (2014). A New Tool for Measuring Customer Satisfaction: the Anchoring Vignette Approach, *Italian Journal of Applied Statistics*, forthcoming.

Paulhus, D.L. (1991). Measurement and control of response bias, in: J.P. Robinson, P.R. Shaver and L.S. Wrightsman (eds): *Measures of Personality and Social Psychological Attitudes*, San Diego, CA: Academic Press, 17–59.

Peracchi, F. and Rossetti, C. (2013). The heterogeneous thresholds ordered response model: identification and inference, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **176**, 703–722.

Rice, N., Robone, S. and Smith, P.C. (2011). Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness, *European Journal of Health Economics*, **12**, 141–162.

Rice, N., Robone, S. and Smith, P.C. (2012). Vignettes and health systems responsiveness in cross-country comparative analyses (with discussion), *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **175**, 337–369.

van Soest, A., Delaney, L., Harmon, C., Kapteyn, A. and Smith, J.P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**, 575–595.

van Soest, A. and Voňková, H. (2014). Testing the specification of parametric models by using anchoring vignettes, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **177**, 115–133.

Voňková, H. and Hullegie, P. (2011). Is the anchoring vignettes method sensitive to the domain and choice of the vignette?, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**, 597–620.

Wand, J. (2013). Credible comparisons using interpersonally incomparable data: Nonparametric scales with anchoring vignettes, *American Journal of Political Science*, **57**, 249–262.

Appendix A. Tables

Table 1. Parameter estimates of the baseline model.

Variable	β	γ^1	γ^2	γ^3	γ^4
Self-assessment equation					
Germany	0.504***	-	-	-	-
Sweden	0.284*	-	-	-	-
The Netherlands	0.197	-	-	-	-
Spain	0.148	-	-	-	-
France	-0.038	-	-	-	-
Greece	-0.671***	-	-	-	-
Belgium	0.296**	-	-	-	-
Female	-0.460***	-	-	-	-
Age: 55-59	-0.085	-	-	-	-
Age: 60+	-0.045	-	-	-	-
Middle education	0.004	-	-	-	-
High Education	-0.025	-	-	-	-
Household size	-0.243	-	-	-	-
Overweight	0.144*	-	-	-	-
Obese	0.178*	-	-	-	-
Chronic diseases	0.513***	-	-	-	-
Symptoms	0.439***	-	-	-	-
Mobility limitat.	0.453***	-	-	-	-
ADL	0.495***	-	-	-	-
IADL	0.442***	-	-	-	-
Grip strength	-0.188***	-	-	-	-
Depression	0.369***	-	-	-	-
Ten words test	-0.444*	-	-	-	-
Constant	-	-0.216	-0.114***	-0.330***	-0.277***
Vignette equation					
Kevin	0.914***				
Anthony	0.701**				
Eve	1.307***				
Female vignette	-0.032				

*Significance levels: *** = 1% level; ** = 5% level; * = 10% level*

Table 2. Parameter estimates of the standard chopit model.

Variable	β	γ^1	γ^2	γ^3	γ^4
Self-assessment equation					
Germany	0.595***	0.068	-0.048	0.136	0.291*
Sweden	-0.097	-0.208*	-0.696***	-0.047	0.464***
The Netherlands	0.286*	-0.035	0.387***	-0.142	0.103
Spain	-0.044	-0.318***	0.161	0.177*	0.812***
France	0.081	0.119	-0.109	0.156*	0.333**
Greece	-0.670***	-0.003	-0.095	-0.027	0.171
Belgium	0.448***	0.059	0.192**	-0.004	0.374**
Female	-0.422***	0.001	0.133*	-0.018	-0.137
Age: 55-59	-0.045	0.012	0.008	0.098*	-0.099
Age: 60+	-0.029	0.003	-0.007	0.112*	-0.008
Middle education	0.014	-0.002	0.016	0.056	-0.025
High Education	-0.008	0.004	0.003	-0.002	-0.002
Household size	-0.172	0.049	0.082	0.480**	-0.160
Overweight	0.153	0.038	-0.080	0.007	-0.063
Obese	0.265**	0.157***	-0.133*	-0.114*	-0.193**
Chronic diseases	0.474***	-0.092*	0.091*	0.110**	0.071
Symptoms	0.424***	-0.078	0.155***	-0.048	0.182**
Mobility limitat.	0.509***	0.096**	-0.142***	0.053	-0.001
ADL	0.507***	-0.043	0.122	-0.066	-0.061
IADL	0.405***	-0.151	0.139	0.081	-0.032
Grip strength	-0.186***	-0.009	0.036	-0.013	-0.003
Depression	0.391***	0.064	-0.129**	0.034	0.050
Ten words test	-0.665**	-0.334**	0.360**	0.190	0.028
Constant	-	-0.041	-0.560***	-0.613***	-0.386
Vignette equation					
Kevin	0.890***				
Anthony	0.665**				
Eve	1.301***				
Female vignette	-0.040*				

*Significance levels: *** = 1% level; ** = 5% level; * = 10% level*

Table 3. Parameter estimates of the chopit with heterogeneity in thresholds model.

Variable	β	γ^1	γ^2	γ^3	γ^4
Self-assessment equation					
Germany	0.599***	0.104	-0.088	0.098	0.229
Sweden	-0.094	-0.186	-0.666***	-0.085	0.403***
The Netherlands	0.272	-0.035	0.383***	-0.137	0.104
Spain	0.001	-0.264**	0.119	0.144	0.775***
France	0.063	0.111	-0.088	0.137	0.270*
Greece	-0.700***	0.003	-0.094	-0.041	0.123
Belgium	0.444***	0.065	0.182*	-0.015	0.330**
Female	-0.417***	0.006	0.119	-0.011	-0.121
Age: 55-59	-0.041	0.015	0.007	0.096*	-0.097
Age: 60+	-0.030	-0.003	0.007	0.101*	-0.013
Middle education	0.043	0.023	0.011	0.041	-0.034
High Education	0.006	0.011	-0.002	-0.011	0.004
Household size	-0.194	0.027	0.090	0.520**	-0.191
Overweight	0.162*	0.046	-0.090*	0.001	-0.041
Obese	0.245**	0.134**	-0.119*	-0.088	-0.165*
Chronic diseases	0.481***	-0.086	0.088	0.104**	0.060
Symptoms	0.429***	-0.070	0.137**	-0.030	0.201**
Mobility limitat.	0.530***	0.113***	-0.163***	0.040	0.001
ADL	0.518***	0.015	0.033	-0.044	-0.037
IADL	0.391***	-0.138	0.142	0.052	-0.031
Grip strength	-0.191***	-0.016	0.044	-0.011	0.002
Depression	0.393***	0.067	-0.123**	0.014	0.023
Ten words test	-0.567**	-0.281*	0.304*	0.199	-0.007
Constant	-	-0.020	-0.543***	-0.636***	-0.397*
Vignette equation					
Kevin	0.957***				
Anthony	0.740**				
Eve	1.354***				
Female vignette	-0.046**				

*Significance levels: *** = 1% level; ** = 5% level; * = 10% level*

Table 4. Parameter estimates of the chopit with sample selection model.

Variable	β	γ^1	γ^2	γ^3	γ^4
Self-assessment equation					
Germany	0.831***	0.351***	-0.093	0.090	0.206
Sweden	-0.079	-0.186	-0.670***	-0.116	0.358***
The Netherlands	0.300*	-0.015	0.381***	-0.121	0.131
Spain	0.061	-0.182	0.125	0.137	0.796***
France	0.422**	0.486***	-0.092	0.132	0.274*
Greece	-0.823***	-0.171*	-0.091	-0.043	0.118
Belgium	0.470***	0.085	0.191**	-0.002	0.348**
Female	-0.416***	0.008	0.108	0.002	-0.118
Age: 55-59	-0.078	-0.022	-0.001	0.099*	-0.096
Age: 60+	0.067	0.109	-0.003	0.096	-0.017
Middle education	0.044	0.037	0.001	0.045	-0.032
High Education	-0.074	-0.066	0.008	-0.010	0.002
Household size	-0.143	0.073	0.080	0.532**	-0.212
Overweight	0.199**	0.089	-0.078	-0.003	-0.048
Obese	0.248**	0.127*	-0.115*	-0.090	-0.183*
Chronic diseases	0.461***	-0.102*	0.082	0.105**	0.075
Symptoms	0.413***	-0.081	0.123**	-0.027	0.216**
Mobility limitat.	0.495***	0.083	-0.147***	0.037	-0.002
ADL	0.576***	0.082	0.042	-0.053	-0.033
IADL	0.414***	-0.108	0.121	0.041	-0.055
Grip strength	-0.202***	-0.029	0.040	-0.009	0.004
Depression	0.378***	0.048	-0.122**	0.013	0.007
Ten words test	-0.682**	-0.355**	0.334**	0.218	0.054
Constant	-	0.025	-0.525***	-0.614***	-0.336
Selection equation					
Germany	-0.717***	-	-	-	-
Sweden	0.034	-	-	-	-
The Netherlands	-0.013	-	-	-	-
Spain	-0.237**	-	-	-	-
France	-1.064***	-	-	-	-
Greece	0.610***	-	-	-	-
Belgium	-0.043	-	-	-	-
Female	0.016	-	-	-	-
Age: 55-59	0.132**	-	-	-	-
Age: 60+	-0.317***	-	-	-	-
Middle education	-0.085	-	-	-	-
High Education	0.210***	-	-	-	-
Household size	-0.068	-	-	-	-
Overweight	-0.175***	-	-	-	-
Obese	0.036	-	-	-	-
Chronic diseases	0.074	-	-	-	-
Symptoms	0.054	-	-	-	-
Mobility limitat.	0.075	-	-	-	-
ADL	-0.187	-	-	-	-
IADL	0.046	-	-	-	-
Grip strength	0.043	-	-	-	-
Depression	0.027	-	-	-	-
Ten words test	0.181	-	-	-	-
Female interviewer	0.139***	-	-	-	-
Age interviewer	0.001	-	-	-	-
Constant	-0.480**	-	-	-	-
Vignette equation					
Kevin	0.280				
Anthony	0.038				
Eve	0.579*				
Female vignette	-0.053***				

*Significance levels: *** = 1% level; ** = 5% level; * = 10% level*

Table 5. Parameter estimates of the longitudinal chopit model.

Variable	β	γ^1	γ^2	γ^3	γ^4
Self-assessment equation					
Germany w1	0.800***	0.105	-0.051	0.072	0.217
Sweden w1	-0.058	-0.193	-0.676***	-0.184	0.402***
The Netherlands w1	0.318	-0.053	0.396***	-0.186*	0.163
Spain w1	0.050	-0.327**	0.129	0.104	0.688***
France w1	0.138	0.220*	-0.124	0.075	0.315**
Greece w1	-0.739***	0.033	-0.093	-0.105	0.098
Belgium w1	0.551***	0.077	0.207**	-0.080	0.327**
Female w1	-0.329**	0.034	0.133**	-0.059	-0.088
Age: 55-59 w1	-0.029	0.019	0.025	0.068	-0.070
Age: 60+ w1	-0.011	-0.046	0.028	0.057	0.063
Middle education w1	-0.020	-0.028	0.038	0.050	-0.047
High Education w1	-0.019	0.002	0.028	-0.004	0.013
Household size w1	-0.031	0.018	0.185	0.320	0.062
Overweight w1	0.217*	0.067	-0.082	-0.042	-0.055
Obese w1	0.336**	0.179**	-0.079	-0.139**	-0.171*
Chronic diseases w1	0.616***	-0.073	0.061	0.104**	0.115
Symptoms w1	0.556***	-0.054	0.102*	-0.026	0.200**
Mobility limitat. w1	0.576***	0.154**	-0.127**	0.026	-0.014
ADL w1	0.500**	0.049	0.105	-0.153	-0.075
IADL w1	0.584***	-0.110	0.074	0.096	-0.058
Grip strength w1	-0.182***	-0.025	0.054**	-0.030	0.018
Depression w1	0.358***	0.003	-0.071	-0.008	0.051
Ten words test w1	-0.753**	-0.503**	0.392**	0.107	0.128
Germany w2	0.373*	-0.203	0.081	-0.077	0.825***
Sweden w2	0.131	0.012	-0.114	-0.316***	0.369*
The Netherlands w2	0.083	-0.018	0.247***	0.040	-0.007
Spain w2	-0.133	-0.045	-0.315***	-0.162*	0.148
France w2	-0.258	0.103	-0.021	-0.025	0.379*
Greece w2	-0.072	0.307***	-0.224***	-0.125	-0.268
Belgium w2	0.305	0.208*	0.101	-0.194**	0.442**
Female w2	-0.291**	0.061	0.071	-0.078	0.284**
Age: 55-59 w2	-0.356***	-0.170***	0.033	-0.099*	-0.156
Age: 60+ w2	-0.097	-0.191***	0.133**	-0.030	-0.100
Middle education w2	0.113	-0.027	0.159***	-0.080	-0.294**
High Education w2	0.079	0.151**	-0.078	0.028	0.025
Household size w2	0.343	-0.411	0.212	0.747***	1.229**
Overweight w2	-0.144	0.048	-0.112**	0.025	-0.087
Obese w2	0.065	0.077	-0.119*	-0.059	-0.228*
Chronic diseases w2	0.440***	-0.001	0.040	-0.065	0.003
Symptoms w2	0.389***	-0.070	0.078	0.019	-0.198
Mobility limitat. w2	0.639***	0.014	-0.041	0.070	0.089
ADL w2	0.589**	-0.040	0.059	-0.021	-0.084
IADL w2	0.444**	0.140	-0.192**	0.021	0.163
Grip strength w2	-0.172***	-0.056*	0.051**	-0.039	0.039
Depression w2	0.483***	0.063	-0.054	0.063	-0.176
Ten words test w2	-0.180	-0.251	0.353**	0.228	-0.218
Constant	-	0.362	-0.410***	-0.010	-0.294
Vignette equation					
Kevin w1	1.563***				
Anthony w1	1.250***				
Eve w1	2.109***				
Kevin w2	1.125***				
Anthony w2	0.819**				
Eve w2	1.599***				
Female vignette	-0.051*				

*Significance levels: *** = 1% level; ** = 5% level; * = 10% level*

An IRT-MIMIC model for the analysis of university student careers

Bruno Bertaccini
Department of Statistics, Informatics, Applications 'G. Parenti' University of Florence
E-mail: brunob@disia.unifi.it

Leonardo Grilli
Department of Statistics, Informatics, Applications 'G. Parenti' University of Florence
E-mail: grilli@disia.unifi.it

Carla Rampichini
Department of Statistics, Informatics, Applications 'G. Parenti' University of Florence
E-mail: rampichini@disia.unifi.it

Summary:

The paper focuses on the analysis of the performance of university students, with reference to first year compulsory courses. The main goal is to compare the exams in terms of difficulty, discrimination and use of the grades. Moreover, the paper aims at assessing how student careers depend on student and course characteristics. The analysis exploits an Item Response Theory approach where exams are treated as items, with a 2-Parameter Logistic model for the probability to pass the exams and a Graded Response Model for the ordinal items representing grades of passed exams. Course characteristics, such as the average student rating on teacher's clarity, directly affect the items, whereas student characteristics, such as the type of high school, indirectly affect the items via the latent ability, even if some direct effects are allowed by fitting a MIMIC model with Differential Item Functioning. The analysis shows that IRT-MIMIC modelling is a flexible and powerful tool giving insights into the peculiarities of the exams and the role of course and student characteristics.

Keywords: 2PL model; academic performance; course evaluation; DIF; Graded Response Model.

1. Introduction

In this paper we analyse the careers of university students during the first year, focusing on compulsory courses. In particular, we consider the probability of passing the exams and the grades obtained in passed exams.

The analysis refers to freshmen enrolled in academic year 2011/2012 at two degree programmes of the School of Economics of the University of Florence, namely Management and Economics. For these students we consider exams of first year compulsory courses passed from January to December 2012.

The main goal of the paper is to compare the exams in terms of difficulty, discrimination and use of the grades. Moreover, the paper aims at assessing how student careers depend on student and course characteristics. This is important for both student tutoring and course organization.

The analysis exploits an Item Response Theory (IRT) approach where exams are treated as items and the student ability is a latent variable. Course characteristics, such as the average student rating on teacher's clarity, directly affect the items, whereas student characteristics, such as the type of high school, indirectly affect the items via the latent ability (standard MIMIC model), even if some direct effects are allowed by fitting a MIMIC model with Differential Item Functioning (DIF). We exploit two versions of this model: first, we analyse the binary items for exams passed or not passed by specifying the IRT part as a 2-Parameter Logistic (2PL) model; then, we analyse the ordinal items for the grades of passed exams by specifying the IRT part as a Graded Response Model (GRM).

Even if IRT-MIMIC modelling is a well-established method in Psychometrics, its application in the analysis of student careers is unconventional and challenging. In this regard, our analysis shows that IRT-MIMIC modelling is a flexible and powerful tool giving insights into the peculiarities of the exams and the role of course and student characteristics.

The paper is organized as follows. Section 2 describes the structure of the two degree programmes under consideration and summarizes the collected data on student careers and course evaluations. Section 3 outlines the model for passing the exams and the model for the grades. Section 4 presents the results of the analysis, and Section 5 offers some concluding remarks.

2. Data

The dataset for the analysis is obtained from the administrative archive on student careers, which includes background characteristics and information on passed exams. The dataset is then enriched with student ratings, which are anonymous and summarized at the course level. The final dataset has one record for each of the 7 compulsory courses. The number of students is 808, yielding 5656 records.

Table 1. First year compulsory courses of the degree programmes in Economics (EC) and Management (MG). University of Florence, A.Y. 2011/2012

Course	Degree prog.	Credits	n. of classes	Enrolled students	Teacher's clarity	Exam	
						%passed	Avg. score
Management	EC, MG	9	4	808	8.28	53.8	25.84
Accounting	EC, MG	9	3	808	8.75	53.1	23.33
Economics MG	MG	6	2	368	6.04	32.6	24.69
Economics EC	EC	9	2	440	7.87	17.5	23.65
History	EC, MG	6	3	808	7.83	63.1	22.73
Private law	EC, MG	9	3	808	7.79	17.2	23.81
Math MG	MG	6	3	368	6.81	21.2	21.79
Math EC	EC	9	2	440	7.62	28.4	23.62
Statistics	EC, MG	9	4	808	8.14	34.2	23.54

The characteristics of the courses are summarized in Table 1. Five courses are common to the two degree programmes, whereas Mathematics and Economics differ between the two degree programmes in terms of both credits and content. All the courses have parallel classes offered to groups of students defined by the first letter of the surname; each class has its own teacher with corresponding student ratings, that are regularly collected to monitor course quality. The questionnaire on student satisfaction is filled before taking the exam through a web system which first requires authentication and then ensures anonymity. Students express ratings on a ten-point scale on several aspects of the course, including the item “Teacher’s clarity” considered here. Table 1 reports the mean rating for each course, averaging over classes.

The last two columns of Table 1 report the percentage of enrolled students who passed the exam within the end of the first year, and the average score for successful students. A student is not forced to enroll in all the compulsory exams within the first year, thus if an exam has not been passed, it can be that either the student did not enroll or the student failed. Unfortunately, failures at the exams are not registered, thus it is not possible to know the reason why an exam has not been passed.

Exams are scored with integer values ranging from 18 to 30, plus ‘30 with honors’, which is scored 31 for the computation of the average.

3. *Model specification*

In order to analyse student careers, the exams of compulsory courses of the first year are treated as a multivariate response. We carry out the analysis in two steps: (1) a model for the probability to pass each exam; (2) a model for the probability to obtain a given grade at each passed exam. In both steps, we view each course as an item of a test in order to exploit IRT modelling (Baker and Kim 2004, Rijmen et al. 2003). In this approach, the items of a student are correlated by means of a latent variable, that can be interpreted as the student ability to pass the exam or to obtain a high grade.

3.1. Model for passing the exams

Let Y_{in}^{P} be a binary variable taking the value 1 if the exam of course i has been passed (hence the superscript P) by student n, $i = 1, \ldots, 9$ and $n = 1, \ldots, 808$. As shown in Table 1, the offered courses are 9, but each student is enrolled only in 7 of them, depending on her degree programme. The response model for the i-th course is:

$$logit[P(Y_{in}^{P} = 1 \mid \mathbf{X}_{in}, \theta_{n}^{P})] = \alpha_{i}^{P} + \boldsymbol{\beta}^{P}\mathbf{X}_{in} + \lambda_{i}^{P}\theta_{n}^{P} \tag{1}$$

where $\mathbf{X}_{in}$ is a vector of covariates with fixed effects $\boldsymbol{\beta}^{P}$, including characteristics of the course and, possibly, course-student interactions. The parameter α_{i}^{P} represents the easiness of the course when $\mathbf{X}_{in} = \mathbf{0}$. The latent variable θ_{n}^{P} can be interpreted as the student ability to pass his/her seven compulsory exams, with discrimination parameter λ_{i}^{P}. The discrimination parameter of the first course (*Management*) is fixed to one for identifiability, i.e. $\lambda_{1}^{P} = 1$.

If covariates $\mathbf{X}_{in}$ are not included, model (1) reduces to a standard 2-parameter logistic model (2PL).

In order to model the relationship between the ability to pass the exams and the observed characteristics of the student, we specify the following structural model:

$$\theta_{n}^{P} = \boldsymbol{\delta}^{P}\mathbf{Z}_{n} + \varepsilon_{n}^{P} \tag{2}$$

where $\mathbf{Z}_{n}$ is a vector of student characteristics, with fixed effects $\boldsymbol{\delta}^{P}$. The residual terms ε_{n}^{P} are assumed to be independent with an identical normal distribution with zero mean and standard deviation τ^{P}. Equations (1) and (2) define a MIMIC model (Jöreskog and Goldberger 1975), belonging to the class of Generalized Linear Latent And Mixed Model (Rabe-Hesketh et al. 2004*a*).

In the standard version, the MIMIC model assumes that student characteristics $\mathbf{Z}_{n}$ affect the probabilities to pass the exams only through the ability, i.e. $\mathbf{Z}_{n}$ only have indirect effects. However, it is interesting to investigate if $\mathbf{Z}_{n}$ also have direct effects. This can be done by including in $\mathbf{X}_{in}$ interaction terms among course indicators (dummy variables) and some of the covariates in $\mathbf{Z}_{n}$. Those interactions can be interpreted as item bias or differential item functioning (DIF, Rijmen et al. 2003).

The model defined by equations (1) and (2) is represented by the path diagram of Figure 1: the mean rating of teacher's clarity is included in the vector of covariates $\mathbf{X}_{in}$ affecting the probabilities to pass exams; on the other hand, high school type (lyceum vs others) and high school grade are included in the vector of covariates $\mathbf{Z}_{n}$ influencing student ability, with DIF for type of high school.

3.2. Model for exam grades

Exams are scored with integer values ranging from 18 to 30, plus '30 with honors'. The observed distribution of exam scores, reported in Figure 2, shows peaks at

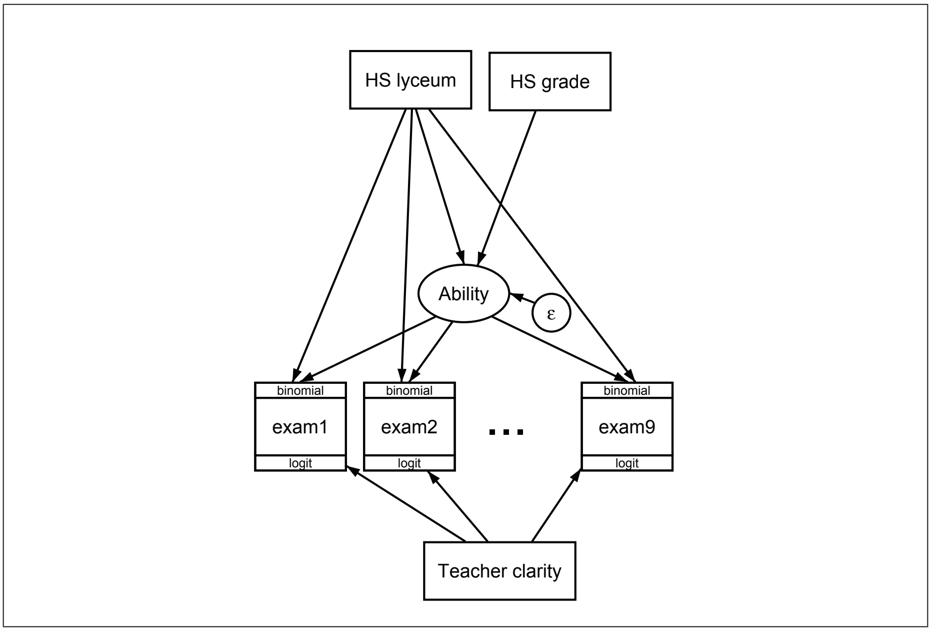


Figure 1. Path diagram of the IRT-MIMIC model (1) and (2), with DIF for 'HS lyceum'.

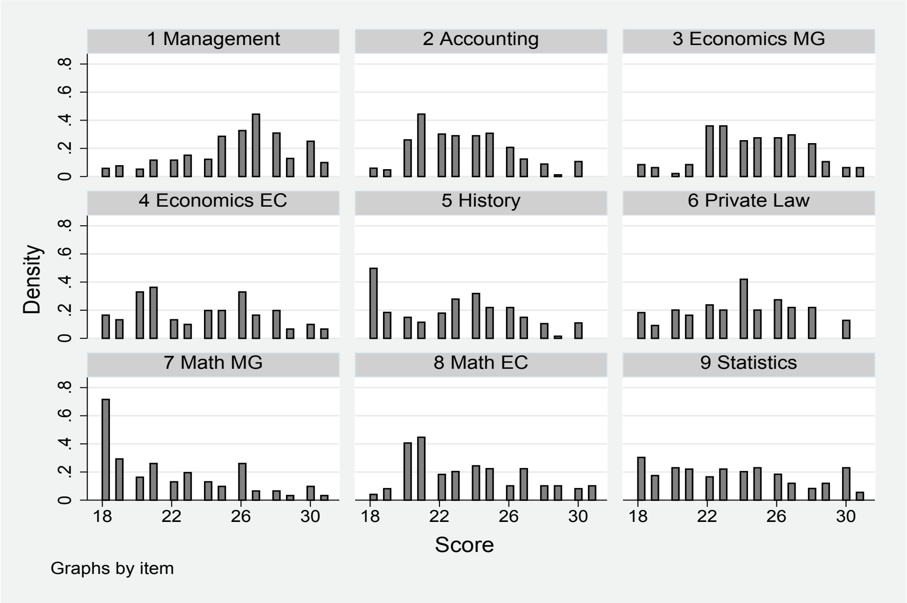


Figure 2. Exam scores of first year compulsory courses, degree programmes in Economics (EC) and Management (MG), University of Florence, A.Y. 2011/2012

the extremes and an irregular use of the scale, e.g. the score 29 is rarely assigned. For these reasons, a linear model would be markedly inappropriate, thus an ordinal response model is required.

Let Y_{in}^{G} be the grade for the exam of course i of student n, with $Y_{in}^{G} = c$ if the student passed the exam with grade c (hence the superscript G). To avoid sparseness, we define the exam grade by aggregating adjacent scores as reported in Table 2, so that Y_{in}^{G} is an ordinal variable with 7 categories.

The response Y_{in}^{G} is not defined if the exam has not been passed, i.e. if $Y_{in}^{P} = 0$, thus the number of observations reduces to 2189 grades from 615 students (in fact, about 24% of the students did not pass any exam).

We specify a cumulative logit model, corresponding to an IRT Graded Response Model (Samejima 1969). Since the ordinal variable has 7 categories, the model for the i-th course ($i = 1, \ldots, 9$) is defined by the following 6 equations:

$$logit[P(Y_{in}^{G} \leq c \mid \mathbf{X}_{in}, \theta_{n}^{G})] = \gamma_{ic}^{G} - \left(\boldsymbol{\beta}^{G}\mathbf{X}_{in} + \lambda_{i}^{G}\theta_{n}^{G}\right), \qquad c = 1, \ldots, 6 \tag{3}$$

where γ_{ic}^{G} is the threshold for the c-th category and $\mathbf{X}_{in}$ is a vector of covariates with fixed effects $\boldsymbol{\beta}^{G}$, including characteristics of the course and, possibly, course-student interactions. Due to the minus before the linear predictor in equation (3), the latent variable θ_{n}^{G} can be interpreted as the student ability to achieve high grades, with dis-

Table 2. Exams passed by grade. First year compulsory courses, degree programmes in Economics and Management, University of Florence, A.Y. 2011/2012

Grade	Score	Freq.	Percent	Cum.
1	18	196	8.95	8.95
2	19 – 21	453	20.69	29.65
3	22 – 23	374	17.09	46.73
4	24 – 25	427	19.51	66.24
5	26 – 27	390	17.82	84.06
6	28 – 29	187	8.54	92.60
7	30, 30 'with honors'	162	7.40	100.00
Total exams passed		2189	100.00	

crimination parameter λ_i^G. As in Sect. 3.1, the discrimination parameter λ_1^G for the first course (*Management*) is set to one for identifiability. In order to estimate all the item-specific thresholds (six for each item), we omit the overall constant from the linear predictor.

Similarly to the model for passing the exams of Sect. 3.1, we specify the following structural model to account for the indirect effects of student characteristics:

$$\theta_n^G = \boldsymbol{\delta}^G \mathbf{Z}_n + \varepsilon_n^G \tag{4}$$

where the vector of student characteristics $\mathbf{Z}_n$ has fixed effects $\boldsymbol{\delta}^G$. The residual terms ε_n^G are assumed to be independent with an identical normal distribution with zero mean and standard deviation τ^G. Equations (3) and (4) define a MIMIC model for the grades.

As discussed in Sect. 3.1, we can look for direct effects of student characteristics (DIF) by including in the vector $\mathbf{X}_{in}$ interaction terms among course indicators (dummy variables) and some of the covariates in $\mathbf{Z}_n$.

The path diagram of model (3) and (4) has the same structure of the IRT-MIMIC model represented in Figure 1; the only difference is that the items are ordinal, thus the response distribution is multinomial with cumulative logit link.

4. Results of the analysis

The models have been fitted using the `gllamm` command of Stata (Rabe-Hesketh et al. 2004*b*), which performs Maximum Likelihood estimation with Adaptive Gaussian Quadrature (we selected 8 quadrature points). To simplify the notation, from now on we omit the superscript P or G in model parameters.

4.1. Results: model for passing the exams

Table 3 reports the estimates of four specifications of the IRT-MIMIC model defined by equations (1) and (2) for the probabilities of passing compulsory exams.

Table 3. Estimates of models for the probabilities of passing first year compulsory exams (5656 exams, 808 students).

	2PL				MIMIC			
Parameter	standard		covariates		without DIF		with DIF	
Fixed effects								
α_1 Management	0.24		0.01		-0.21		-0.14	
α_2 Accounting	0.20		-0.09		-0.37		0.37	
α_3 Economics MG	-2.21	***	-2.08	***	-2.40	***	-2.29	***
α_4 Economics EC	-4.01	***	-4.18	***	-4.60	***	-4.94	***
α_5 History	0.82	***	0.65	***	0.46	**	0.52	***
α_6 Private law	-2.66	***	-2.83	***	-3.07	***	-2.98	***
α_7 Math MG	-2.50	***	-2.46	***	-2.72	***	-3.15	***
α_8 Math EC	-1.39	***	-1.55	***	-1.82	***	-3.87	***
α_9 Statistics	-1.60	***	-1.82	***	-2.13	***	-2.35	***
β_1 Teacher's clarity			0.18	*	0.19	*	0.21	*
β_2 Accounting×lyc							-1.39	***
β_3 Economics EC×lyc							1.35	*
β_4 Math MG×lyc							1.01	**
β_5 Math EC×lyc							3.34	***
β_6 Stat×lyc							0.75	**
Discrimination								
λ_1 Management	1.00	fixed	1.00	fixed	1.00	fixed	1.00	fixed
λ_2 Accounting	1.28	***	1.28	***	1.30	***	1.70	***
λ_3 Economics MG	1.63	***	1.68	***	1.75	***	1.77	***
λ_4 Economics EC	1.81	***	1.82	***	1.87	***	1.76	***
λ_5 History	0.79	***	0.79	***	0.78	***	0.79	***
λ_6 Private law	0.94	***	0.95	***	1.00	***	0.99	***
λ_7 Math MG	1.01	***	1.02	***	1.09	***	1.05	***
λ_8 Math EC	0.83	***	0.84	***	0.92	***	1.14	***
λ_9 Statistics	1.43	***	1.45	***	1.49	***	1.44	***
Structural model								
δ_1 HS grade					0.10	***	0.10	***
δ_2 HS lyceum					1.31	***	1.13	***
τ Ability SD	2.33	***	2.32	***	1.90	***	1.92	***
n. of parameters	18.00		19.00		21.00		26.00	
$logL$	-2681.82		-2679.10		-2579.78		-2506.59	

legend: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

We first consider the simple 2PL model without covariates. For a student with average ability, the easiest exam is *History*, and the hardest one is *Economics EC*. The probability to pass an exam depends on the ability of the student through the discrimination parameter: for example, *Statistics* has estimated discrimination $\widehat{\lambda}_9 = 1.43$, namely for a given increase in the student ability, the logit of the probability of success increases 43% more for *Statistics* than for the reference course, i.e. *Management* ($\lambda_1 = 1$). Figure 3 reports the item characteristic curves, describing how $P(Y_{in} = 1)$ depends on student ability (in standard scale, i.e. θ_n/τ): note that, due to different discriminations, the ranking of the exams in terms of difficulty varies across the range of student ability.

The estimated standard deviation of student ability is large ($\widehat{\tau} = 2.33$). For exam-

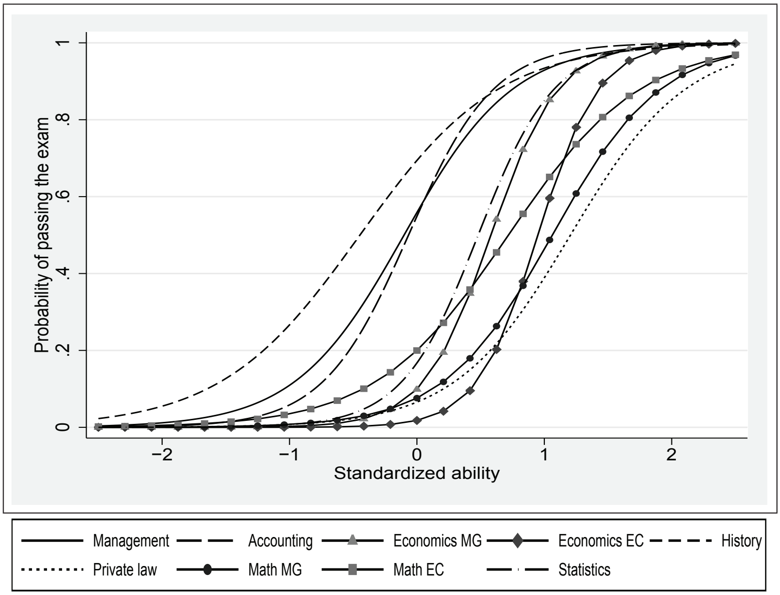


Figure 3. Standard 2PL item characteristic curves

ple, considering the exam of *Statistics*, a student with average ability ($\theta_n = 0$) has a probability of success equal to 17%, while if the ability increases by one standard deviation (i.e., $\theta_n = 2.33$) the probability of success becomes 85%, obtained by inverting equation (1): $1/[1 + \exp(-\widehat{\alpha}_9 - \widehat{\lambda}_9\theta_n)] = 1/[1 + \exp(1.60 - 1.43 \times 2.33)] = 0.85$. It is worth to note that the large value of the estimated standard deviation of the ability is a consequence of the high percentage of students who did not pass any exam (24%). Indeed, if those students are removed from the dataset, the estimated standard deviation reduces to 1.05.

Column *2PL with cov.* of Table 3 refers to the 2PL model including the average student rating on teacher's clarity, centered on the value 7, which is the integer closest to the observed midrange. This covariate is specific for each teacher, so it takes a different value for each class of the course. Its effect is significant and positive: the higher the teacher's clarity, the higher the probability to pass the exam. For example, considering the exam of *Statistics*, a student with average ability has a probability of success equal to 17% if the course has a mean rating of 7; this probability raises to 23% if the course rating is 9, while it goes down to 12% if the course rating is 5.

The MIMIC models reported in Table 3 are designed to account for the effect of student characteristics on the ability. High school grade (centered on the mid-point 80) and high school type have significant effects: students with a better grade and coming from a Lyceum have higher ability. The introduction of the two student covariates reduces the residual variance of the ability by 33%.

The last model is a MIMIC with DIF, showing that school type has also a direct effect on the probability of success for some of the courses: the better performance of students coming from Lyceum is attenuated for *Accounting*, and magnified for quantitative courses, i.e. *Economics EC*, *Math EC*, *Math MG* and *Statistics*.

For example, let us consider a course with average student rating on teacher's clarity equal to 7, and a student with all covariates equal to zero (HS grade=80, HS type not Lyceum) and having an average residual ability ($\varepsilon_n = 0$): for this student, the estimated probability of passing the exam of *Statistics* equals 9%; if this student has a 10-point increase in the HS grade (90 out of 100), the probability raises to 29%, whereas if the student comes from a Lyceum (keeping HS grade at 80), the probability raises to 51%, which is obtained as:

$$\frac{1}{1+e^{-(\widehat{\alpha}_9+\widehat{\beta}_6+\widehat{\lambda}_9\widehat{\delta}_2)}} = \frac{1}{1+e^{-(-2.35+0.75+1.44\times 1.13)}} = 0.51 \qquad (5)$$

The total effect of Lyceum on the probability of passing *Statistics*, under the given conditions, is $51\% - 9\% = 42\%$. This total effect is the sum of indirect effect and direct effect (DIF). The indirect effect is obtained as $33\% - 9\% = 24\%$, where 33% is derived from equation (5) without the interaction term ($\widehat{\beta_6} = 0.75$). The direct effect is $42\% - 24\% = 18\%$.

The direct effect (DIF) of Lyceum is significant for 5 exams, and it is positive for all these exams, but *Accounting*. Indeed, with a similar computation as above, the total effect of Lyceum on the probability of passing *Accounting* (12%) decomposes into a positive indirect effect (32%) and a negative direct effect (-20%). Therefore, the better performance of students from a Lyceum is attenuated for *Accounting*, since their advantage due to a higher overall ability is partially counterbalanced by a lack of subject-specific background.

4.2. Results: model for exam grades

Table 4 reports the estimates of four specifications of the IRT-MIMIC model defined by equations (3) and (4) for the grades of passed compulsory exams. The seven-point grading scale is defined in Table 2. The item-specific intercepts have been omitted so that all the item-specific thresholds can be estimated. Table 5 reports the estimated thresholds for the first model (standard GRM); the estimated thresholds for the other models are similar.

In order to facilitate comparisons, the parameters of the model for the grades (Table 4) are the same (apart from the intercepts) as those of the model for passing the exams (Table 3), even if some effects are not significant. Note that the model for the grades is fitted on a smaller set of students (615 instead of 808), since students who did not pass any exam are automatically excluded.

Table 4. Estimates of models for grades on first year compulsory exams (2189 exams, 615 students).

	GRM				MIMIC			
Parameter	standard		covariates		without DIF		with DIF	
Thresholds (6×9)	(see Tab.5)		(not reported)		(not reported)		(not reported)	
Fixed effects								
β_1 Teacher's clarity			-0.05		-0.03		-0.02	
β_2 Accounting×lyc							-0.64	*
β_3 Economics EC×lyc							-0.45	
β_4 Math MG×lyc							0.76	
β_5 Math EC×lyc							0.20	
β_6 Stat×lyc							0.07	
Discrimination								
λ_1 Management	1.00	fixed	1.00	fixed	1.00	fixed	1.00	fixed
λ_2 Accounting	1.74	***	1.72	***	1.68	***	1.88	***
λ_4 Economics MG	1.52	***	1.51	***	1.67	***	1.69	***
λ_3 Economics EC	1.46	**	1.45	**	1.28	**	1.38	**
λ_5 History	0.97	***	0.97	***	1.19	***	1.20	***
λ_6 Private law	0.78	**	0.77	**	0.97	***	0.95	***
λ_8 Math MG	0.68	*	0.68	*	0.56	*	0.62	*
λ_7 Math EC	1.38	***	1.35	***	1.25	***	1.31	***
λ_9 Statistics	1.68	***	1.66	***	1.27	***	1.32	***
Structural model								
δ_1 HS grade					0.06	***	0.06	***
δ_2 HS lyceum					0.57	***	0.66	***
τ Ability SD	0.93	***	0.93	***	0.74	***	0.72	***
n. of parameters	63		64		66		71	
$logL$	-3760.52		-3760.35		-3675.83		-3670.07	

legend: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

The estimated discrimination parameters of the two models are in good agreement (e.g. the correlation is 0.75 for the standard versions), which indicates that the implicit grade corresponding to a failed examination is assigned in a way that is somewhat consistent with the use of the grading scale for a successful examination.

The estimated standard deviation of the ability in the model for the grades (0.93) is considerably lower than in the model for passing the exams (2.33). However, as noted before, in the model for passing the exams the standard deviation of the ability reduces to 1.05 if we omit students who did not pass any exam.

Interestingly, the course quality assessed through the average student ratings on teacher's clarity has a significant effect on the probability of passing the exams, but not on the grades.

In the MIMIC part, the effects of Lyceum and HS grade and their contribution to the reduction of the residual variance of the ability (-37%) are similar to those of the model for passing the exams.

Only the interaction term (DIF) between Lyceum and *Accounting* is significant in the model for the grades, with the same sign as in the model for passing the exams. The other interaction terms are not significant, anyway note that in the model for the grades the sample sizes are substantially reduced (see Table 1).

Table 5. Estimated thresholds for standard GRM, grades on first year compulsory exams.

Item	γ_{1i}	γ_{2i}	γ_{3i}	γ_{4i}	γ_{5i}	γ_{6i}
Management	-4.03	-2.20	-1.37	-0.49	0.99	2.19
Accounting	-4.64	-0.99	0.39	1.99	3.52	4.44
Economics MG	-3.63	-2.31	-0.15	1.04	2.67	4.50
Economics EC	-2.55	0.07	0.56	1.43	2.86	4.21
History	-1.61	-0.59	0.26	1.38	2.64	3.49
Private law	-2.42	-0.87	-0.02	1.11	2.37	3.53
Math MG	-0.72	0.60	1.22	1.73	2.87	3.50
Math EC	-4.48	-0.32	0.47	1.51	2.48	3.52
Statistics	-2.30	-0.43	0.42	1.45	2.38	3.27

The estimated thresholds in Table 5 convey information about how the grading scale is used by teachers of compulsory exams, with reference to a student with average ability ($\theta_n = 0$). In general, the grade tends to be lower as the thresholds become larger. It is worth to note that the estimated thresholds of the exams are not related by a simple shift, thus there is evidence of a different use of the scale.

To convert the thresholds into probabilities, we exploit the following formula, derived from model (3):

$$\begin{aligned} P(Y_{in} = c \mid \theta_n) &= P(Y_{in} \le c \mid \theta_n) - P(Y_{in} \le c-1 \mid \theta_n) \\ &= \frac{1}{1+e^{-(\gamma_{ic} - \lambda_i \theta_n)}} - \frac{1}{1+e^{-(\gamma_{i,c-1} - \lambda_i \theta_n)}} \end{aligned}$$

for $c = 1, \dots, 7$, posing $P(Y_{in} \le c-1 \mid \theta_n) = 0$ if $c = 1$. For a student with average ability ($\theta_n = 0$) these probabilities depend only on the thresholds. Table 6 reports the predicted probabilities $P(Y_{in} = c \mid \theta_n = 0)$ using the estimated thresholds of Table 5.

Table 6. Estimated probabilities for standard GRM (student with average ability), grades on first year compulsory exams.

	$P(Y_{in} = c \mid \theta_n = 0)$						
Item	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$	$c = 7$
Management	0.02	0.08	0.10	0.18	0.35	0.17	0.10
Accounting	0.01	0.26	0.33	0.28	0.09	0.02	0.01
Economics MG	0.03	0.06	0.37	0.28	0.20	0.05	0.01
Economics EC	0.07	0.45	0.12	0.17	0.14	0.04	0.01
History	0.17	0.19	0.21	0.23	0.13	0.04	0.03
Private law	0.08	0.21	0.20	0.26	0.16	0.06	0.03
Math MG	0.33	0.32	0.13	0.08	0.10	0.02	0.03
Math EC	0.01	0.41	0.19	0.20	0.10	0.05	0.03
Statistics	0.09	0.30	0.21	0.21	0.11	0.05	0.04

The wide variability of the first threshold across exams in Table 5 reflects the heterogeneity in the use of the minimum grade, which is very common for *Math MG* (33%), but rare for *Accounting* (1%), as shown by Table 6. However, we cannot claim that the grades at *Accounting* are generally higher than the grades at *Math MG* since grade 7 is less likely for *Accounting*. In other words, the teachers of *Accounting* tend to rule out extreme grades.

In order to compare grades across exams for a student with ability different from the average, we can rely on item characteristic curves. Figure 4 reports, as a function of the standardized ability θ_n/τ, the item characteristic curves for the probabilities to get the following grades: the minimum grade $(Y = 1 \leftrightarrow score = 18)$, a low grade $(Y \leq 2 \leftrightarrow score \leq 21)$, a high grade $(Y \geq 6 \leftrightarrow score \geq 28)$, the maximum grade $(Y = 7 \leftrightarrow score = 30$ or '30 with honors'$)$. Due to exam-specific discriminations and thresholds, the ranking of the exams in terms of the probability to get a given grade varies across the range of student ability. For example, the probability of get the minimum grade at *Math MG* is quite large across the whole range of ability, whereas at *Statistics* the corresponding probability is very large for low ability and negligible for high ability.

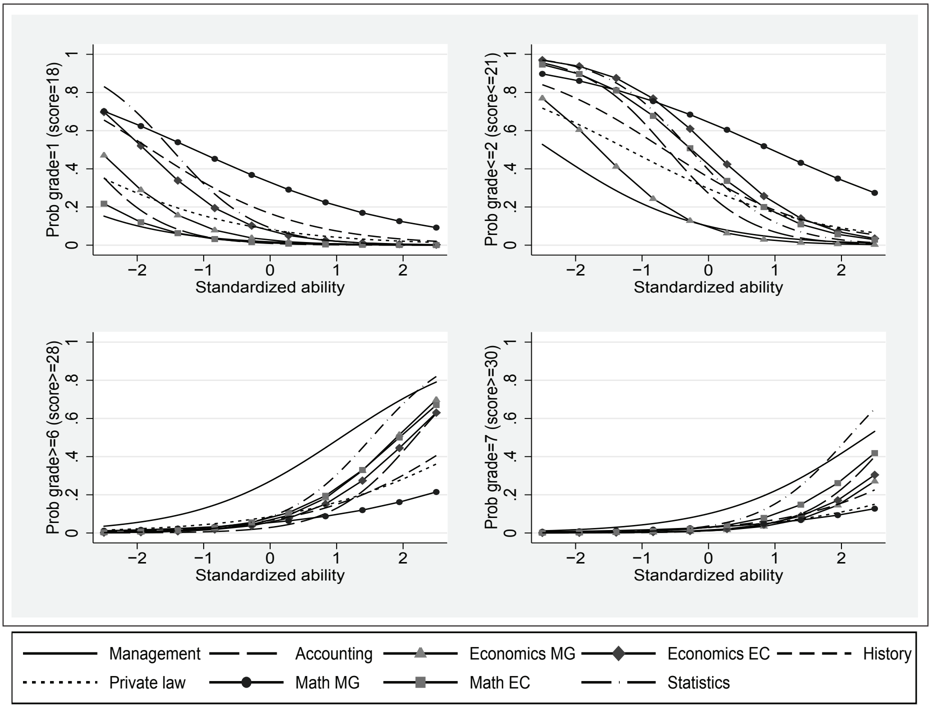


Figure 4. Standard GRM item characteristic curves for some probabilities

The MIMIC models reported in Table 4 allow to take into account the effect of student characteristics on the ability. As for the model for passing the exams, high school grade (centered on the mid-point 80) and high school type have significant effects: students with a better grade and coming from a Lyceum have higher ability. The last model is a MIMIC with DIF, which is significant only for *Accounting*.

In order to illustrate how to interpret the MIMIC parameters, let us consider the probability of getting a grade higher than 3 (i.e. a score ≥ 24, which is the mid-point of the scale) for a course with average student rating on teacher's clarity equal to 7, and a student with all covariates equal to zero (HS grade= 80, HS type= not Lyceum). Considering a student with average residual ability ($\varepsilon_n = 0$), the predicted probability of getting a grade > 3 in *Statistics* is 32%; if this student has a 10-point increase in the HS grade (90 out of 100), the probability raises to 49%, whereas if the student comes from a Lyceum (keeping HS grade= 80), the probability raises to 55%. This probability is obtained as one minus the cumulative probability for $c = 3$ given by inverting equation (3):

$$1 - \frac{1}{1 + e^{-[\widehat{\gamma}_{3,9} - (\widehat{\beta}_6 + \widehat{\lambda}_9 \widehat{\delta}_2)]}} = 1 - \frac{1}{1 + e^{-[0.42 - (0.07 + 1.32 \times 0.66)]}} = 0.55 \tag{6}$$

Thus, for *Statistics* the total effect of Lyceum is $55\% - 32\% = 23\%$ (considering a student with HS grade=80). The interaction term β_6 is not significant for *Statistics*; anyway, as in the model for passing the exams, the total effect of Lyceum can be decomposed into an indirect effect (21%) and a direct effect (2%).

5. *Final Remarks*

The paper considered the performance of students on compulsory first year exams at the School of Economics and Management of the University of Florence. We first fitted models for the probability of passing the exams and then models for the grades obtained on passed exams.

All the models have an IRT structure in order to measure student ability alongside with exam-specific difficulty and discrimination power. Moreover, we extended the models with a structural part (MIMIC) to estimate the effects of student's characteristics (Lyceum and HS grade) on the abilities of passing the exams and obtaining a high grade. The contribution of such characteristics is substantial, since the residual variance of the ability reduces by about one third in both models. We also looked for direct effects of these covariates by adding Differential Item Functioning (DIF): the better performance of students from a Lyceum is attenuated for *Accounting* and magnified for quantitative exams.

As regard course quality, assessed through the average student ratings on teacher's clarity, our analysis has found a significant effect on the probability of passing the exams, but not on the grades. In other words, a better teacher is not associated with a better performance of successful students, but with a higher success rate and this could help to reduce the drop-out rate. The estimation of the effect of the average student ratings is free from selection bias since students are assigned to classes depending on the first letter of the surname, even if there could be some bias due to non-response. However, it is hard to reliably assess the impact of a specific intervention in the course organization since the student ratings are a summary of a complex entity, thus the impact should be evaluated through an experiment.

The grades are modelled through a Graded Response Model, which allows to understand how the grading scale is used by the teachers and to carry out fair comparisons among the exams, thus locating exams with an anomalous pattern (e.g. *Math MG*).

Our approach based on IRT-MIMIC models represents a noteworthy advance with respect to traditional approaches relying on a summary measure of student performance, such as the the number of passed exams, the number of gained credits or a proficiency indicator combining credits and grades (for a short review see Grilli et al., 2013). Indeed, IRT modelling gives insights into the peculiarities of the exams and the role of course and student characteristics. This information is valuable in order to design and implement policies to improve the degree programme organization and to tailor the student tutoring service.

Acknowledgements: This research was partially supported by the FIRB 2012 project "Mixture and latent variable models for causal inference and analysis of socio-economic data", grant n. RBFR12SHVV_003.

References

Baker, F.B., Kim, S. (2004). Item response theory: parameter estimation techniques. New York: Dekker.

Grilli L., Rampichini C., Varriale R. (2013). Predicting students' academic performance: a challenging issue in statistical modelling. Cladag 2013 - 9th Meeting of the Classification and Data Analysis Group. Modena, September 18 - 20, 2013. Book of Abstracts, Eds. Minerva T., Morlini I., Palumbo F..

Jöreskog, K.G. and Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, **70**, 631–639.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004a). Generalized multilevel structural equation modeling. *Psychometrika*, **69**, 167–190,

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004b). *GLLAMM Manual*. Free content U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160.

Rijmen, F., Tuerlinckx, F., De Boeck, P., and Kuppens P. (2003). A Nonlinear Mixed Model Framework for Item Response *Theory. Ann. Psychological Methods*, **8**, 185–205.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. Psychometrika Monographs, 34(Suppl. 4).

QdS – Vol. 15, 2013

Nonlinear CUB models: some stylized facts

Marica Manisera
University of Brescia
E-mail: manisera@eco.unibs.it

Paola Zuccolotto
University of Brescia
E-mail: zuk@eco.unibs.it

Summary: The Nonlinear CUB models have been recently introduced with the aim of generalizing the standard CUB in the context of rating data modelling. In this paper the stylized facts concerning the main features of the Nonlinear CUB models are established by means of an extended systematic analysis of a great number of different models. Results provide interesting insights on this new class of models and suggestions about the future theoretical developments.

Keywords: CUB models; Nonlinear CUB models; Rating data; Likert-type scales; Latent variables; Transition probability; Transition plot.

1. Introduction

Statistical analyses in several fields often deal with rating data, used to investigate the individuals' perceptions, attitudes, behaviours, cognitions. Rating data are usually collected by means of a questionnaire involving categorical ordinal items, i.e. questions whose possible responses are measured on an ordinal scale. In the literature, several methods and techniques have been proposed to model rating data, taking into account their categorical ordinal nature (see Agresti, 2010; Tutz, 2012). Among them, a different paradigm is given by the CUB models (D'Elia and Piccolo, 2005; Piccolo, 2006; Piccolo and D'Elia, 2008; Iannario and Piccolo, 2012), introduced in 2003 with the name MUB (Piccolo, 2003). Since then, the CUB models have been developed in several directions and many papers concerning inferential issues, identifiability problems, fitting measures, computational strategies and software routines have been published (Iannario, 2009, 2010, Iannario and Piccolo, 2010, 2012, 2014). In addition, the CUB models have been extended in several directions, for example to consider subjects' and objects' covariates (Iannario 2007, 2008; Piccolo, 2013), the so-called shelter effect, resulting in a very high frequency on a given response category (Iannario, 2012a), the possible presence of a hierarchical structure in the data (Iannario, 2012b), multimodal response distributions deriving from a latent class structure (Grilli, Iannario, Piccolo and Rampichini, 2013). The interest towards the CUB models has increased also from the point of view of applications, because they can be used effectively in different contexts, for example linguistics (Balirano and Corduas, 2008), risk analysis (Cerchiello, Iannario and Piccolo, 2010), marketing (Iannario, Manisera, Piccolo e Zuccolotto, 2012), medicine (D'Elia, 2008), sensometrics (Piccolo and D'Elia, 2008).

A possible generalization of the CUB models is the so-called Nonlinear CUB (NLCUB), a new class of models recently proposed in order to deal with the unequal spacing of the response categories in the respondent's mind (Manisera and Zuccolotto, 2014). The unequal spacing of categories has been translated into the concept of nonlinearity, defined as the presence of non-constant transition probabilities, i.e. the probabilities of moving from one rating to the next one during a decision process where the expressed rating derives from a step-by-step mechanism. NLCUB, differently from CUB, can be used to model rating data with non-constant transition probabilities. Simulation studies and real data analyses (Manisera and Zuccolotto, 2013, 2014) show promising results that encourage further research.

The aim of this paper is to present some stylized facts concerning the NLCUB models, deriving from an extended systematic study performed in order to investigate their behaviour. Based on the findings of this study, we draw some interesting conclusions on the nonlinearity patterns expressed by different NLCUB models and useful suggestions concerning their estimation procedure.

The paper is organized as follows: in Section 2 we describe the basic features of the CUB and NLCUB models. In particular, the concept of transition probability is defined in Subsection 2.2 and its formulation is derived for both CUB and NLCUB models, while the parameter estimation procedure of NLCUB models is briefly recalled in Subsection 2.3. In Section 3 we introduce the concepts of linearity and nonlinearity of the decision process underlying the responses on a rating scale. Also, we propose a nonlinearity index able to measure the degree of nonlinearity in the decision process. In Section 4 the results of a wide systematic study are presented and summarized by some stylized facts. Section 5 concludes the paper.

2. CUB and Nonlinear CUB models

The CUB models have been introduced in the literature to analyse ordinal data and fit in the latent variable framework. With the CUB models, rating or ranking data are modelled by a mixture of a Uniform and a Shifted Binomial random variables: the observed rating r ($r = 1, \dots, m$) is a realization of the discrete random variable R

whose probability distribution is given by

$$Pr\{R = r|\theta\} = \pi Pr\{V(m,\xi) = r\} + (1-\pi)P\{U(m) = r\} \tag{1}$$

with $r = 1,\ldots,m$, $\theta = (\pi,\xi)'$, $\pi \in (0,1]$, $\xi \in [0,1]$. The number of possible response categories m is a given and known integer. For a given m, $V(m,\xi)$ is a Shifted Binomial random variable, with trial parameter m and success probability $1-\xi$, modelling the *feeling* component, and $U(m)$ is a discrete Uniform random variable defined over the support $\{1,\ldots,m\}$, aimed to model the *uncertainty* component. The CUB models are identifiable for $m > 3$ (Iannario, 2010).

The Nonlinear CUB models (NLCUB), introduced by Manisera and Zuccolotto (2014), are a generalization of the CUB models. In detail, with NLCUB the discrete random variable *R* generating the observed rating r has a probability distribution depending on a new parameter T, $T \geq m-1$ and given by

$$Pr\{R = r|\theta\} = \pi \sum_{y\in l^{-1}(r)} Pr\{V(T+1,\xi) = y\} + (1-\pi)P\{U(m) = r\} \tag{2}$$

where l is a function mapping from $(1,\ldots,T+1)$ into $(1,\ldots,m)$. In detail, l is defined as

$$l(y) = \begin{cases} 1 & \text{if} \quad y \in [y_{11},\ldots,y_{g_1 1}] \\ 2 & \text{if} \quad y \in [y_{12},\ldots,y_{g_2 2}] \\ \vdots & \vdots \quad \vdots \\ m & \text{if} \quad y \in [y_{1m},\ldots,y_{g_m m}] \end{cases} \tag{3}$$

where y_{hs} is the h-th element of $l^{-1}(s)$, and

$$(y_{11},\ldots,y_{g_1 1},y_{12},\ldots,y_{g_2 2},\ldots,y_{1m},\ldots,y_{g_m m}) = (1,\ldots,T+1).$$

We denote with $g_s = |l^{-1}(s)|$, where $|\cdot|$ denotes the cardinality of a set, the number of "latent" values to which rating s corresponds based on l. The values $g_1,\ldots,g_m$ univocally determine the function l and can be considered as parameters of the model. We have $T = g_1 + \ldots + g_m - 1$.

When $T = m-1$ and $g_s = 1$ for all $s = 1,\ldots,m$, then the proposed model collapses into the standard CUB model.

2.1. The general framework for the decision process

In Manisera and Zuccolotto (2014) the NLCUB formulation is derived as a special case of a more general framework, proposed to describe the decision process (DP) driving individuals' responses to survey questions with ordered response levels. This general model assumes the presence of two different approaches, which compose the DP and, borrowing the CUB terminology, are called feeling and uncertainty approach, respectively. The feeling approach proceeds through T consecutive steps, called feeling path. At each step, an elementary judgment is given. The rating of the feeling path results from these elementary judgments that are, firstly, summarized and, secondly, transformed into a Likert-scaled rating. The uncertainty approach consists of a random judgment that can be given by the respondent due to the indecision in choosing the ordinal response, depending on a great variety of possible reasons, e.g. unconscious willingness to delight the interviewer, difficulty in evaluating some specific objects using limited information, partial understanding, lack of self-confidence, laziness, boredom, etc. In the end, the expressed rating can derive from the feeling or the uncertainty approach with given probabilities. Some existing statistical models can be viewed as special cases of this general framework.

The most interesting feature of this DP is the mechanism that, along the feeling path, generates the rating according to the feeling approach. We address the reader to the seminal paper on NLCUB models for a formal statistical description and two illustrative examples that highlight the difference between the CUB and NLCUB models. Here we limit ourselves to provide an intuitive explanation. First of all, the difference between the DPs of NLCUB and CUB models only pertain the feeling approach. In both models, the idea is that the elementary judgement given at each step of the feeling path can be viewed as a quick and instinctive "Yes/no" response to a very simple question. For example, when a respondent is asked to express his/her agreement with a certain statement by using a Likert scale from 1 to $m = 5$, the simple question can be "Do I have a positive sensation about this statement? Yes or no?". The sequence of elementary judgements obtained in the feeling path is a sequence of "Yes" and "No" responses that reflect the set of positive and negative sensations that disorderly come to mind during the reasoning, according to the individual's experience about the latent trait being evaluated. The main difference between CUB and NLCUB is that:

1. in the DP of CUB models, the number of steps in the feeling path (that is the number of simple questions) is $T = m - 1 = 4$ and the last rating of the feeling path is given by 1 plus the total number of "Yes" responses. This last rating follows a Shifted Binomial distribution, since the basic judgments are realizations of *iid* Bernoulli random variables;

2. in the DP of NLCUB models, the number of steps in the feeling path is $T > m-1$ and the last rating of the feeling path is still based on the total number of "Yes" responses, but in an unbalanced way. As an example, we can have $T = 9$ and the total number of "Yes" responses can be transformed into the last rating of the feeling path by the rule represented in Table 1, which shows that, for example, rating 2 is reached with one, two, three or four "Yes" responses and moving from rating 1 to rating 2 is much more easier than moving from rating 2 to rating 3.

Then, in both CUB and NLCUB the final response can be either the rating deriving from the feeling approach or a random rating resulting from the uncertainty approach,

Table 1. DP of NLCUB models - Feeling approach (example with $m = 5$ and $T = 9$)

$T = 9$ $(> m - 1)$ elementary judgments: “Positive sensation? Yes or no?”

Number of “Yes” responses	0	1	2	3	4	5	6	7	8	9
Corresponding rating	1		2				3		4	5

with probabilities π and $1 - \pi$, respectively. It is easy to see that the expressed ratings derived from the mechanism in 1. and in 2. follow distribution (1) and (2), respectively. In particular, in (2) the asymmetric correspondence between the total number of “Yes” responses and the rating of the feeling approach is accounted for by the function l and the values g_s, which denote the number of positive sensations needed to move to the next rating (in the above example, $g_1 = 1$, $g_2 = 4$, $g_3 = 3$, $g_4 = 1$ and $g_5 = 1$).

2.2. Transition probabilities for CUB and Nonlinear CUB

The way respondents achieve, moving through T steps, the formulation of a rating in the feeling approach is called feeling path. The examples of DP considered in the previous section show that, in the end of the feeling path, the respondent (unconsciously) considers the total number of “Yes” responses (i.e. the total number of positive sensations that came into his/her mind) and decides which rating should be assigned, according to some rule. As a matter of fact, we can imagine that the same reasoning is made at each step of the feeling path. In other words, at each step (i.e. for each new basic judgment he/she expresses), the respondent considers the number of “Yes” responses collected up to that moment and formulates a provisional rating, which will be updated at the next step, until the T-th step has been reached.

Within this framework, we can express the so-called transition probabilities $\phi_t(s)$, i.e. the probability of moving to provisional rating $s + 1$ at step $t + 1$ of the feeling path, given that the provisional rating at step t is s, $s = 1, \ldots, m - 1$. Transition probabilities depend on the function l, i.e. on the rule according to which the respondents transform the number of “Yes” responses into the rating during the feeling path. Therefore, transition probabilities describe the respondents’ state of mind about the response scale used to express judgments in the feeling approach.

For ease of interpretation, the average transition probability $\phi(s)$, obtained averaging $\phi_t(s)$ over t, is generally used. It indicates the “perceived closeness” between ratings s and $s+1$ and can be transformed into a “perceived distance” $\delta_s = h(\phi(s))$ by means of a proper function (usually $\delta_s = -\log(\phi(s))$). These quantities are the basis for constructing the so-called transition plot, useful to detect whether the ratings are perceived by respondents as equally spaced or not. In the transition plot, a broken line joins the points $(s, \tilde{\phi}(s-1))$, $s = 1, \ldots, m$, $\tilde{\phi}(0) = 0$, and $\tilde{\phi}(s-1) = (\delta_1 + \cdots + \delta_{s-1})/(\delta_1 + \cdots + \delta_{m-1})$ for $s = 2, \ldots, m$. Figure 1 represents two examples of linear (left) and nonlinear (right) transition plot.

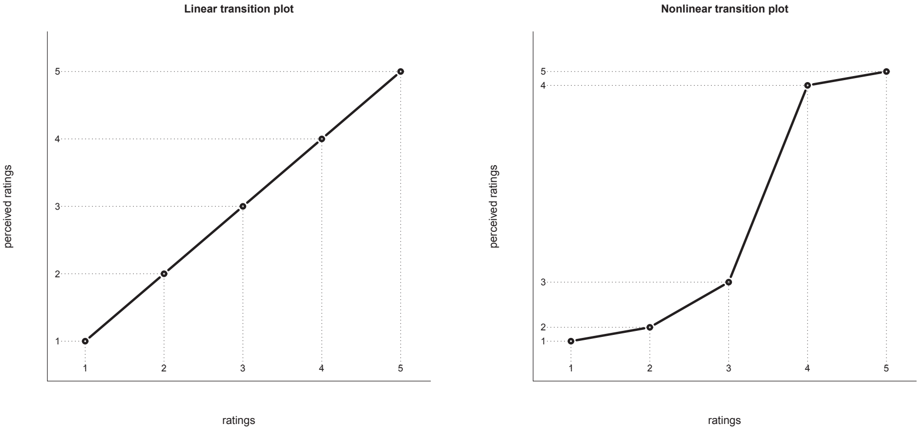


Figure 1. Examples of linear (left) and nonlinear (right) transition plot ($m = 5$)

By construction, the y-axis in the transition plot ranges in $[0, 1]$. A linear transition plot suggests that the ratings are perceived as equally-spaced in the respondents' mind (Figure 1, left) while a nonlinear transition plot accounts for unequally-spaced perceived ratings (Figure 1, right).

Starting from the transition probabilities, we can also define the expected number μ of one-rating-point increments during the feeling path and the unconditional probability of increasing one rating point in one step of the feeling path $\phi = \mu/T$.

Manisera and Zuccolotto (2014) derive $\phi_t(s)$, $\phi(s)$, μ and ϕ for CUB and NLCUB models. In the CUB models, the transition probabilities are constant over t, s and given by

$$\phi_t(s) = 1 - \xi \quad \forall t, s. \tag{4}$$

with $s = 1, \ldots, m-1$, $t = 1, \ldots, m-1$ and $\phi_0 = \phi_0(1) := 1 - \xi$. In addition, we also have $\phi = \phi_t(s) = 1 - \xi$ and $\mu = (m-1)(1-\xi)$. In other words, in the CUB models $1 - \xi$, that is the *feeling* parameter, indicates the probability of increasing one rating point in one step of the feeling path.

In the NLCUB models, the transition probabilities result

$$\phi_t(s) = (1-\xi) \frac{\binom{t}{w_{g_s s}} (1-\xi)^{w_{g_s s}} \xi^{t - w_{g_s s}}}{\sum_{h=1}^{g_s} \binom{t}{w_{hs}} (1-\xi)^{w_{hs}} \xi^{t - w_{hs}}} \tag{5}$$

where $w_{hs} = y_{hs} - 1$ and with $s = 1, \dots, m-1$, $w_{1s} \leq t < T$, $\phi_0 = \phi_0(1) := 0$ if $g_1 > 1$ and $\phi_0 = \phi_0(1) := 1 - \xi$ if $g_1 = 1$. When $T = m - 1$ and $g_s = 1$ for all s, formulas (4) and (5) coincide. In NLCUB, the expected number of one-rating-point increments during the feeling path is given by

$$\mu = \phi_0 + (1-\xi) \sum_{t=1}^{T-1} \sum_{s=1}^{m-1} \binom{t}{w_{g_s s}} (1-\xi)^{w_{g_s s}} \xi^{t - w_{g_s s}} \tag{6}$$

and $1 + \mu$ is the expected rating of the feeling approach, without the effect of the uncertainty approach.

2.3. Parameter estimation of the Nonlinear CUB models

In this paragraph we briefly recall the procedure proposed to estimate the parameters of a NLCUB model. It's worth pointing out that estimating a NLCUB model implies estimating both the parameters π, ξ and the parameters $g_1, \dots, g_m$ describing the function l. Therefore, the transition probabilities and the shape of the transition plot are estimated from the data.

Given a random sample of n expressed ratings $\mathbf{s} = (s_1, \dots, s_n)$, the loglikelihood function L of a NLCUB model for fixed $\mathbf{g} = (g_1, \dots, g_m)$ can be written as

$$L(\xi, \pi | \mathbf{g}; \mathbf{s}) = \sum_{i=1}^{n} \log \left\{ \pi \left[\sum_{h=1}^{g_{s_i}} \binom{T}{w_{hs_i}} (1-\xi)^{w_{hs_i}} \xi^{T - w_{hs_i}} \right] + (1-\pi)\frac{1}{m} \right\} \tag{7}$$

with $T = g_1 + \cdots + g_m - 1$. We obtain the estimates $\hat{\boldsymbol{\theta}} = (\hat{\xi}, \hat{\pi}, \hat{\mathbf{g}})$ by the following procedure:

- fix a maximum value T_{max} for T;
- considering all the possible configurations of $g_1, \dots, g_m$ such that $g_1 + \cdots + g_m \leq T_{max} + 1$, compute

$$\hat{\mathbf{g}} = (\hat{g}_1, \dots, \hat{g}_m) = \arg\max_{\mathbf{g}} \left\{ \max_{\xi, \pi} L(\xi, \pi | \mathbf{g}; \mathbf{s}) \right\};$$

- maximize (7) with respect to ξ and π to get

$$\hat{\xi}, \hat{\pi} = \arg\max_{\xi, \pi} L(\xi, \pi | \hat{\mathbf{g}}; \mathbf{s}).$$

The number of possible configurations of $g_1, \dots, g_m$ to be considered in the estimation procedure clearly depends on the values of m and T_{max}. For example, for $m = 5$ and $T_{max} = 8, 9, 10, 11$ we have $126, 252, 462, 687$ possible configurations of

$g_1, \ldots, g_m$, respectively. Estimation, along with other inferential issues, are the main challenges of the NLCUB models and further research is being devoted to refine some points, as discussed in Manisera and Zuccolotto (2014). With reference to the choice of T_{max}, which could appear discretionary, some considerations are discussed in Subsection 4.3 of this paper.

3. Linear and nonlinear decision processes

The decision process underlying the individuals' responses on a rating scale has been defined to be linear or nonlinear according to whether the transition probabilities $\phi_t(s)$ are constant on non-constant for different t and s. This implies that, for linear processes, the transition plot shows a straight line, since the probability of increasing one rating point in the next step of the feeling path is constant for every rating in every step (as in the example of Figure 1, left).

Manisera and Zuccolotto (2014) derive, under some general assumptions, a sufficient condition for linearity and show that CUB (*i*) is a particular case of the general framework and (*ii*) meets the sufficient condition for linearity. The NLCUB models, instead, are a nonlinear variant of the general model and this is a reason for their name. A graphical explanation is also possible, since the transition plot of the NLCUB models generally shows a nonlinear broken line, giving interesting insights on the way the respondents perceive the response scale and, in particular, the distance among the response categories (as in the example of Figure 1, right).

Starting from the above definition of linearity, in this paper we propose to measure the degree of nonlinearity expressed by a NLCUB model as the standard deviation of the transition probabilities. Formally, we define the following nonlinearity index:

$$\lambda(\xi, \mathbf{g}) = \sigma(\phi_t(s))/\max(\sigma) \tag{8}$$

where $\sigma(\phi_t(s))$ is the standard deviation of $\phi_t(s)$, $\forall t, s \in \Phi$ with Φ denoting the set containing all the pairs $(t, s) : \exists\ \phi_t(s)$. The value $\max(\sigma)$ can be obtained as follows. Let $|\Phi| = k$, where $|\cdot|$ denotes the cardinality of a set. For odd k, $\max(\sigma)$ is the value of $\sigma(\phi_t(s))$ in the extreme situation where $k/2$ probabilities $\phi_t(s)$ equal 0 and the remaining $k/2$ probabilities equal 1; in this case with simple algebra we obtain $\sigma(\phi_t(s)) = \sqrt{1/2 - 1/4} = \sqrt{1/4}$. For even k, $\max(\sigma)$ is reached when either $(k-1)/2$ probabilities $\phi_t(s)$ equal 0 and the remaining $(k+1)/2$ probabilities equal 1 or $(k+1)/2$ probabilities $\phi_t(s)$ equal 0 and the remaining $(k-1)/2$ probabilities equal 1. In these two cases we have $\sigma(\phi_t(s)) = \sqrt{(k+1)/2k - (k+1)^2/4k^2}$ and $\sigma(\phi_t(s)) = \sqrt{(k-1)/2k - (k-1)^2/4k^2}$, respectively. In both cases we finally obtain $\sigma(\phi_t(s)) = \sqrt{1/4 - 1/4k^2)}$. Therefore, we have

$$\max(\sigma) = \begin{cases} \sqrt{1/4} & \text{if} \quad k \quad \text{is} \quad \text{odd} \\ \sqrt{1/4 - 1/4k^2} & \text{if} \quad k \quad \text{is} \quad \text{even} \end{cases} .$$

The index $\lambda(\xi, \mathbf{g})$ is normalized in [0,1] (or [0,100] if expressed in percentage) and can be interpreted as the proportion of nonlinearity in the NLCUB model respect to its maximum. The nonlinearity index λ is expressed as a function of $(\xi, \mathbf{g})$ and does not depend on π, because the transition probabilities only pertain the feeling approach.

4. Stylized facts

In this section, we empirically observe the statistical features of several different NLCUB models, obtained by varying the parameters in the parameter space so as to systematically explore a huge number of possible combinations. In the end, we draw some stylized facts considering the following issues:

- the extrapolation of some particular cases concerning the existence of linear DPs within the NLCUB framework;
- the analysis of some evidence about the nonlinearity pattern of different NLCUB models;
- the possibility to draw some remarks about the choice of T_{max} for estimation purposes.

In order to gain insights about these issues, we have computed the values of $\phi_t(s)$, $\phi(s)$ $(t = 1, 2, \ldots, T-1; s = 1, \ldots, m-1)$ and μ for all the NLCUB models obtained crossing the experimental conditions reported in Table 2 (altogether, 4,752 different combinations of m, T, ξ). For each case, we have investigated all the possible configurations of $(g_1, \ldots, g_m)$, so that a total number of 13,695,858 NLCUB models have been considered in this systematic study.

Table 2. Experimental conditions

Parameter	Explored values
m	$4, 5, 6, 7$
T	$m-1, m, m+1, \ldots, 3m-1$
ξ	$0.01, 0.02, \ldots, 0.98, 0.99$

For illustrative purposes, we report the transition plots for 36 selected combinations of m, T, ξ (Figures 2 - 5; for each case, all the combinations of $(g_1, \ldots, g_m)$). At a first sight we notice that (1) each combination seems to contain (at least) one linear transition plot, (2) the shapes of the transition plots tend to become "more nonlinear" with increasing values of T and decreasing values of ξ. These two rough remarks will be more deeply analysed in the next three subsections.

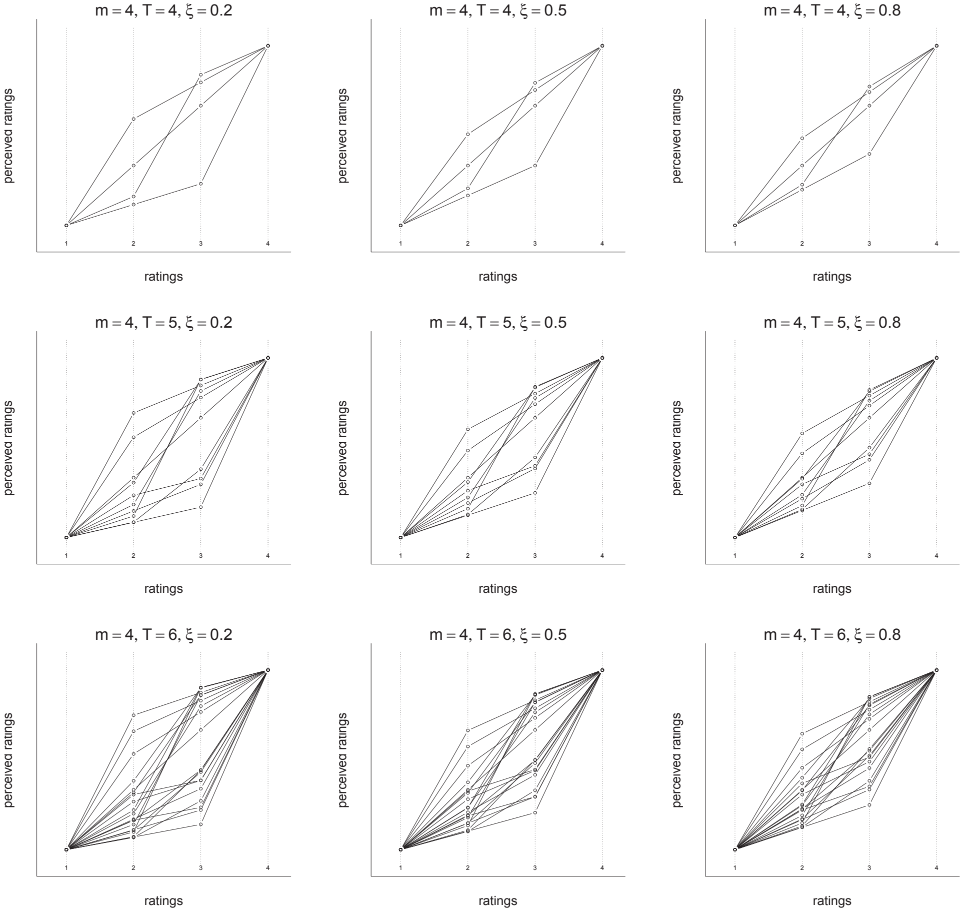


Figure 2. Transition plots for the cases with $m = 4$*,* $T = 4, 5, 6$ *(top, middle, bottom),* $\xi = 0.2, 0.5, 0.8$ *(left, middle, right)*

4.1. Linear DPs within the NLCUB framework

Within the NLCUB framework, the sufficient condition for linearity in Manisera and Zuccolotto (2014) is satisfied only by the configuration $g_1, \ldots, g_m = (1, \ldots, 1)$, that is, when the NLCUB collapses into a classical CUB model. However, we are aware that other linear DPs may exist within the NLCUB framework, as the above mentioned condition is not necessary. Our empirical investigation has found that, for each combination of m and T, a linear DP is generated by the configuration of $g_1, \ldots, g_m$ such that $g_s = 1$ with $s = 1, \ldots, m - 1$ and $g_m = T - m + 2$, whatever the value of ξ. As for future

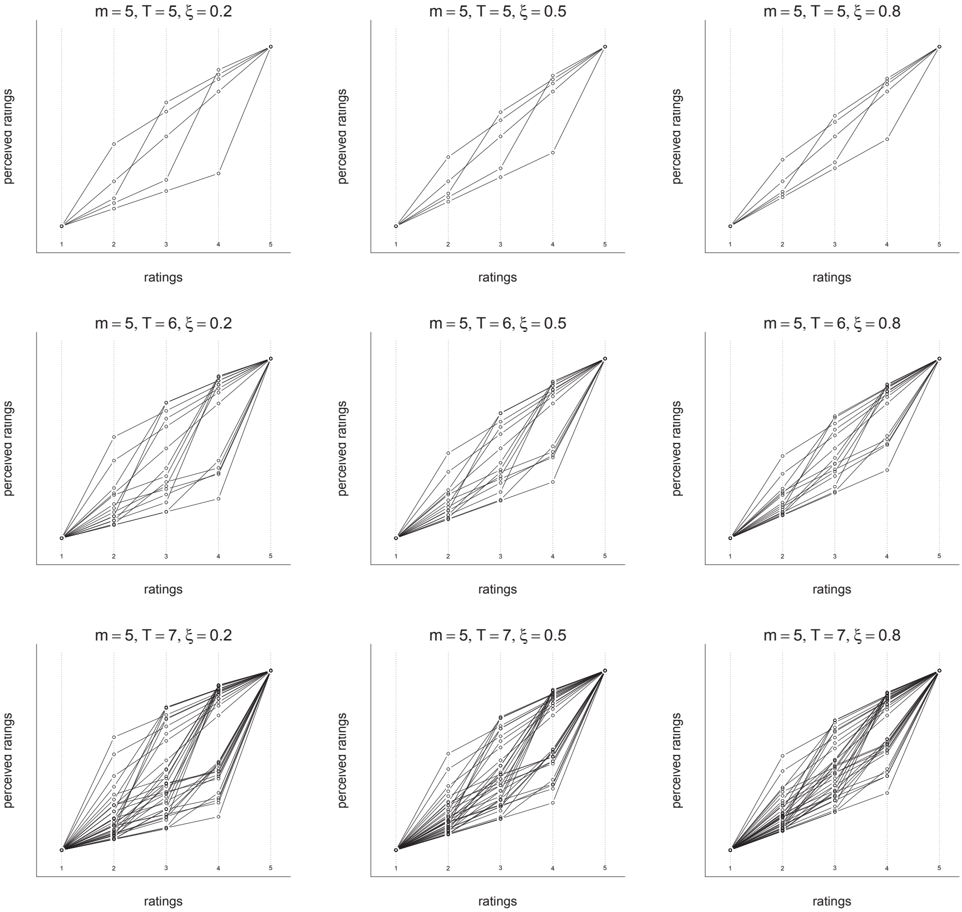


Figure 3. Transition plots for the cases with $m = 5$*,* $T = 5, 6, 7$ *(top, middle, bottom),* $\xi = 0.2, 0.5, 0.8$ *(left, middle, right)*

research, this constitutes a clear suggestion for trying to define a sufficient and necessary condition. Although all these different DPs meet the definition of linearity, their feeling paths work according to different mechanisms, so that the same values of ξ correspond to different values of μ (see Figure 6). In detail, with high values of T, μ tends to remain fixed at its highest value until ξ reaches a given threshold.

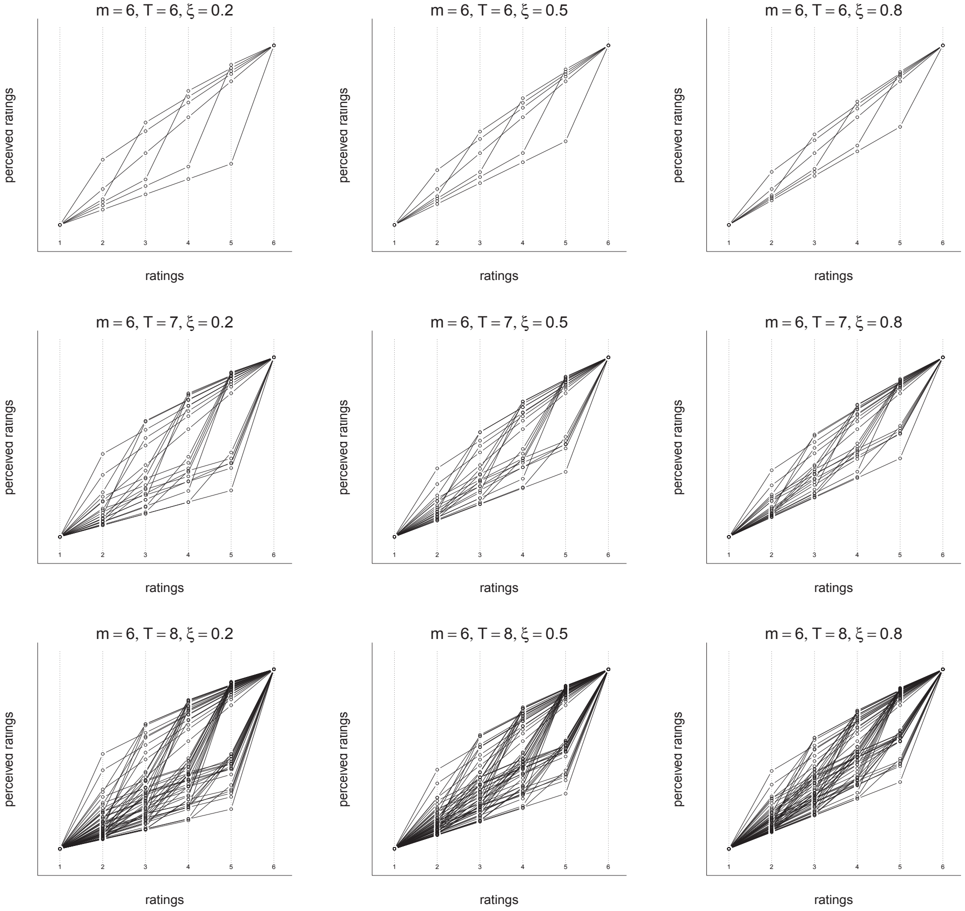


Figure 4. Transition plots for the cases with $m = 6$*,* $T = 6, 7, 8$ *(top, middle, bottom),* $\xi = 0.2, 0.5, 0.8$ *(left, middle, right)*

4.2. Empirical evidence about nonlinearity

In this subsection we explore the relationships between the nonlinearity index $\lambda(\xi, \mathbf{g})$ and the parameters ξ and T. Figure 7 shows the overall plot of $\lambda(\xi, \mathbf{g})$ versus ξ and the corresponding partial plots for some selected values of T, with $m = 4$, the remaining three cases ($m = 5, 6, 7$) being substantially similar.

The graphs clearly show that the highest levels of nonlinearity can be reached with low values of ξ, provided that T is large enough (from $T = 7$ onwards, in this case) and that higher values of T allow $\lambda(\xi, \mathbf{g})$ to cover a wider range of values. The linear

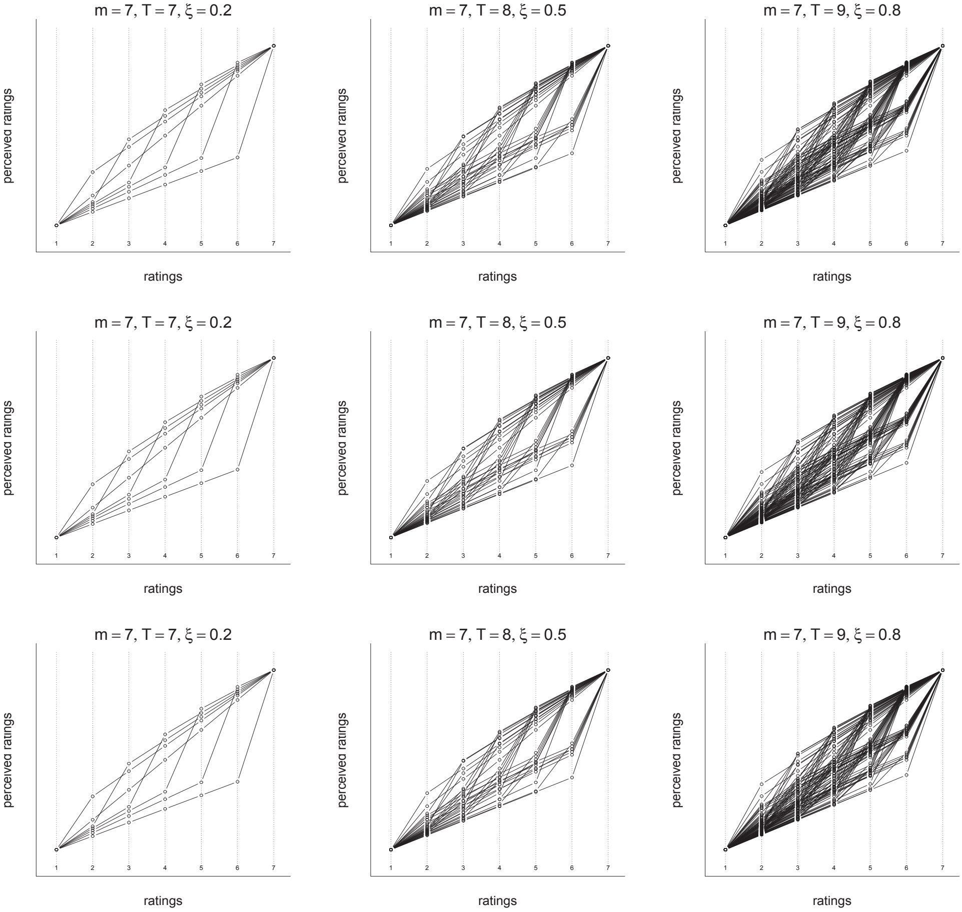


Figure 5. Transition plots for the cases with $m = 7$*,* $T = 7, 8, 9$ *(top, middle, bottom),* $\xi = 0.2, 0.5, 0.8$ *(left, middle, right)*

correlation between $\lambda(\xi, \mathbf{g})$ and ξ results -0.9206, -0.9529, -0.9664, -0.9754 for $m = 4, 5, 6, 7$, respectively. The relationship between $\lambda(\xi, \mathbf{g})$ and T can be evaluated by inspecting the boxplots in Figure 8, showing that when m increases, we need higher and higher values of T to reach the maximum level of nonlinearity. On the other hand, the median values of $\lambda(\xi, \mathbf{g})$ seem to increase very slowly after the first values of T.

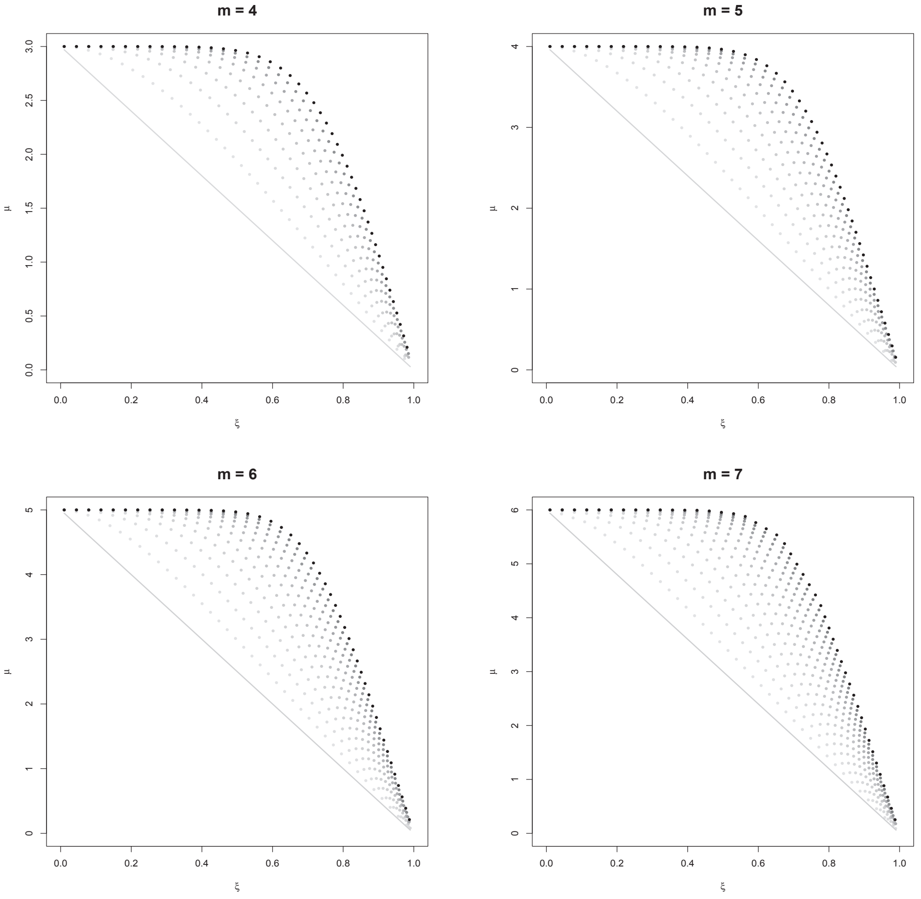


Figure 6. Relationship between ξ and μ (gray-level scale proportional to the values of T; solid line denoting the configuration with $T = m - 1$, i.e. the CUB model) for $m = 4, 5, 6, 7$

4.3. Remarks about the choice of T_{max}

The estimation procedure for the NLCUB models requires the definition of the maximum value T_{max} of T. This choice is rather crucial as high values of T may cause both identifiability and overfitting problems (Manisera and Zuccolotto, 2014), which can be kept under control by forcing T within a given range. The choice of a relatively small value for T_{max} is also justified from the point of view of the unconscious DP, since the commitment of respondents in formulating judgments is generally moderate, so it is rea-

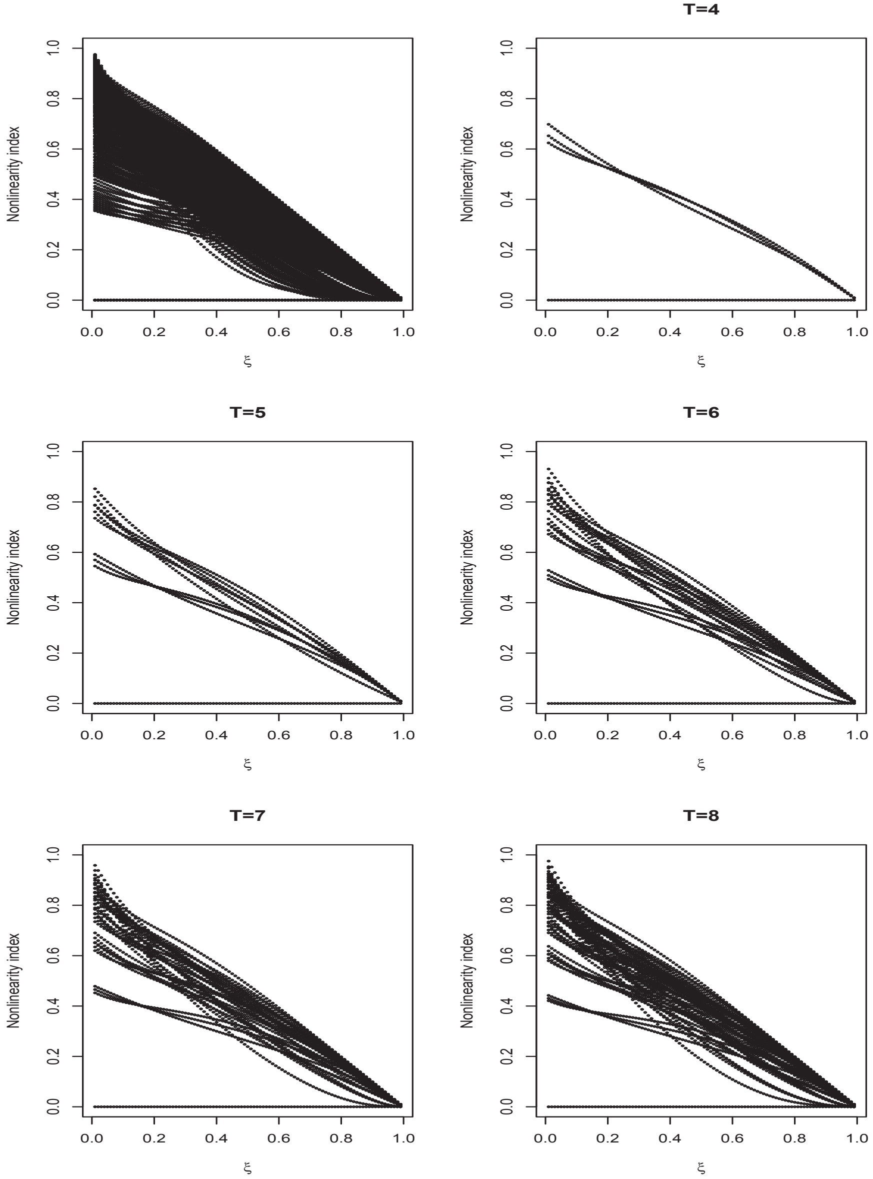


Figure 7. Overall plot of $\lambda(\xi, \mathbf{g})$ versus ξ (top left panel) and partial plots for some selected values of T $(m = 4)$

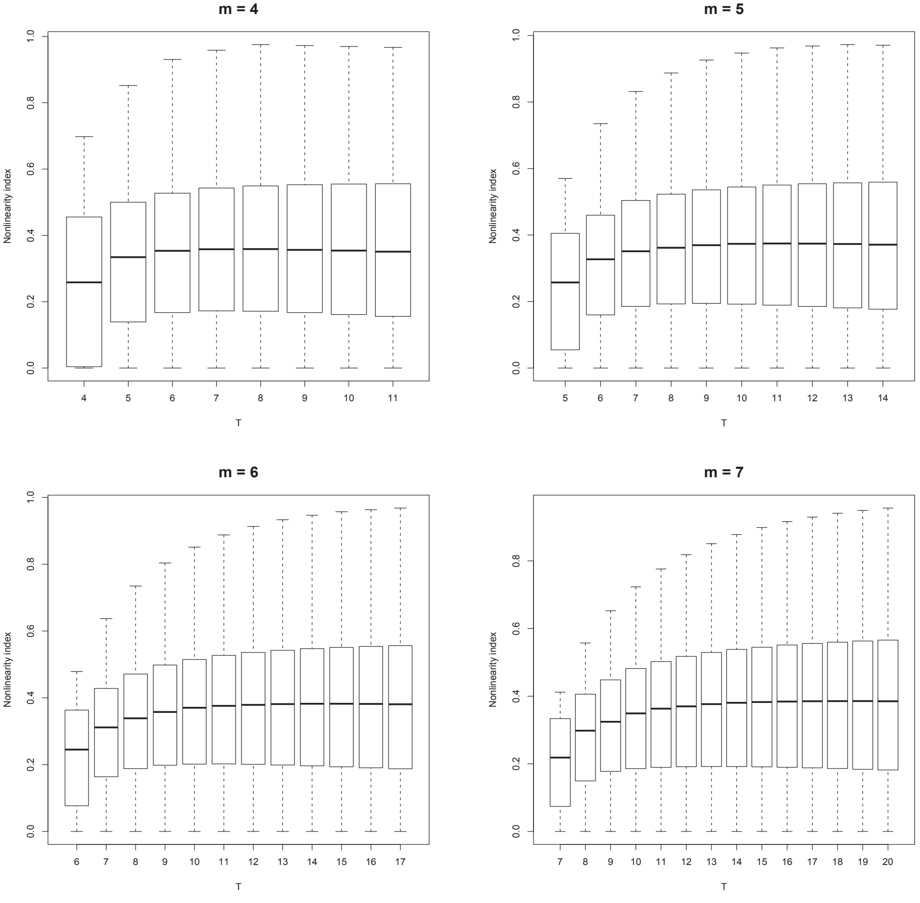


Figure 8. Boxplots of $\lambda(\xi, \mathbf{g})$ *given* T *for* $m = 4, 5, 6, 7$

sonable to assume a limited number of steps in the feeling path, whatever the complexity of the evaluated item.

The results of the systematic analysis carried out in this work can provide some useful suggestions about the choice of T_{max}. In summary, we have to balance two opposite needs:

- to define a model with flexibility enough to reproduce several nonlinear patterns: this requires to fix a high value for T_{max}, which allows the nonlinearity index $\lambda(\xi, \mathbf{g})$ to cover a wider range of values, as pointed out in Subsection 4.2;
- to pay attention on identifiability and overfitting problems: this requires to fix a

low value for T_{max}.

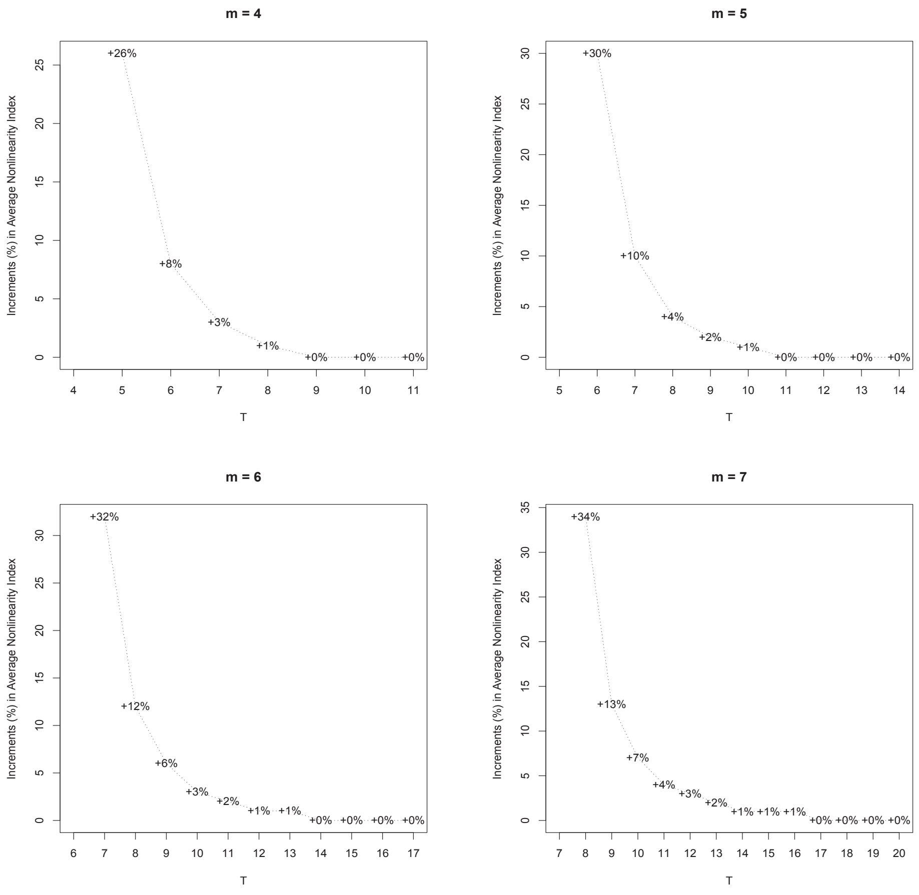


Figure 9. Increments in the average of $\lambda(\xi, \mathbf{g})$ given T, when moving from $T - 1$ to T, for $m = 4, 5, 6, 7$

Figure 9 shows how the average of $\lambda(\xi, \mathbf{g})$ given T increases when moving from $T - 1$ to T. We immediately note that the increments tend to be negligible after the first values of T. We feel that a good balance between the above mentioned opposite needs can be found when T_{max} approximately equals $2m$. In addition, if we establish that NLCUB models should be flexible enough to guarantee a nonlinearity index with an acceptably large range (from 0 to, at least, 85-95%) for varying m, $T_{max} = 2m - 1$ seems preferable (Table 3). Being aware that the choice of T_{max} is, to a certain extent, discretionary, we are convinced that T_{max} could be conveniently be set at $2m - 1$. This corresponds to select a Shifted Binomial random variable with twice the categories of the response scale. Additionally, although the estimation procedure of the NLCUB model is fairly not time-consuming, this choice allows to keep the number of possible configurations of $g_1, \ldots, g_m$ reasonably low, so reducing computational time.

Table 3. Values of the nonlinearity index (in %) for some m and T; only the values in [85%,95%] are displayed

T	$m = 4$	$m = 5$	$m = 6$	$m = 7$
$2m - 3$	85			
$2m - 2$	93	89	85	
$2m - 1$	95	93	89	85
$2m$		95	91	88
$2m + 1$			93	90
$2m + 2$			94	92
$2m + 3$			95	93
$2m + 4$				94
$2m + 5$				95
$2m + 6$				95

5. Conclusions

In this paper we have presented a systematic analysis of some main features of the Nonlinear CUB models (NLCUB), aimed at giving insights on the behaviour of this new class of models and suggestions about the future theoretical developments.

In detail, we have explored three issues, concerned with (1) the existence of linear DPs within the NLCUB framework, (2) the nonlinearity patterns expressed by different NLCUB models, (3) the choice of the value T_{max} in the estimation procedure.

The computational method exploited in this study was not, as usual, simulation. In fact, we have derived the theoretical statistical features of all the NLCUB models obtained by varying the parameters values, so that over 13 millions different models have been included in the analysis.

About point (1), we have demonstrated that the CUB case is not the unique linear DP within the NLCUB framework, thus confirming the need of devoting future research to the definition of a sufficient and necessary condition for linearity, whose possible formulation can be conjectured relying on the presented empirical evidences.

Points (2) and (3) are strictly connected each other. In fact, we have found that both the parameters T and ξ play an important role in the nonlinearity pattern of the NLCUB models. This evidence, although being of limited importance with reference to ξ, whose parameter space is restricted to $[0, 1]$, is very meaningful for what concerns T. It is then able to give some suggestions about the choice of T_{max} in the estimation procedure (at least for the explored values of m, which are, however, the most common ones in real situations).

Starting from this systematic study, future research can be devoted to derive a theo-retical formalization of the obtained empirical evidence.

Acknowledgements

Research funded by STAR project (University of Naples Federico II - CUP: E68C13000020003).

References

Agresti, A. (2013). *Categorical Data Analysis*, 3^{rd} edition, J. Wiley & Sons, New York.

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

Cerchiello, P., Iannario, M., Piccolo, D. (2010). Assessing risk perception by means of ordinal models, in: M. Corazza, C. Pizzi (eds.): *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, Springer-Verlag, pp.75–83.

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

Grilli, L., Iannario, M., Piccolo, D., Rampichini, C. (2013). Latent class cub models, *Advances in Data Analysis and Classification*, **8**, 105-119.

Iannario M. (2007) Dummy variables in CUB models, *Statistica*, LXVIII, 2.

IannarioM. (2008) A class of models for ordinal variables with covariates effects, *Quaderni di Statistica*, **10**, 53–72.

Iannario, M. (2009). Fitting measures for ordinal data models, *Quaderni di Statistica*, **11**, 39–72.

Iannario, M. (2010). On the identifiability of a mixture model for ordinal data, *Metron*, **LXVIII**, 87–94.

Iannario, M. (2012a). Modelling shelter choices in a class of mixture models for ordinal responses, *Statistical Methods and Applications*, **20**, 1–22.

Iannario, M. (2012b). Hierarchical CUB Models for Ordinal Variables, *Communication in Statistics - Theory and Methods*, **41**, 3110–3125.

Iannario, M., Manisera, M., Piccolo, D., Zuccolotto, P. (2012). Sensory analysis in the food industry as a tool for marketing decisions, *Advances in Data Analysis and Classification*, **6**, 303–321.

Iannario, M., Piccolo, D. (2014). A Short Guide to CUB 3.0 Program. Available at https://www.researchgate.net/publication/260959050.

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

Iannario, M., Piccolo, D. (2012). CUB Models: Statistical Methods and Empirical Evidence, in: R. S. Kenett, S. Salini (eds.): *Modern Analysis of Customer Surveys*, NY: Wiley, pp. 231–258.

Manisera, M., Zuccolotto, P. (2013). Nonlinear CUB models. In: T. Minerva, I. Morlini, F. Palumbo (eds.): *Book of Abstracts Cladag 2013, 9th Meeting of the Classification and Data Analysis Group*, CLEUP, pp. 288–291.

Manisera, M., Zuccolotto, P. (2014). Modeling rating data by Nonlinear CUB models, *Computational Statistics and Data Analysis*, **78**, 100-118.

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. (2013). Inferential issues on CUBE models with covariates, *Communication in Statistics - Theory and Methods*, in press.

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

Tutz, G. (2012). Regression for categorical data. Cambridge University Press. Cambridge.

QdS – Vol. 15, 2013

On the use of the generalized linear exponential cluster-weighted model to asses local linear independence in bivariate data

Antonio Punzo
Department of Economics and Business, University of Catania
E-mail: antonio.punzo@unict.it

Salvatore Ingrassia
Department of Economics and Business, University of Catania
E-mail: s.ingrassia@unict.it

Summary: The generalized linear exponential cluster-weighted model is a recent mixture-based approach which allows for flexible clustering and distribution estimation of a bivariate random vector composed by a response and by a covariate, regardless from the support of these variables. Examples concern a count response and a covariate taking values on the positive real line. With respect to the exponential-exponential latent class model, which is based on the assumption of local independence, the present approach assumes that, in each mixture component, there is a (generalized) linear dependence of the response given the covariate. Since the two models can be considered as nested, in this paper the BIC will be adopted to select the best assumption for data at hand. The procedure is illustrated through an application to real data from a survey on fair-trade coffee consumers interviewed at stores.

Keywords: Cluster-weighted models; Generalized linear models; Local independence; Mixture models with random covariates; Model-based clustering; Mixed-type data.

1. Introduction

Finite mixture models are commonly employed in statistical modeling with two different purposes (Titterington *et al.*, 1985, pp. 2–3). In *indirect applications*, they are used as semiparametric competitors of nonparametric density estimation techniques (see Titterington *et al.*, 1985, pp. 28–29, McLachlan and Peel, 2000, p. 8 and Escobar and West, 1995). On the other hand, in *direct applications*, finite mixture models are considered as a powerful device for clustering, classification, and discriminant analysis, by assuming that each mixture component represents a group (or cluster) in the original data (see Fraley and Raftery, 1998 and McLachlan and Basford, 1988). The areas of application of mixture models is huge and range from biology and medicine (see Schlattmann, 2009) to economics and marketing (see Wedel and Kamakura, 2001); an overview is given in McLachlan and Peel (2000) and Frühwirth-Schnatter (2006).

The framework is represented by data arising from a bivariate random vector $(X,Y)'$, taking value on a subset of $\mathbb{R}\times\mathbb{R}$ and having joint distribution $p\left(x,y\right)$, where Y is the response variable and X is the (random) covariate. The family of mixture models with random covariates (see, *e.g.*, Hennig, 2000), to which the cluster-weighted model (CWM; Gershenfeld 1997) belongs, constitutes a flexible frame to analyze such data. In particular, the CWM factorizes $p\left(x,y\right)$, in each mixture component, into the product between the conditional distribution of $Y|X=x$ and the marginal distribution of X by assuming a parametric functional dependence for $E\left(Y|x\right)$. Some recent works about the CWM can be found in Ingrassia *et al.* (2012, 2014), Punzo (2014), and Subedi *et al.* (2013).

This paper focuses on the generalized linear exponential CWM (GLECWM; Punzo and Ingrassia, 2013) which considers an exponential family distribution for both $Y|x$ and X in each mixture component. This implies the possibility to jointly model two variables defined on different supports such as, for example, a count response (via a Poisson distribution) and a strictly positive covariate (by a gamma density). Furthermore, since in regression terms the exponential family is strictly related to the generalized linear models, this means that the functional relationship for $E\left(Y|x\right)$ within each mixture component is modeled by a generalized linear model. Hence, the resulting approach is based on the assumption of local (generalized) linear dependence of Y on x. As a special case, when the component slopes are assumed to be null, the exponential-exponential latent class model (EELCM; Punzo and Ingrassia, 2013) is obtained. The EELCM fits $p\left(x,y\right)$ by a mixture in which each component joint distribution is factorized as the product of the univariate exponential family distributions chosen for X and Y. Thus, the EELCM is based on the stronger assumption of local independence (see Vermunt and Magidson, 2002 and Hennig and Liao, 2013).

By considering the Bayesian information criterion (BIC; Schwarz, 1978), the idea of the present paper is to evaluate if for (heterogeneous) data at hand the assumption of local independence (as induced by the EELCM), with respect to the weaker assumption of local (generalized) linear dependence (as induced by the GLECWM), is too strong.

The paper is organized as follows. In Section 2, the GLECWM is summarized. In Section 3, the EM algorithm for maximum likelihood parameters estimation is described (see also Section 4 for computational details). The BIC is recalled in Section 5 and, based on a real data set, the BIC-based approach above outlined is illustrated in Section 6. Finally, discussion and suggestions for further work are presented in Section 7.

2. Genesis and formulation of the model

2.1. The cluster-weighted model

Let

$$p(x,y;\boldsymbol{\psi}) = \sum_{j=1}^{k} \pi_j p\left(x,y;\boldsymbol{\xi}_j\right) \tag{2.1}$$

be the finite mixture of distributions, with k components, used to estimate $p(x,y)$. In (2.1), $p\left(x,y;\boldsymbol{\xi}_j\right)$ is the parametric distribution (with respect to $\boldsymbol{\xi}_j$) associated with the jth component, π_j is the weight of the jth component, with $\pi_j > 0$ and $\sum_{j=1}^{k} \pi_j = 1$, and $\boldsymbol{\psi} = \left(\boldsymbol{\pi}', \boldsymbol{\xi}'\right)'$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_k)'$, contains all of the unknown parameters of the mixture. Model (2.1) implicitly assumes that the component joint distributions belong to the same parametric family. The mixture model (2.1) is called a CWM when:

- for each j, there is a parametric function

  $$E\left(Y|x,j;\boldsymbol{\beta}_j\right) = \mu_{Y|j}\left(x;\boldsymbol{\beta}_j\right)$$

  relating the expected value of the response Y to the covariate, where $\boldsymbol{\beta}_j$ are regression parameters.

- the jth component joint distribution is factorized as

  $$p\left(x,y;\boldsymbol{\xi}_j\right) = p\left(y|x;\boldsymbol{\xi}_{Y|j}\right) p\left(x;\boldsymbol{\xi}_{X|j}\right),$$

  where $\boldsymbol{\xi}_j = \left\{\boldsymbol{\xi}_{X|j}, \boldsymbol{\xi}_{Y|j}\right\}$, with $\boldsymbol{\xi}_{Y|j}$ containing $\boldsymbol{\beta}_j$.

Hence, the CWM has the form

$$p(x,y;\boldsymbol{\psi}) = \sum_{j=1}^{k} \pi_j p\left(y|x;\boldsymbol{\xi}_{Y|j}\right) p\left(x;\boldsymbol{\xi}_{X|j}\right). \tag{2.2}$$

2.2. Exponential family and “link” with generalized linear models

A random variable Z is in the exponential family if it has density function $p(z)$ of the form

$$p(z;\theta,\phi) = \exp\left\{\frac{z\theta - b(\theta)}{a(\phi)} + c(z;\phi)\right\}, \tag{2.3}$$

for specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. In particular, if ϕ is known, this is an exponential family with canonical parameter θ which is function of the location parameter of the distribution. It may or may not be a two-parameter exponential family if ϕ is unknown. The function $b(\cdot)$ is of special importance because it describes the relationship between the mean value and the variance in the distribution (see, *e.g.*, McCullagh and Nelder, 1989, pp. 28–29). Moreover, $E(Z) = \mu = b'(\theta)$ and $Var(Z) = b''(\theta)\, a(\phi)$, where primes denote differentiation with respect to θ. The parameter ϕ is called the *dispersion parameter* and $b''(\theta)$ is called the *variance function*. The variance function is often written as $V(\mu) = b''(\theta)$, where the notation $V(\mu)$ does not mean "the variance of μ"; rather, $V(\mu)$ indicates how the variance depends on μ in the distribution, and μ is in turn a function of θ (see also Olsson, 2002, pp. 37–40).

It is well known that (2.3) is strictly related to the generalized linear models. Here, a monotone and differentiable link function $g(\cdot)$ is introduced which relates the expected value μ, of the response Z, to the covariate X through the relation

$$g(\mu; \boldsymbol{\beta}) = \eta = \beta_0 + \beta_1 x. \tag{2.4}$$

In (2.4), for simplicity, we consider $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Since the interest is now in the parameters $\boldsymbol{\beta}$, the distribution of Z given x will be denoted by $p(z|x; \boldsymbol{\beta}, \phi)$. The choice of the link function depends on the type of data. However, certain link functions are, in some sense, "natural" for certain distributions and they are called *canonical links*. In particular, the canonical link is the function $g(\cdot)$ such that $g(\mu; \boldsymbol{\beta}) = \theta$. Table 1 summarizes all the quantities discussed so far for a few well-known distributions in the exponential family. It should be noted, however, that there is no guarantee that the canonical links will always provide the "best" model for a given set of data. In any particular application the data may exhibit peculiar behavior, or there may be theoretical justification for choosing links other than the canonical ones. For example, in the case of the gamma distribution, the domain of the canonical link function is not the same as the permitted range of the mean. In particular, the linear predictor may be negative, which would give an impossible negative mean. When maximizing the likelihood, precautions must be taken to avoid this. An alternative is to use a noncanonical link function like the "log".

Table 1. Characteristics of some common distributions in the exponential family

	Gaussian	gamma	Poisson	binomial
Notation	$N(\mu, \sigma^2)$	$G(\mu, \nu)$	$P(\mu)$	$B(m, p)/m$
Support of Z	$(-\infty, \infty)$	$(0, \infty)$	$\{0, 1, \ldots\}$	$\{0/m, 1/m, \ldots, m/m\}$
$a(\phi)$	σ^2	ν^{-1}	1	$1/m$
$b(\cdot)$	$\theta^2/2$	$-\ln(-\theta)$	$\exp(\theta)$	$\ln[1 + \exp(\theta)]$
$c(z; \phi)$	$-\frac{1}{2}\left[\frac{z^2}{\phi} + \ln(2\pi\phi)\right]$	$\nu \ln(\nu z) - \ln(z) - \ln[\Gamma(\nu)]$	$-\ln(z!)$	$\ln\binom{m}{mz}$
$\mu(\theta)$	θ	$-\theta^{-1}$	$\exp(\theta)$	$\frac{\exp(\theta)}{1 + \exp(\theta)}$
Canonical link	identity	μ^{-1}	log	logit
$Var(Z)$	σ^2	μ^2/ν	μ	$p(1-p)/m$
$V(\mu)$	1	μ^2	μ	$\mu(1-\mu)$

2.3. The generalized linear exponential CWM

Consider the general formulation of a CWM in (2.2) and assume that i) the distribution of X, given the component j, is within the exponential family (2.3), and ii) a generalized linear model for Y on x for each j. Thus we obtain the model

$$p\left(x,y;\boldsymbol{\psi}\right)=\sum_{j=1}^{k}\pi_j p\left(y\left|x;\boldsymbol{\beta}_j,\phi_{Y|j}\right.\right)p\left(x;\theta_{X|j},\phi_{X|j}\right), \tag{2.5}$$

which will be referred to as the *generalized linear exponential CWM*. With respect to model (2.2) we have: $\boldsymbol{\xi}_{Y|j}=\left(\boldsymbol{\beta}_j',\phi_{Y|j}\right)'$ and $\boldsymbol{\xi}_{X|j}=\left(\theta_{X|j},\phi_{X|j}\right)'$, with $\boldsymbol{\xi}_Y=\left(\boldsymbol{\xi}_{Y|1}',\dots,\boldsymbol{\xi}_{Y|k}'\right)'$ and $\boldsymbol{\xi}_X=\left(\boldsymbol{\xi}_{X|1}',\dots,\boldsymbol{\xi}_{X|k}'\right)'$. Note that the number of free parameters in (2.5) naturally depends on the exponential family distributions adopted. Sufficient conditions for the identifiability of model (2.5), under the the four distributions (binomial, Poisson, gamma, and Gaussian) of Table 1, are given in Punzo and Ingrassia (2013). Roughly speaking, the model is identifiable when the related mixture of generalized linear models is identifiable. In addition, if a binomial distribution is used for $Y|x$ (X) in each mixture component, then we have also to require that $k\leq\left(m_Y+1\right)/2$ ($k\leq\left(m_X+1\right)/2$), where m_Y (m_X) is the maximum value the binomial distribution for $Y|x$ (X) can assume.

If we assume in (2.5) that $\beta_{11}=\dots=\beta_{1k}=0$ (local independence), then the *exponential-exponential latent class model* (EELCM; Punzo and Ingrassia, 2013), of equation

$$p\left(x,y;\boldsymbol{\pi},\boldsymbol{\xi}_X,\widetilde{\boldsymbol{\xi}}_Y\right)=\sum_{j=1}^{k}\pi_j p\left(y;\theta_{Y|j},\phi_{Y|j}\right)p\left(x;\theta_{X|j},\phi_{X|j}\right), \tag{2.6}$$

is obtained. In (2.6), it results $\widetilde{\boldsymbol{\xi}}_{Y|j}=\left(\theta_{Y|j},\phi_{Y|j}\right)'$ and $\widetilde{\boldsymbol{\xi}}_Y=\left(\widetilde{\boldsymbol{\xi}}_{Y|1}',\dots,\widetilde{\boldsymbol{\xi}}_{Y|k}'\right)'$. Hence, given a value for k, model (2.6) is nested in model (2.5).

3. The EM algorithm for maximum likelihood estimation

Given n observed pairs $\left(x_1,y_1\right)',\dots,\left(x_n,y_n\right)'$ from $\left(X,Y\right)'$, the observed-data log-likelihood for the generalized linear exponential CWM, when k is supposed to be pre-assigned, can be written as

$$l\left(\boldsymbol{\psi}\right)=\sum_{i=1}^{n}\ln\left[\sum_{j=1}^{k}\pi_j p\left(y_i\left|x_i;\boldsymbol{\beta}_j,\phi_{Y|j}\right.\right)p\left(x_i;\theta_{X|j},\phi_{X|j}\right)\right].$$

The EM algorithm (Dempster *et al.*, 1977) can be used to maximize $l\left(\boldsymbol{\psi}\right)$ in order to find maximum likelihood (ML) estimates for the d unknown parameters of the GLECWM.

Once k is assigned, the EM algorithm basically takes into account the complete-data log-likelihood

$$\begin{aligned} l_c\left(\boldsymbol{\psi}\right) &= \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\ln\left(\pi_j\right) + \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\ln\left[p\left(y_i\left|x_i;\boldsymbol{\beta}_j,\phi_{Y|j}\right.\right)\right] \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij}\ln\left[p\left(x_i;\theta_{X|j},\phi_{X|j}\right)\right], \end{aligned} \tag{3.1}$$

where $z_{ij} = 1$ if $(x_i, y_i)'$ comes from component j and $z_{ij} = 0$ otherwise. The E and M steps of the algorithm can be detailed as follows.

3.1. E-step

The E-step, on the $(q+1)$th iteration ($q = 0, 1, \ldots$), requires the calculation of the expectation of $l_c\left(\boldsymbol{\psi}\right)$ given the observed data $(x_1, y_1)', \ldots, (x_n, y_n)'$ and given also the provisional estimate $\boldsymbol{\psi}^{(q)}$, of $\boldsymbol{\psi}$, arising from the previous M-step. As $l_c\left(\boldsymbol{\psi}\right)$ is linear in the unobservable data z_{ij}, the E-step simply requires the calculation of the current conditional expectation of Z_{ij} given the observed sample, where Z_{ij} is the random variable corresponding to z_{ij}. In particular, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$, it follows that

$$E_{\boldsymbol{\psi}^{(q)}}\left[Z_{ij}\left|(x_i, y_i)'\right.\right] = \tau_{ij}^{(q)} = \frac{\pi_j^{(q)} p\left(y_i\left|x_i;\boldsymbol{\beta}_j^{(q)},\phi_{Y|j}^{(q)}\right.\right) p\left(x_i;\theta_{X|j}^{(q)},\phi_{X|j}^{(q)}\right)}{p\left(x_i, y_i;\boldsymbol{\psi}^{(q)}\right)}, \tag{3.2}$$

which corresponds to the posterior probability that the unlabeled observation $(x_i, y_i)'$ belongs to the jth component of the mixture, using the current fit $\boldsymbol{\psi}^{(q)}$ for $\boldsymbol{\psi}$.

3.2. M-step

In the M-step, on the $(q+1)$th iteration ($q = 0, 1, \ldots$), to maximize the expectation of l_c with respect to $\boldsymbol{\psi}$, the values z_{ij} in (3.1) are simply replaced by their current expectations $\tau_{ij}^{(q)}$ obtained in (3.2). This leads to

$$\begin{aligned} E\left[l_c\left(\boldsymbol{\psi}\right)\right] &= \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{ij}^{(q)}\ln\left(\pi_j\right) + \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{ij}^{(q)}\ln\left[p\left(y_i\left|x_i;\boldsymbol{\beta}_j,\phi_{Y|j}\right.\right)\right] \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{ij}^{(q)}\ln\left[p\left(x_i;\theta_{X|j},\phi_{X|j}\right)\right]. \end{aligned} \tag{3.3}$$

Since the three terms on the right-hand side have zero cross-derivatives, they can be maximized separately. Let us set $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_k\right)'$, $\boldsymbol{\phi}_Y = \left(\phi_{Y|1}, \ldots, \phi_{Y|k}\right)'$, $\boldsymbol{\theta}_X = \left(\theta_{X|1}, \ldots, \theta_{X|k}\right)'$ and $\boldsymbol{\phi}_X = \left(\phi_{X|1}, \ldots, \phi_{X|k}\right)'$. The maximum of equation (3.3) with respect to $\boldsymbol{\pi}$, subject to the constraints on those parameters, is obtained by maximizing the augmented function

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{ij}^{(q)} \ln\left(\pi_j\right) - \lambda\left(\sum_{j=1}^{k}\pi_j - 1\right), \tag{3.4}$$

where λ is a Lagrangian multiplier. Setting the derivative of equation (3.4) with respect to π_j equal to zero and solving for π_j yields

$$\pi_j^{(q+1)} = \sum_{i=1}^{n} \tau_{ij}^{(q)} \Big/ n, \quad j = 1, \ldots, k.$$

Maximizing (3.3) with respect to $\boldsymbol{\beta}$ (and, eventually, to $\boldsymbol{\phi}_Y$) is equivalent to independently maximizing each of the k expressions

$$l_c^{(Y,j)} = \sum_{i=1}^{n} \tau_{ij}^{(q)} \ln\left[p\left(y_i \left| x_i; \boldsymbol{\beta}_j, \phi_{Y|j}\right.\right)\right], \quad j = 1, \ldots, k. \tag{3.5}$$

The maximization of (3.5) is equivalent to the maximization problem of the generalized linear model (for the complete data), except that each observation $(x_i, y_i)'$ contributes to the log-likelihood for each component with a known weight $\tau_{ij}^{(q)}$. Maximization of (3.5), with respect to $\boldsymbol{\beta}_j$ (and, eventually, to $\phi_{Y|j}$ if the adopted distribution is not fixed in the exponential family) can be carried out numerically; details can be found in Wedel and DeSarbo (1995) and Wedel and Kamakura (2001, pp. 120–124) since the GLECWM shares this part of complete-data log-likelihood with the finite mixture of generalized linear models discussed in these references.

Finally, maximizing (3.3) with respect to $\boldsymbol{\theta}_X$ (and, eventually, to $\boldsymbol{\phi}_X$) is equivalent to independently maximizing each of the k expressions

$$l_c^{(X,j)} = \sum_{i=1}^{n} \tau_{ij}^{(q)} \ln\left[p\left(x_i; \theta_{X|j}, \phi_{X|j}\right)\right], \quad j = 1, \ldots, k. \tag{3.6}$$

The maximization of (3.6) is equivalent to the maximization problem of the exponential family (for the complete data), except that each observation x_i contributes to the log-likelihood for each component with a known weight $\tau_{ij}^{(q)}$, which is obtained in the preceding E-step. Maximization of (3.6), with respect to $\theta_{X|j}$ (and, eventually, to $\phi_{X|j}$) can be carried out, as before, numerically.

3.3. Classification of units

In the framework of model-based clustering, the fitted model can be used to classify the n observations via the maximum *a posteriori* (MAP) classification induced by the final estimates (denoted, as usual, by hats)

$$\text{MAP}\left(\widehat{\tau}_{ij}\right) = \begin{cases} 1 & \text{if } \max_h \left\{\widehat{\tau}_{ih}\right\} \text{ occurs at component } j \\ 0 & \text{otherwise} \end{cases}, \tag{3.7}$$

$i = 1, \ldots, n$ and $j = 1, \ldots, k$. Note that the MAP classification will be used in the analyses of Section 6.

4. Computational issues

Code for the EM algorithm described in Section 3 is written in the R computing environment (R Core Team, 2013), and it is available at `http://www.dei.unict.it/punzo`. In its actual version, the algorithm considers the simple four distributions in Table 1; they are able to cover four different data supports, as highlighted in the second row of Table 1. Hence, we obtain sixteen different bivariate distributions for $(X, Y)'$. For the gamma distribution note that, as motivated at the end of Section 2.2, the "log" link is considered. Parameter recovery of the EM algorithm discussed in Section 3 is investigated, via simulations, by Punzo and Ingrassia (2013).

4.1. EM initialization

Before running the EM algorithm, the choice of the starting values constitutes an important issue. The standard initialization consists in selecting a value for $\boldsymbol{\psi}^{(0)}$. An alternative approach (see McLachlan and Peel, 2000, p. 54) consists in performing the first M-step by specifying, in equation (3.2), the values of $\tau_{ij}^{(0)}$, $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Among the possible initialization strategies (see Biernacki *et al.* 2003, Karlis and Xekalaki 2003, and Bagnato and Punzo, 2013 for details) – according to the R-package **flexmix** (Leisch, 2004 and Grün and Leisch, 2008) which allows to estimate finite mixtures of generalized linear models – a random initialization is repeated t times from different random positions and the solution maximizing the observed-data log-likelihood among these t runs is selected. In each run, the n vectors $\boldsymbol{\tau}_i^{(0)} = \left(\tau_{i1}^{(0)}, \ldots, \tau_{ik}^{(0)}\right)'$ can be randomly generated in a "soft" way, that is by generating k positive values summing to one, or alternatively in a "hard" way by randomly drawn a single observation from a multinomial distribution with probabilities $(1/k, \ldots, 1/k)'$.

4.2. Convergence criterion

The Aitken acceleration procedure (Aitken, 1926) is used to estimate the asymptotic maximum of the log-likelihood at each iteration of the EM algorithm. Based on this estimate, a decision can be made regarding whether or not the algorithm has reached convergence; that is, whether or not the log-likelihood is sufficiently close to its estimated asymptotic value. The Aitken acceleration at iteration $q+1$, $q=0,1,\ldots$, is given by

$$a^{(q+1)}=\frac{l^{(q+2)}-l^{(q+1)}}{l^{(q+1)}-l^{(q)}},$$

where $l^{(q+2)}$, $l^{(q+1)}$, and $l^{(q)}$ are the log-likelihood values from iterations $q+2$, $q+1$, and q, respectively. Then, the asymptotic estimate of the log-likelihood at iteration $q+2$ (Böhning et al., 1994) is given by

$$l_{\infty}^{(q+2)}=l^{(q+1)}+\frac{1}{1-a^{(q+1)}}\left(l^{(q+2)}-l^{(q+1)}\right).$$

In the analyses in Section 6, we follow Subedi *et al.* (2013) and stop our algorithms when $l_{\infty}^{(q+2)}-l^{(q+1)}<\epsilon$, with $\epsilon=0.05$.

4.3. Standard errors of the estimates

Once the EM algorithm is run, the covariance matrix of $\widehat{\boldsymbol{\psi}}$ is calculated using the inverted negative Hessian matrix. The Hessian matrix is computed by the function `hessian()` in the R-package **numDeriv**. In particular, `hessian()` is evaluated on the observed-data log-likelihood in correspondence to the solution provided by the EM algorithm.

5. Model selection and clustering performance

The GLECWM, in addition to $\boldsymbol{\psi}$, is also characterized by the number of components k. So far, this quantity has been treated as *a priori* fixed. Nevertheless, for practical purposes, choosing a relevant model needs its choice. In model-based clustering, model selection criteria are commonly used with this end. Among them, we will adopt the Bayesian information criterion (BIC; Schwarz, 1978)

$$\text{BIC}=2l\left(\widehat{\boldsymbol{\psi}}\right)-d\ln\left(n\right),$$

where d represents the number of free parameters of the model. The performance of the BIC, in comparison with other famous likelihood-based information criteria, is evaluated by simulations in Punzo and Ingrassia (2013).

In order to evaluate the clustering performance in cases in which the true classification is known, the adjusted Rand index (ARI; Hubert and Arabie, 1985), and the misclassification rate (proportion of misallocated observations), will be taken into account. We recall that the ARI has an expected value of 0 and perfect classification would result in a value equal to 1. The ARI and the misclassification rate will be respectively computed according to the functions `adjustedRandIndex()` and `classError()` of the **mclust** package for R.

6. Real data analysis

We present an application of our approach in modeling a real data set based on a sample of $n = 224$ fair-trade coffee consumers interviewed at stores. These data were first considered in Cicia *et al.* (2010). The variables of interest are: *importance that respondents attribute to price in their fair-trade coffee purchase* (Y^*; measured on a 1–7 Likert scale where 1 indicated "completely unimportant" and 7 "extremely important") and *number of fair-trade coffee packages over 10 purchases* (X^*; those who did not consume coffee were excluded from the survey leading to values ranging from 1 to 10). Cicia *et al.* (2010), by using the CUB model proposed in D'Elia and Piccolo (2005), studied the probability mass function of Y^* as function of a dichotomous version of X^* assuming value 1 if $X^* \leq 4$, and 0 otherwise. In particular, the authors conducted two separated analyses, one for consumers with $X^* \leq 4$, and the other for consumers with $X^* > 4$, so separating *a priori* with respect to a possible group variable. In our analysis, X^* and Y^* are respectively replaced by $X = X^* - 1$ and $Y = Y^* - 1$. In each mixture component of the CWM, a Binomial distribution is adopted with the parameter "number of replications" (see Table 1) fixed at $m_X = 9$, for X, and $m_Y = 6$ for Y. In contrast with Cicia *et al.* (2010), we consider the relation between Y and X without imposing an *a priori* group structure. The aim is also to compare our clustering/classification results with respect to the dichotomization of these authors.

Observed data are displayed in Figure 1 by the CW-plot, a graphical representation proposed in Ingrassia *et al.* (2014) to highlight the key aspects on which the CWM is based on. The top panel of the CW-plot in Figure 1 shows the bar plot of the empirical probability mass function of X. The scatter plot of the data is displayed in the bottom panel; here, the size of each point is proportional to the joint frequency of the pair (x, y), for $x = 0, 1, \ldots, 9$ and $y = 0, 1, \ldots, 6$. This also gives information about the probability mass of the pairs. The models fitted on these data are the Binomial-Binomial LCM and the Binomial-Binomial CWM. In order to fulfill the identifiability constraint given at the end of Section 2.3, the considered values for k range from 1 to 3. Table 2 gives the BIC values for the two models. For both of them, the choice $k = 2$ is the best one and a strong improvement, of the considered criterion, is obtained by moving from $k = 1$ to $k = 2$; this justifies the presence of a group structure. The CW-plots of the best BIC models are given Figure 2. The two approaches provide similar results, especially

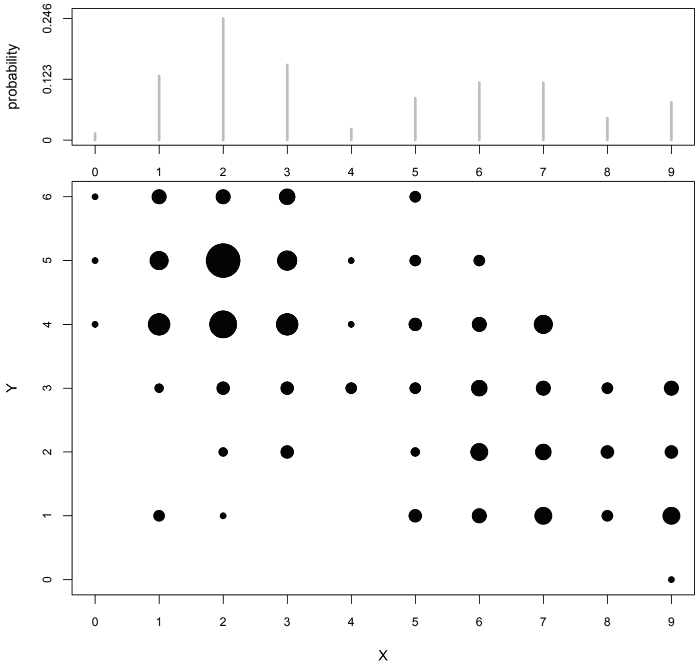
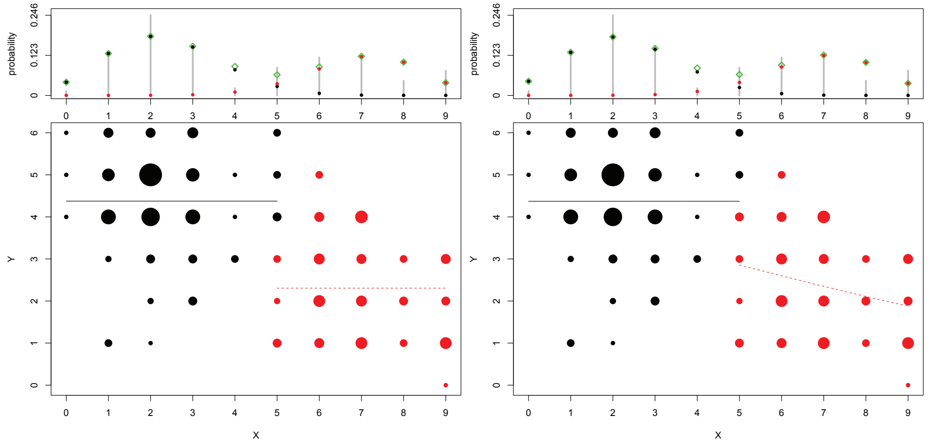


Figure 1. CW-plot of the real data ($n = 224$). In the scatter plot, the point size is proportional to the number of occurrences.

k \ models	Binomial-Binomial LCM	Binomial-Binomial CWM
1	-2119.759	-1990.296
2	-1709.480	-1715.345
3	-1709.673	-1731.885

Table 2. BIC values for the fitted models.

with reference to the regression model in the "black" group. Moreover, for both models, the obtained classification is sufficiently in agreement with the dichotomization of Cicia *et al.* (2010), as highlighted by the summary measures of Table 3. A slight higher agreement is obtained for the Binomial-Binomial CWM. Finally, since the Binomial-Binomial LCM can be seen as nested in the Binomial-Binomial CWM, likelihood-based quantities, such as the BIC, can be compared even if referred to different models. Thus, based on Table 2, we can select both the best number of mixture components and the best model. In our case, the best configuration is represented by the Binomial-Binomial



(a) Binomial-Binomial LCM ($k = 2$)

(b) Binomial-Binomial CWM ($k = 2$)

Figure 2. CW-plot of the fitted models on the real data.

measures \ models	Binomial-Binomial LCM	Binomial-Binomial CWM
ARI	0.749	0.812
misclassification rate	0.067	0.049

Table 3. Classification results with $k = 2$.

LCM with $k = 2$ components; furthermore, apart from the case $k = 1$, the Binomial-Binomial LCM is always better than its CWM counterpart for each considered value of k. For these data we can conclude that the stronger assumption of local independence can be assumed.

7. Discussion and future work

The generalized linear exponential cluster-weighted model of Punzo and Ingrassia (2013) constitutes a flexible mixture-based approach to model heterogeneous bivariate data (also arising by two variables defined on a different support). In this paper it has been considered, as a benchmark model, to investigate the assumption of local independence of the nested exponential-exponential latent class model. Based on a real data set illustrated in Cicia *et al.* (2010), investigation has been coped with the use of the BIC. Future work could involve a different approach based on the likelihood-ratio test; with this aim, since in the mixture context the χ^2 reference distribution gives reasonable approximation for the likelihood-ratio statistic only under more stringent conditions (see, *e.g.*, Lo, 2008), a parametric bootstrap approach may be considered.

References

Aitken, A. (1926). On Bernoulli's numerical solution of algebraic equations. In *Proceedings of the Royal Society of Edinburgh*, volume 46, pages 289–305.

Bagnato, L. and Punzo, A. (2013). Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. *Computational Statistics*, **28**(4), 1571–1597.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, **41**(3-4), 561–575.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**(2), 373–388.

Cicia, G., Corduas, M., Del Giudice, T., and Piccolo, D. (2010). Valuing consumer preferences with the cub model: A case study of fair trade coffee. *International Journal on Food System Dynamics*, **1**(1), 82–93.

D'Elia, A. and Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**(3), 917–934.

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

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**(430), 577–588.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *Computer Journal*, **41**(8), 578–588.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer, New York.

Gershenfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**(1), 18–24.

Grün, B. and Leisch, F. (2008). **FlexMix** version 2: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, **28**(4), 1–35.

Hennig, C. (2000). Identifiablity of models for clusterwise linear regression. *Journal of Classification*, **17**(2), 273–296.

Hennig, C. and Liao, T. F. (2013). How to find an appropriate clustering for mixed type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**(3), 1–25.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via the cluster-weighted approach with elliptical distributions. *Journal of Classification*, **29**(3), 363–401.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics and Data Analysis*, **71**, 159–182.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, **41**(3–4), 577–590.

Leisch, F. (2004). **FlexMix**: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, **11**(8), 1–18.

Lo, Y. (2008). A likelihood ratio test of a homoscedastic normal mixture against a heteroscedastic normal mixture. *Statistics and Computing*, **18**(3), 233–240.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, Boca Raton, 2nd edition.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and Applications to clustering*. Marcel Dekker, New York.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, New York.

Olsson, U. (2002). *Generalized Linear Models: An Applied Approach*. Lightning Source Incorporated, Sweden.

Punzo, A. and Ingrassia, S. (2013). Clustering Bivariate Mixed-Type Data via the Cluster-Weighted Model. Submitted to *Computational Statistics*.

Punzo, A. (2014). Flexible mixture modeling with the polynomial Gaussian cluster-weighted model. *Statistical Modelling*, **14**(3), 1–35.

Schlattmann, P. (2009). *Medical Applications of Finite Mixture Models*. Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

R Core Team (2013). R*: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, **7**(1), 5–40.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.

Vermunt, J. K. and Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars and A. L. McCutcheon, editors, *Applied Latent Class Analysis*, pages 89–106, Cambridge. Cambridge University Press.

Wedel, M. and DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, **12**(1), 21–55.

Wedel, M. and Kamakura, W. (2001). *Market Segmentation: Conceptual and Methodological Foundations*. Kluwer Academic Publishers, Boston, MA, USA, 2nd edition.

Clustering quantified ordinal data distributions

Rosanna Verde, Antonio Irpino and Antonio Balzanella
Department of Political Sciences "Jean Monnet", Second University of Naples
E-mails: rosanna.verde@unina2.it; antonio.irpino@unina2.it;
antonio.balzanella@gmail.it

Summary: This paper introduces a strategy for clustering grouped categorical ordinal data based on the partition of the set of distributions obtained by a quantification of ordinal categorical variables. The analyzed data are issued by the 2003 edition of the International Social Survey Programme studying the feelings of national identity and involving about 46 thousands respondents in 36 different countries. The ordinal categorical variables, corresponding to the judgment of each respondent to several questions, are measured on Likert-type scales. We propose to quantify them according to a procedure of Optimal Scaling, the Categorical Principal Component Analysis (CATPCA). From the results of the quantification step, we consider the distribution of individuals belonging to each country on the first two axes, for performing a partitioning of the countries. The main novelty of our proposal is that we use a Dynamic Clustering Algorithm which partitions the set of distributions describing the different countries, rather than the means of the country distributions. In the conclusions, we compare the proposed approach with a clustering algorithm performed on the means of the country distributions, in order to point out the advantages in considering distributions in the analysis.

Keywords: Categorical Principal Component Analysis; Optimal Scaling; Ordinal Data; Histogram-valued data; Clustering distribution data

1. Introduction

Researches, especially in social sciences field, usually work on survey data with a high number of respondents. An important task is to classify individuals into homogeneous classes, in order to deduce similar behaviors and identify typologies of respondents. In this area, researches are often conducted through questionnaires measuring judgments on ordinal scales. Thus, the recorded data often consists of qualitative or categorical variables that describe an individual through a limited number of categories. In the literature, it is known that working on non-numeric data implies "uncertainty" in the measurement scale: the zero point is uncertain; the relationships between categories is often no-well defined and the mutual distance between ordinal categories might be unknown. For this reason an important development in categorical and ordinal data analysis has been the optimal assignment of quantitative values to qualitative scale. In the literature, there are many optimal scaling (optimal scoring) approaches to deal with multivariate categorical data. The main references are to Hayashi (1952) and to Gifi's school (Gifi, 1981, 1990) but we also mention the papers by de Leeuw (1990) and Heiser and Meulman, (1994). With the aim of clustering survey data, through algorithms like k-means, or more in general, dynamic clusters methods, it is required a step of categorical (ordinal) variables quantification. In particular, in this paper we focus on the analysis of data issued by the International Social Survey Programme [ISSP, 2003] concerning the feeling of national identity. It involves about 46 thousands respondents in 36 different countries all over the world. The choice of concentrating our attention to this survey data, although it is not very recent, is motivated by the current debate, lively in Italy and in the other European countries, about the immigration problem after the tragic events of Lampedusa. Our proposal is based on the quantification of ordinal categorical variables by means of the Categorical Principal Component Analysis (CATPCA) technique. The choice of using this method is consistent with the approach proposed by Meulman (Meulman et al., 2004) in the framework of ALSOS methods for categorical variable quantification. Our aim is to introduce a new strategy for the representation of groups of individuals by using a symbolic data approach which is further development of what has been proposed in the original paper by Meulman. The optimal scaling CATPCA technique has been widely illustrated in (Meulman et al., 2004) and it is implemented in the main statistical software (like SPSS, SAS). It is performed on three groups of variables related to the feeling of national identity of the respondents and their opinion about immigration.

Starting from the CATPCA results, our main contribute consists in quantifying the profiles of all the individuals (respondents belonging to different countries) with respect to the quantification vectors and then, to represent the empirical distributions, for each country, of their values on the first two factorial axes. We consider the empirical distributions, as histogram-valued data, according to the symbolic data definition (Bock and Diday, 2000). In this way, we take into consideration not only the mean-points of the observations in each country, as in CATPCA analysis, but the quantified-values distributions. In the context of Symbolic Data Analysis (SDA), many techniques have been developed for analyzing distributional, or histogram-valued, data, like clustering methods (Verde and Irpino, 2008), regression model (Billard and Diday, 2006; Dias and Brito, 2011; Irpino and Verde, 2013a), forecasting model (Arroyo, 2009; Arroyo, Maté, 2008) and factorial analysis (Verde and Irpino, 2013b). We perform a Dynamic Clustering Algorithm (DCA) of histogram valued data, according to the method proposed by Irpino et al. (2006). Each symbolic data corresponds to a country described by two empirical distributions of the individual scores on the first and second factorial variables. Being obtained by an optimal scaling procedure based on PCA, the two factorial variables are assumed to be uncorrelated by definition. The results of the clustering algorithm allow us to identify similar behavior of the citizen in different countries about their feeling on immigration and their national identity sentiment. The main advantage of using distributions to represent countries, rather than syntheses of the quantified opinions such as mean-points, is the possibility to take into consideration the variability of the distributions of the individual scores on the reduced space of the quantified variables, which highlights the diversity of the opinions inside the same country. Moreover, the DCA provides a description of each cluster by means of prototypal distributions (for each variable) which synthesize the characteristics of the distributions of the countries belonging to the cluster. As illustrated in (Irpino et al., 2006; Verde and Irpino, 2007), the DCA is based on a suitable distance between distributions, the Wasserstein metric (Wasserstein, 1969), which has been also proposed in other analysis contexts involving histogram-valued data (Dias and Brito, 2011; Arroyo, 2009). The ℓ_2 norm of the Wasserstein distance, also known as Mallows distance (Mallows, 1972) can be interpreted as an Euclidean distance between quantile functions, the inverse of the empirical cumulated distribution functions associated to histogram-valued data. This metric presents many properties, as demonstrated in Arroyo (2009), in Irpino et al. (2006) and in Irpino and Romano, (2007), so it is preferred to some other metric or dissimilarity measures between distributions. Finally, the representation of the results on the factorial plane gives an interpretation of the clusters according to the new quantified variables. Then, we point out the main differences in the results of our approach with respect to the one based on the mean-points of the countries (see Meulman, 2004). The paper is structured as follows: the first section recalls the Optimal scaling procedure (CATPCA) for a quantification of the ordinal variables by means of a monotone regression as well as by a non linear transformation using spline functions; Section 2 introduces the histogram-valued data, according to the symbolic data definition. In this section it is also discussed the way to construct the input data as the empirical distributions of the individual scores on the reduced subspace of the quantified variables; the third section illustrates the dynamic clustering algorithm for histogram-valued data and the suitable ℓ_2 Wasserstein distance introduced for comparing distributional data; the last section presents the proposed strategy by using a large-scale multivariate data set from the ISSP (2003) concerning feelings of national identity. Then, a comparison with the results of a classical CATPCA approach is performed. The conclusions open new perspectives on the clustering techniques for non-numeric ordinal data.

2. Optimal Scaling Transformation: Categorical Principal Component Analysis

To cope with the problems related to the uncertainty of the measurements recorded on categorical or ordinal scale variables (already mentioned in the introduction), we proceed to the quantification of variables by means of an optimal scaling approach. When data are expressed by categorical or nominal levels, a suitable nonlinear transformation

method is the nonlinear Principal Component Analysis. It converts every category to a numeric value, in accordance with the level of analysis chosen for the variables, using optimal quantification. A first version of this method was described by Guttman (1941), other major contributions in the literature are from Kruskal (1965), Shepard (1966), Kruskal and Shepard (1974), Young et al. (1978), and Winsberg and Ramsay (1983) (for an overview, see Gifi, 1990). The Categorical Principal Component Analysis (CATPCA) based on nonlinear transformation was introduced by Meulman and Heiser (1999) and implemented in SPSS statistical software (Meulman et al., 2004). In nonlinear PCA the optimal quantification step and the linear PCA are performed simultaneously through the minimization of a least-squares loss function. The two analyses are alternated through the use of an iterative algorithm that converges to a stationary point in which the optimal quantifications of the categories do not change anymore.
The nonlinear PCA is based on the classical scheme (Gifi, 1990) looking for the direct quantification of the categories of categorical variables and the scores of the individuals.

Let Y_m be a categorical variable, described by C_m categories, where $m = 1, \ldots, M$ is the index of the variable. We denote with $\mathbf{G}_m$ the *indicator matrix* (according to the term used for the first time by de Leeuw in 1968, but also known as *attribute*, *dummy*, *incidence* matrix) with N rows, with N the number of individuals, and C_m columns. The elements of $\mathbf{G}_m$ assume the values 0 or 1 according to the category chosen by the individuals. Assuming that each individual can choose only one category of response for each variable, the $\mathbf{G}_m$ is a complete disjunctive table.
With $\mathbf{y}_m$ is denoted the quantification vector for the C_m categories, such that the transformed variable is: $\mathbf{q}_m = \mathbf{G}_m\mathbf{y}_m$.
The quantified variable $\mathbf{q}_m$ becomes a single vector of dimension N which assigns a numerical value to each individual with respect to categorical variable. Define $\mathbf{x}$ the vector of the quantification of the individuals as the mean vector of all the $\mathbf{q}_m$: $\mathbf{x} = M^{-1}\sum_{m=1}^{M}\mathbf{q}_m$; for same direct quantification $\mathbf{y}_m$ of the categories, $\mathbf{x}$ is the vector of the induced scores of the N individuals.
We still define the induced quantification of a category as the average of the scores of those individuals that are mapped to such category:

$$\mathbf{y}_m = \mathbf{D}_m^{-1}\mathbf{G'}_m\mathbf{x} \tag{1}$$

where $\mathbf{D}_m^{-1}$ is the inverse of the matrix $\mathbf{D}_m = \mathbf{G'}_m\mathbf{G}_m$ of the weights of the categories, given by the frequencies of the categories (the latter assumes that there are no categories with zero frequency). The two procedures require that the solution for direct quantification of the individuals $\mathbf{x}$ must be proportional to the induced individual scores, and vice versa, that the direct quantification $\mathbf{y}_m$ of the categories must be proportional to the induced category quantification $\mathbf{D}_m^{-1}\mathbf{G'}_m\mathbf{x}$. This discussion is related to just one solution of direct quantification of individuals $\mathbf{x}$ and of direct quantification $\mathbf{y}_m$ of the categories of the m-th categorical variable.
The here recalled nonlinear PCA procedure allows us to obtain P different solutions. This implies that the category quantifications are collected in a matrix $\mathbf{Y}_m$ of dimension

($C_m \times P$), and **X** ($N \times P$) is the matrix of the N individual scores on the P dimensional representation space. We define $\mathbf{a}_m$ as the vector of dimension C_m of the coordinates of the categories of the variable Y_m on the same subspace and **A** as the matrix of the the coordinates of the categories of the M variables.

The CATPCA objective function can be written as follows:

$$L(\mathbf{Q}, \mathbf{X}, \mathbf{A}) = M^{-1} \sum_{m=1}^{M} \|\mathbf{q}_m - \mathbf{X}\mathbf{a}_m\|^2 \,; \tag{2}$$

where **Q** is the matrix of the quantified variables $\mathbf{q}_m$. The optimal quantification is given by the C_m values of the vector $\mathbf{y}_m$. The objective function can be also written as:

$$L(\mathbf{y}, \mathbf{X}, \mathbf{A}) = M^{-1} \sum_{m=1}^{M} \|\mathbf{G}_m\mathbf{y}_m - \mathbf{X}\mathbf{a}_m\|^2 \,; \tag{3}$$

under the constrains: $\mathbf{u'G_m y_m = 0}$ and $\mathbf{y'_m D_m y_m = 1}$, having indicated with $\mathbf{u}$ the unitary vector of dimension m and **y** the vector of the $\mathbf{y}_m$ quantifications of the M variables.
An approximation of the nonlinearly scaled variables $\mathbf{q}_m = \mathbf{G}_m\mathbf{y}_m$, in the P-dimensional subspace, is given by the projection of the N object points **X** on the vector $\mathbf{a}_m$.

Being the categorical variables Y_m (for $m = 1, \dots, M$) discrete ordinal variables, the quantification solutions $\mathbf{y}_m$ must belong to a cone in a C_m-dimensional space.
Nonlinear PCA procedures (e.g. CATPCA) find solutions by:

- *weighted monotone regression process*, which makes $\mathbf{y}_m$ monotonically increasing. The weights are the diagonal elements of the matrix $\mathbf{D}_m$;

- alternatively, a *regression monotone spline* is used for a direct quantification of ordinal categories. The spline transformation is computed as a weighted regression (with the same weights as the monotone regression) of $\mathbf{y}_m$ on the I-spline basis $\mathbf{S}_m$: $\mathbf{y}*_m = d_m + \mathbf{S}_m\mathbf{b}_m$. For the spline ordinal scaling level the elements $\mathbf{b}_m$ are restricted to be no negative.

The procedure starts with an estimate of $\mathbf{y}_m$ satisfying the constrains. Then, it computes $\mathbf{G}_m\mathbf{y}_m$ and minimizes $SSQ(\mathbf{Q} - \mathbf{XA})$ on **X** and **A**. Given **X** and **A**, a new $\mathbf{y}_m$ is computed, for each m, by a normalized cone regression.

The main results of CATPCA are represented by the graphical representations: a biplot can display the individuals, as points, and the transformed variables, as vectors. Furthermore, groups of individuals can be highlighted by the joint representation of their centroids (mean-points) and the most characterizing category points of variables.

3. Representation of groups of individuals by Histogram-valued data

In this section, we introduce a way to represent groups of individuals, on the achieved dimensionally reduced subspace, as symbolic data. In SDA, *symbolic data* are individual descriptions of *symbolic objects*. They assume a set of numbers or categories (weighted or not) of set-valued variables, or *symbolic variables*; whereas classical data take only a single value (a number or a category) for each variable. Bock and Diday (2000), in the reference book of SDA, defined *symbolic variables* as follows:

Let E be a set of objects, a variable Z is termed *set-valued* with domain $\mathcal{Z}$, if for all $i \in E$,

$$\begin{aligned} Z : E \to D \\ i \mapsto z(i) \end{aligned} \tag{4}$$

where the description D is defined by $D = \mathcal{D}(\mathcal{Z}) = \{U \neq \emptyset | U \subseteq \mathcal{Z}\}$.

A set-valued variable Z is called *multi-valued* if its description set D_c is the set of all finite subsets of the underlying domain $\mathcal{Z}$; such that $|z(i)| < \infty$, for all $i \in E$.

A set-valued variable Z is called *categorical multi-valued* if it has a finite set $\mathcal{Z}$ of categories and *quantitative multi-valued* if the values $z(i)$ are finite sets of real numbers.

A set-valued variable Z is called *interval-valued* if its description set D_I is the set of intervals of $\Re$.

A *modal variable* Z on a set E of objects with domain $\mathcal{Z}$ is a mapping $z(i) = (S(i), \pi_i), \forall i \in E$, where: π_i a measure (frequency, probability or weight) on the domain $\mathcal{Z}$ of the possible observation values (completed by a σ-field), and $S(i) \subseteq \mathcal{Z}$ is the support of π_i in the domain $\mathcal{Z}$. The description set of a modal variable is denoted with D_m.

The *Histogram valued* is a particular case of modal variable when we assume that:

the support $S(i) = [\underline{z}_i; \overline{z}_i]$ is bounded in $\Re$ and is partitioned into a set of n_i intervals, or bins, I_{li}, as follows:
$S(i) = \{I_{1i}, \ldots, I_{li}, \ldots, I_{n_i i}\}$, where $I_{li} = [\underline{z}_{li}, \overline{z}_{li})$ (for $l = 1, \ldots, n_i$), i.e.:

$$\begin{array}{ll} i. & I_{li} \cap I_{l'i} = \emptyset; \;\; l \neq l' \;\; ; \\ ii. & \bigcup\limits_{l=1,\ldots,n_i} I_{li} = S(i) \end{array}$$

Histogram definition supposes that each interval I_{li} is uniformly dense. In such a way, it is possible to define the description of the object i as follows:
$z(i) = \{(I_{li}, \pi_{li}) \mid \forall I_{li} \in S(i); \; \pi_{li} = \Psi_i(\underline{z}_{li} \leq y \leq \overline{z}_{li}) = \int\limits_{I_{li}} \psi_i(\nu) d\nu \geq 0\}$, where

$\int\limits_{S(i)} \psi_i(\nu)d\nu = 1.$

According to our proposal, we represent groups of individuals by the means of the histogram of the individuals scores on the two first principal axes (stored in the first two columns of $\mathbf{X}$). Thus, we denote by Z_1 and Z_2 two histogram variables and by $z_1(i) = (I_{1li}, \pi_{1li})$ (where $\forall I_{1li} \in S_1(i)$) and $z_2(i) = (I_{2li}, \pi_{2li})$ (where $\forall I_{2li} \in S_2(i)$) the histograms of the i-th group for the first and second principal axis, respectively. Being the factorial axes orthogonal for construction, the two new variables Z_1 and Z_2 are uncorrelated. The S_1 and S_2 are respectively the support of Z_1 and Z_2. Furthermore, we have chosen to partition the support of each histogram in 50 intervals (bins), bounded by the 50 quantiles of the distribution.

4. Dynamic Clustering algorithm for Histogram-valued data

Dynamic Clustering Algorithm (DCA), here proposed for partitioning the set E of histogram data, consists in a generalization to the case of multi-valued data, of the well known method of the "Nuées Dinamiques" introduced by Diday (1971). In its classical formulation, the DCA represents a general reference for partitioning algorithms.

It is based on two alternating steps: a representation and an allocation step. DCA looks for the partition $\mathcal{P} \in \mathcal{P}_K$ of E in K classes among all the possible partitions $\mathcal{P}_K$, and the vector $L \in L_K$ of K representative elements of the classes $C_k \in \mathcal{P}$ such that the fitting criterion Δ between L and $\mathcal{P}$ is minimized:

$$\Delta(\mathcal{P}^*, L^*) = Min\{\Delta(\mathcal{P}, L) \mid \mathcal{P} \in \mathcal{P}_K, L \in L_K\}. \tag{5}$$

Such a criterion is defined as the sum of dissimilarity, or distance measures $\delta(z_i, G_k)$ between each element z_i of $C_k \in \mathcal{P}$ and the class representation $G_k \in L$:

$$\Delta(\mathcal{P}, L) = \sum_{k=1}^{K} \sum_{y_i \in C_k} \delta(y_i, G_k).$$

- *Initialization*. The algorithm starts from a random partition $\mathcal{P}^{(o)}$ of the set E in a fixed number K of clusters and compute the representative element G_k associated with a class C_k as an element of the description space of E. When the data are represented by histograms, G_k, defined as prototype of the cluster (Verde, Irpino, 2007), is still a histogram.

- The *allocation step* consists in allocating the elements z_i to a cluster C_k according to the minimum distance $\delta(z_i, G_k)$ from the prototype G_k:
z_i is assigned to the cluster C_k if $\delta(z_i, G_k) < \delta(z_i, G_{k'})$ for all $k' \neq k$.

- The *representation step* computes the prototypes G_k of the clusters C_k of the new partition $\mathcal{P}^{(1)}$.

The two steps are alternated until the partitions and the prototypes do not change anymore.

The main choice concerns the measure $\delta(.)$ of dissimilarity, or the distance, to compare histogram data. As proposed by the authors in previous papers, a suitable comparison measure is the ℓ_2 Wasserstein distance (also known as Mallows distance (1972)), for the important properties that it presents.

As in the original DCA, the consistence between the distance measure and the prototype function used for representing the clusters, guarantees the convergence of the algorithm to a stationary point.

4.1. Wasserstein metric for histogram data

The ℓ_p Wasserstein distance [12] is a distance function defined between the probability distributions of two random variables X_1 and X_2, on a given metric space $\mathcal{S}$. The minimal $\ell_1 - metric$ was introduced and investigated already in 1940 by Kantorovich for compact metric spaces. In 1914, Gini introduced the same metric in a discrete setting on the real line, Salvemini (1943), in the discrete case, and Dall'Aglio (1956), in the general case, proved the basic representation of ℓ_p norm between the quantile functions of the two random variables X_1 and X_2. Denoting with $F(X_1)$ and $G(X_2)$ the distribution functions of X_1 and X_2 respectively, the ℓ_p Wasserstein distance is expressed as follows:

$$d_W(X_1, X_2) := \left[\int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right|^p dt \right]^{\frac{1}{p}}.$$

Mallows (1972) proposed a metric corresponding to the ℓ_2 version of the Wasserstein metric. In particular, in our analysis we focus our attention on such ℓ_2-norm distance because it can be interpreted as the Euclidean distance between quantile functions.

The main computational drawbacks are related to the invertibility of the distribution functions. But, as we will show in the following, this problem can be addressed, when data are histograms, by introducing an exact and efficient way to compute this distance.

Given a histogram description $z(i)$, with H_i be the number of weighted intervals (bins):

$$z(i) = \left\{ (I_{1i}, \pi_{1i}), ..., (I_{hi}, \pi_{hi}), ..., (I_{H_i i}, \pi_{H_i i}) \right\},$$

we define the quantities w_{li} in order to represent the cumulative weights associated with the elementary intervals of $z(i)$:

$$w_{li} = \begin{cases} 0 & l = 0 \\ \sum\limits_{h=1,\ldots,l} \pi_{hi} & l = 1, \ldots, H_i \end{cases} . \tag{6}$$

Using (6), and assuming a uniform density for each I_h, we can describe the empirical distribution function as:

$$\Psi_i(z) = w_i + (z - \underline{z}_{\,li}) \frac{w_{li} - w_{l-1i}}{\overline{z}_{\,li} - \underline{z}_{li}} \quad \textit{iff } \underline{z}_{\,li} \le z \le \overline{z}_{li}.$$

Then, the inverse distribution function is a piecewise function defined as follows:

$$\Psi_i^{-1}(t) = \underline{z}_{\,li} + \frac{t - w_{l-1i}}{w_{li} - w_{l-1i}} \left(\overline{z}_{li} - \underline{z}_{li}\right) \;\; w_{l-1i} \le t < w_{li}.$$

To compute the distance between two histogram descriptions $z(i)$ and $z(j)$ we need to identify a set of common uniformly dense intervals. Let w be the set of the cumulated weights of the two distributions: $w = [w_0, ..., w_l, ..., w_s]$ where: $w_0 = 0 \;\; w_s = 1$ and $\pi_l = w_l - w_{l-1}$. To solve the problem of a common set w of cumulated weights (relative frequencies) associated to the s-th quantiles of the two distributions, we consider equi-depth histograms. The weights $\pi_l = w_l - w_{l-1}$ associated to the intervals (bins) I_{li} are all equals (for $l = 1, \ldots, s$) to the relative frequency $\frac{1}{s}$.

The squared distance between the two histogram descriptions is computed as:

$$d_M^2(z(i), z(j)) := \sum_{l=1}^{s} \int_{w_{l-1}}^{w_l} \left(\Psi_i^{-1}(t) - \Psi_j^{-1}(t)\right)^2 dt. \tag{7}$$

Each couple (w_{l-1}, w_l) allows us to identify two uniformly dense intervals, one for i and one for j, having respectively the following bounds:

$$I_{li} = [\underline{z}_{li}, \overline{z}_{li}] \quad \text{and} \quad I_{lj} = [\underline{z}_{lj}, \overline{z}_{lj}].$$

Recalling that $\Psi_u^{-1}(w_{l-1}) = \underline{z}_{lu}$; $\Psi_u^{-1}(w_l) = \overline{z}_{lu}$ (for $u = i, j$), for each interval it is possible to compute the centers and the radii, as follows:

$$c_{lu} = (\overline{z}_{lu} + \underline{z}_{lu})/2 \quad r_{lu} = (\overline{z}_{lu} - \underline{z}_{lu})/2.$$

Because intervals are uniformly distributed, we may express them as function of their centers and radii and rewrite equation (7) as:

$$d_W^2(z(i), z(j)) = \sum_{l=1}^{s} \pi_l \left[(c_{li} - c_{lj})^2 + \frac{1}{3} (r_{li} - r_{lj})^2 \right]. \tag{8}$$

The proposed distance can be easily extended to the case of P variables. Under the hypotheses of independence of the P variables, the multivariate version of $d_W^2(z(i), z(j))$ is additive on the Z_p, with $p = 1, \ldots, P$ variables.

4.2. Using ℓ_2 Wasserstein distance in dynamic clustering algorithm

Given a set of n histogram data, we define the *centroid* of the elements in E as a histogram itself (or prototype of the global set E). According to the criterion function minimized in DCA, the prototypal histogram $Z(b)$ can be computed minimizing the following (sum of distances) function:

$$f(c_{1b}, r_{1b}, \ldots, c_{sb}, r_{sb}) = \sum_{i=1}^{n} \sum_{j=1}^{s} \pi_j \left[(c_{ji} - c_{jb})^2 + \frac{1}{3} (r_{ji} - r_{jb})^2 \right]. \tag{9}$$

Once fixed s (and so also π_j) to be equal to the cardinality of the elementary intervals of the union of the supports of the $z(i)$'s, the support of $z(b)$ can be expressed as a vector of s pairs (c_{jb}, r_{jb}). Function in (9) holds a minimum when the following first order conditions are satisfied:

$$\begin{cases} \frac{\partial f}{\partial c_{jb}} = -2\pi_j \sum\limits_{i=1}^{n} [(c_{ji} - c_{jb})] = 0 \\ \frac{\partial f}{\partial r_{jb}} = -\frac{2}{3}\pi_j \sum\limits_{i=1}^{n} [(r_{ji} - r_{jb})] = 0 \end{cases}$$

for each $j = 1, \ldots, s$. It is easy to prove that function (9) is minimum when:

$$c_{jb} = n^{-1} \sum_{i=1}^{n} c_{ji} \quad ; \quad r_{jb} = n^{-1} \sum_{i=1}^{n} r_{ji}.$$

The barycenter (*prototype*) of the n histogram data is expressed as follows:

$$Z(b) = \{([c_{1b} - r_{1b}; c_{1b} + r_{1b}], \pi_1); \ldots; \ldots; ([c_{sb} - r_{sb}; c_{sb} + r_{sb}], \pi_s)\}. \tag{10}$$

The identification of a barycenter allows us to show a second property of the proposed distance: it is possible to express a measure of inertia of data using d_W^2. The total inertia, with respect to the barycenter $z(b)$ of a set of n histogram data, is given by the following quantity:

$$TI = \sum_{i=1}^{n} d_W^2 z(i), z(b).$$

Furthermore, according to the Huygens', Total inertia can be decomposed into the sum of Within- and Between-clusters inertia. Let us consider a partition of E into K clusters. For each cluster C_k, with $k = 1, .., K$, a histogram barycenter, denoted by $z(b_k)$, is computed by a local optimization of (9). Minimizing the following function:

$$f(c_{1b_1}, r_{1b_1}, \ldots, c_{sb_K}, r_{sb_K}) = n^{-1} \sum_{k=1}^{K} |C_k| \sum_{j=1}^{m} \pi_j \left[(c_{jb_k} - c_{jb})^2 + \frac{1}{3} (r_{jb_k} - r_{jb})^2 \right] \tag{11}$$

where $|C_k|$ is the cardinality of cluster C_k. It is possible to prove that functions (9) and (11) reach the a minimum for the same $z(b)$. The last result permits to obtain the following decomposition of the total inertia[1]:

$$TI = WI + BI = \sum_{k=1}^{K} \sum_{i \in C_k} d_M^2(z(i), z(b_k)) + \sum_{k=1}^{K} |C_k| d_M^2(z(b_k), z(b)). \quad (12)$$

5. Application on the ISSP 2003 dataset: main results

An example of Principal Component Analysis with nonlinear optimal scaling transformations (CATPCA) is presented by Meulman et al (2004). The authors use a large-scale multivariate data set from International Social Survey Programme (ISSP, 1995) concerning the feeling of national identity from about 28,500 respondents in 23 different countries.

Following that example, we have performed a similar analysis on the dataset of a successive edition of the same international survey: the ISSP 2003[2]. We have considered only three groups of variables (almost the same of those analyzed by Meulman et al., 2004). The first group of four variables (section Q.2 of the questionnaire) indicates how close the respondents feel toward their town or city (a), their county (b), their country (c), and their continent (d). Answers are recoded by a score ranging from 1 (*not close at all*) to 4 (*very close*). The second group of eight variables (section Q.3 of the questionnaire) is related to how the respondents consider important the following things about their national identity: a. *to have born in [Country]*; b. *to have citizenship*; c. *to have there lived for most of one's life*; d. *to be able to speak the country language*; e. *to be a [of the dominant religion of the country]*; f. *to respect [Country] national political institutions and laws*; g. *to feel [Country nationality]*; h. *to have [Country nationality] ancestry*. Data are recoded by a score ranging from 1 (*not important at all*) to 4 (*very important*). The third set of variables (section Q.10 of the questionnaire) concerns opinions about immigrants, asking the respondents' agreement, on a scale from 1 (*strongly disagree*) to 5 (*strongly agree*), about the following statements: a. *Immigrants increase crime rates*, b. *Immigrants are generally good for the economy*, c. *Immigrants take jobs away from people who were born [in this country]*, d.*Immigrants make [this] Country more open to new ideas and cultures*; e. *Government spends too much money assisting the immigrants*. Finally, respondents were asked to scale themselves with respect to the statement (section Q.11 of the questionnaire): *The number of immigrants to [my Country] nowadays should be: reduced a lot* (1) ... *increased a lot* (5).

[1] For the sake of brevity, we do not report here the proof.

[2] The ISSP is a continuous annual cross-national data collection project that has been running since 1985. More detailed information about the ISSP data service are available on the ISSP Internet pages provided by the Zentral Archiv, Cologne: "http://www.tarki.hu/en/research/issp/". We refer to the Questionnaire ISSP - 2003 National Identity (II)

The number of effective respondents to this survey was 45,993 in 36 countries. However, since the number of respondents in each country was proportional to the population size, a random resampling from the original data has been performed in order to give the same weight to all the countries in the analysis. The analysis has been conducted on 18,000 individuals corresponding to 500 individuals for each one of the 36 countries. The list of the analysed variables is presented in tab. 1, while the list of the respective categories is in tab. 2.

Var-id	Label	Variable
Q2a	Close-City	How close do you feel to: Your town - city
Q2b	Close-Province	How close do you feel to: Your [county], province
Q2c	Close-Country	How close do you feel to: [Country]
Q2d	Close-Continent	How close do you feel to: continent
Q3a	Imp-born	Important: to have been born in [Country]
Q3b	Imp-nationality	Important: To have [Country Nationality] citizenship
Q3c	Imp-live	Important: To have lived in [Country] for most of one's life
Q3d	Imp-language	Important: To be able to speak [Country language]
Q3e	Imp-religion	Important: To be a [religion]
Q3f	Imp-politics	Important: To respect [Country Nationality] political institutions and laws
Q3g	Imp-identity	Important: To feel [Country Nationality]
Q3h	Imp-ancestry	Important: To have [Country Nationality] ancestry
Q10a	I-Crime	Immigrants increase crime rates
Q10b	I-Economy	Immigrants are generally good for [Country's] economy
Q10c	I-Job	Immigrants take jobs away from people who were born in
Q10d	I-Culture	Immigrants improve [Country Nationality] society by bringing in new ideas and cultures
Q10e	I-Assisting	Government spends too much money assisting immigrants
Q11	N-Immigrants	Number of immigrants coming to Country

Table 1. List of the variables of analysis (Var-id "Variable identifier"; Label"short name of the variable"; Variable "Name of the variable")

The optimal scaling transformation of the 18 categorical ordinal variables is performed by CATPCA, using SPSS Software. The CATPCA, as illustrated above, achieves a quantification of the ordinal variables by a monotone regression or by a spline regression under the constrains of positive coefficients in order to obtain solutions in a cone-space. Then, the CATPCA performs a reduction of the space of the transformed variables by extracting the factorial axes maximally correlated with the quantified variables. Different graphical representations can be shown in order to analyze the structure of correlation between the variables in the analysis, the position of the categories of the quantified variable along the straight lines trough the origin which represent the variables on the factorial plane. The number of objects in the ISSP data set is too large to inspect the relations between object and variables or categories, so the individual points are represented by the centroids of the respondents of each country.

The fig. 1 shows the plot of the variables on the first factorial plane (axis 1 and 2) of the CATPCA. The first axis (that explains the 26.39% of the total inertia of the

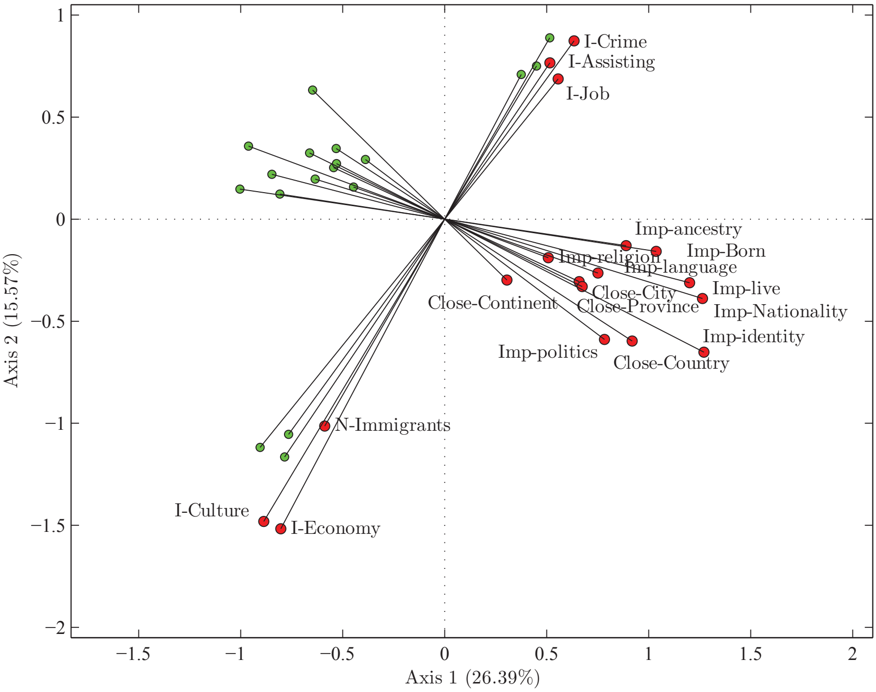


Figure 1. Quantified variables representation on the first factorial plane of CATPCA. (Labels of variables are set in correspondence of the lowest values of the quantified categories, corresponding to the original categories "Not very close", "Not very important", "Disagree Strongly" and "Reduced a lot")

Id-variables	Categories
Q2a Q2b Q2c Q2d	(1) Very close; (2) Close; (3) Not very close; (4) Not close at all
Q3a Q3b Q3c Q3d Q3e Q3f Q3g Q3h	(1) Very important; (2) Fairly important; (3) Not very important; (4) Not important at all
Q10a Q10b Q10c Q10d Q10e	(1) Agree strongly; (2) Agree; (3) Neither agree nor disagree; (4) Disagree; (5) Disagree strongly
Q11	(1) Increase a lot; (2) Increase a little; (3) Remain the same; (4) Reduced a little; (5) Reduced a lot

Table 2. List of the categories of the variables (Id-var. "Identifier of the variables"; Categories "Categories of the variables")

quantified variables) is strongly correlated to the quantified statements related to the National Identity (Q2-how close the respondents feel toward their city, province, country, continent; Q3 - how the respondents consider important to be born in their country, to have citizenship, to have lived in their country for most of one's life, to be able to speak the country language, to be of a prevalent religion of the country, to respect the political institution and the laws, to feel their country nationality, to have ancestry of the same country nationality): the negative versus of the axis identifies strong *nationalist* positions expressed by the quantified categories *Very close* and *Very important* while the positive side of the axis corresponds to less *nationalist* behaviors, as reveal the position on the plane of the quantified categories *Not very close* and *Not very important*. The second axis (that explains the 15.57% of the total inertia of the quantified variables) is then correlated to the quantified variables expressing the tolerance of the respondents about the immigrates (Q10 - Immigrants: increase crime rates, are generally good for country's economy, take jobs away from people who were born in country, improve [Country Nationality] society by bringing in new ideas and cultures; Government spends too much money assisting immigrants; Q11 - Judgements about the Number of immigrants coming to Country). It opposes positions more favorable to immigration (positive side of the axis) versus positions much more critical toward the increasing of the immigration in the country of respondent.

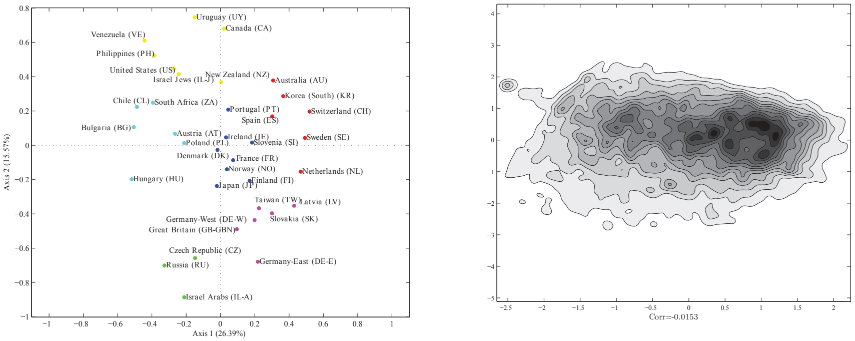


Figure 2. CATPCA: (Right panel) Centroid of countries on the first factorial plane. (Left panel) contour plot and correlation coefficient of the bivariate kernel density estimates of the distribution of points projected on the first factorial plane.

5.1. Clustering quantified ordinal distributions

In Meulman and al. (2004), using *Country* variable as supplementary one, a representation of the centroids on the factorial plane is proposed. A straightforward way of clustering countries could be a k-means of the Country-related centroids. However, the use of centroids instead of the whole Country-related data implies that, for each country, the distribution of points of a particular country is symmetric around its centroid and it is not significatively different in scale and shape with respect to the other country distributions. The Country-related centroids on the first factorial plane are plotted on the left panel of Fig.2, while on the right panel it is represented the contour plot of a two dimensional kernel density estimate (using a Gaussian kernel) of the distribution of the individual points. Looking at the left panel, as it is expected, it is not easy to identify Country-related clusters and also similarities among country distributions.

If we look at the contour plot of each countries, as reported in Figs. 3, 4 and 5 (we reported also the correlation index of each bivariate density), we can observe that: the distributions appear to show a low correlation within each country, except for South Africa, the distributions are skewed. In this case, using the centroids as representatives of each country could be a too restrictive choice, because they do not sufficiently synthesize the information of the country distribution. In such a case, we propose to use marginal distributions for each country and cluster them according to the Wasserstein distance between distributions. In this case, we use a histogram representation for the two marginal distributions associated with each country according to the CATPCA scores on the first and second factorial axis.

Starting from the individual scores on the first and second factorial axis, each country is represented as an equi-depth histogram of 50 bins. For each histogram we have

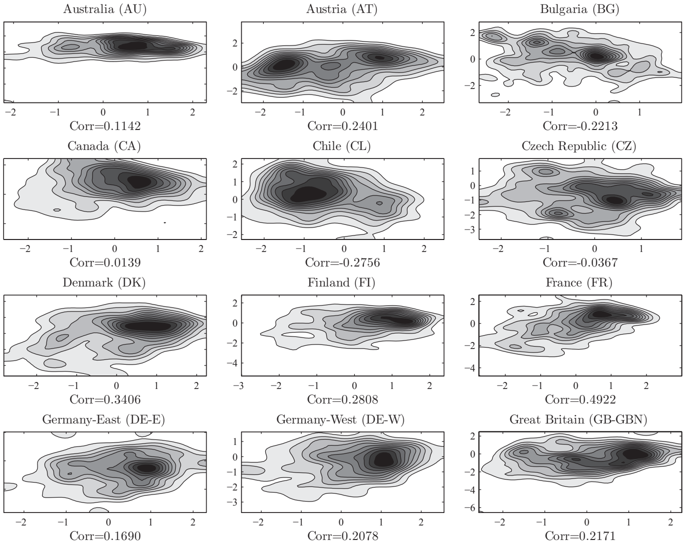


Figure 3. CATPCA: kernel density estimates for each country on the first factorial plane (first group of 12 countries alphabetically ordered).

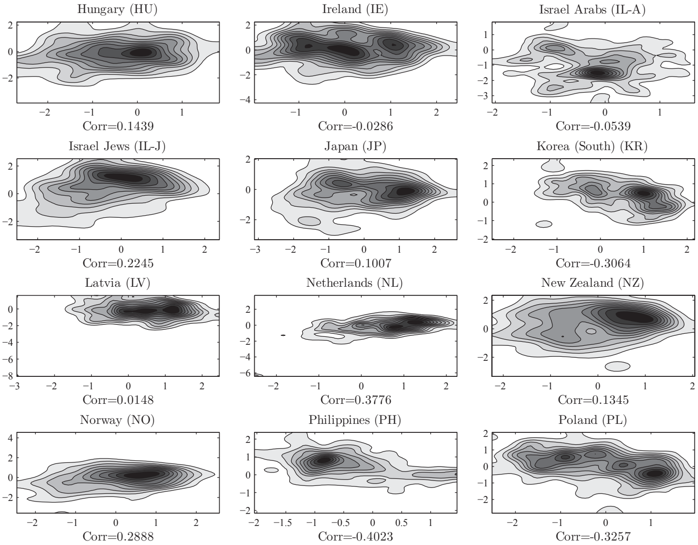


Figure 4. CATPCA: kernel density estimates for each country on the first factorial plane (second group of 12 countries alphabetically ordered).

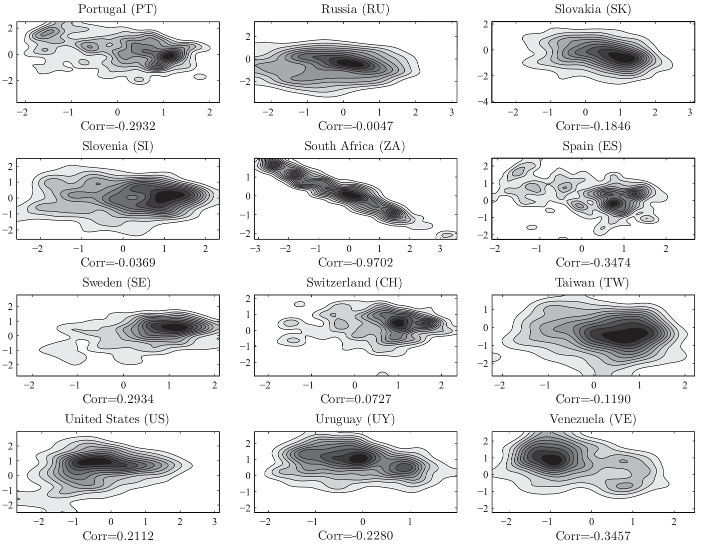


Figure 5. CATPCA: kernel density estimates for each country on the first factorial plane (last group of 12 countries alphabetically ordered).

computed the distribution function and the corresponding quantile function. Then, we perform on these data, a k-means (on the centroids) and a dynamic clustering algorithm (using distributions) in order to compare the results of the two partitioning approaches of the set of aggregated data and to point out the main advantage of considering the distributions rather than the simple centroids of the set of respondents belonging to the different countries. Both in k-means and DCA, it is needed to fix in advance the number of the clusters partitioning the set of data. We chose to fix the number of clusters to 6 on the basis of the Calinski Harabasz index (1974) computed on the k-means of the centroids, and varying the number of clusters from 2 to 8 (that we consider as a reasonable limit for clustering 36 objects). Main results are reported in table 3.

The partition obtained by k-means algorithm is presented in tab. 4 and the optimal value of the optimized criterion is 1.1273 (in tab. 3) corresponding to the Within Inertia. A measure of the partition quality, expressed as $QPI = 1 - W/T$, is equal to 0.87. The quality partition index QPI can be considered as the generalization of the ratio between the Between-inertia and the Total-inertia.

Centroids of the countries are projected on the factorial plane as shown in fig. 6. Countries belonging to the same cluster are connected. According to the correlation between the first axis and the quantified variables which has permitted to give an interpretation of the new factorial variables, it is possible to describe Cluster 1 as that constituted by the *most nationalist countries* (like: Hungary, Bulgaria, Chile, South

No. of clusters	Explained inertia	Within (W)	Between (T)	Total (T)	CH-index
2	0.4551	4.9545	4.1384	9.0929	28.3900
3	0.6614	3.0784	6.0145	9.0929	32.2300
4	0.7506	2.2675	6.8253	9.0929	32.1000
5	0.8249	1.5914	7.5015	9.0929	36.5300
6	0.8760	1.1273	7.9656	9.0929	**42.3900**
7	0.8958	0.9472	8.1457	9.0929	41.5600
8	0.9132	0.7886	8.3042	9.0900	42.1100

Table 3. Calinski Harabasz index for determining the number of clusters in k-means algorithm

Cluster 1	Cluster 2	Cluster 3
Austria (AT) Bulgaria (BG) Chile (CL) Hungary (HU) Poland (PL) South Africa (ZA)	Canada (CA) Israel Jews (IL-J) Philippines (PH) United States (US) Uruguay (UY) Venezuela (VE)	Czech Republic (CZ) Israel Arabs (IL-A) Russia (RU)
Cluster 4	Cluster 5	Cluster 6
Denmark (DK) Finland (FI) France (FR) Ireland (IE) Japan (JP) Norway (NO) Slovenia (SI)	Australia (AU) Korea (South) (KR) Netherlands (NL) Spain (ES) Sweden (SE) Switzerland (CH)	Germany-East (DE-E) Germany-West (DE-W) Great Britain (GB-GBN) Latvia (LV) New Zealand (NZ) Portugal (PT) Slovakia (SK) Taiwan (TW)

Table 4. K-means partition of the countries in K=6 clusters

Africa and Austria and Poland) on the basis of the average of the opinions given by their respondents; a similar interpretation can be given for Cluster 2, especially for the position of some countries (like: Venezuela, Philippines, United States, Israel Jews) - less, Uruguay and in a middle position Canada and New Zealand. Countries belonging to Cluster 3 (like: Russia, Israel Arabs and Czech Republic) also can be interpreted as closer to nationalist positions than the more open ones. Opposed along the first factorial axis, we find Cluster 5 of the countries the *less nationalist* (like: The Netherlands, Sweden, Switzerland, South-Korea, Australia and Spain) and also Cluster 6, especially for some countries (like: Latvia, Taiwan, Germany West and East) and less for Great Britain which is in an intermediate position. Finally, Cluster 4 is that of the countries which assume (in average) a middle position with respect to their feeling about national identity.

Clusters 2, 3, and 6 can be better interpreted according to the meaning given to the second axis: *tolerance of the respondents (in average) about the immigrates*. Countries in Cluster 2 (like: Uruguay, Canada, Venezuela, Philippines, United States, Israel Jews) are much more favorable to immigration in opposition to the those in Cluster 3 (like: Russia, Israel Arabs and Czech Republic) and ine Cluster 6 (like: Germany West and East, Great Britain, Latvia, Taiwan).

In order to take into account the distributions of the opinions expressed by the respondents of each country in the clustering procedure, a dynamic clustering on histogram-valued data is performed by considering a number of clusters equal to six. The partition obtained by DCA is shown in tab. 5. The inertia explained by the histogram-valued prototypes of the dynamic clustering is equal to 75.71% of the total inertia, that amounts to 10.8163. According to the formula 12, it is due to the variability of the centers and the variability of the radii of the bins of the histograms. In such a way, it takes into accounts the internal variability of the empirical distributions and not only of the centroids (like in k-means on the centroids). It is worth noting that the total inertia is computed by using the ℓ_2 Wasserstein distance based on the squared differences between the quantile functions of the country respondents (associated with the histogram representations) and the barycenter of all the quantile distributions. That is different from considering the total inertia computed as squared differences between the means of the countries and the global means (of the all the means of the countries).

In order to show the different structure of the clusters obtained by the k-means algorithm, in fig. 7 are represented the means of the clustered countries connected to the respective barycenters. For facilitating the interpretation, we have not represented the histograms in this graph.

Barycenters of clusters, given by the histogram-valued prototypes, are represented, for each dimension (axis 1 and axis 2), in fig. 8.

Moreover, in fig. 9 the prototypes are represented on the factorial plane in false colors. The different intensity (from dark to light) of the colors is related to the density of the distributions (from high-density = dark to low-density=light).

It is worth to note that the prototypes give a synthesis of the empirical distributions in

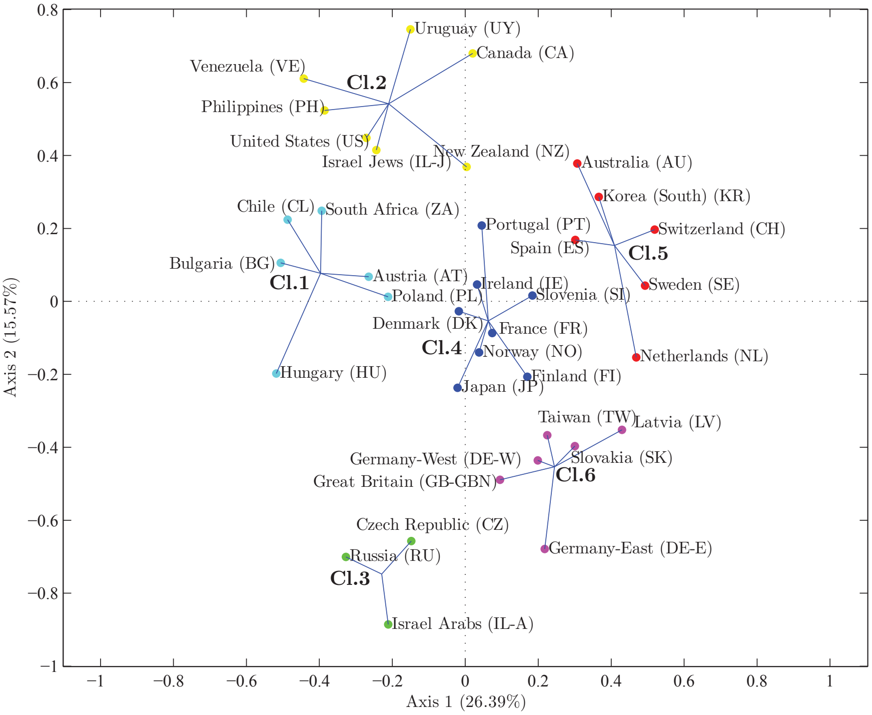


Figure 6. Country means representation on the factorial plane (axis 1 and 2) by CAT-PCA. Countries are connected according to their belonging to the clusters obtained by k-means algorithm on the means of the country - with K=6.

Cluster 1	Cluster 2	Cluster 3
Austria (AT)	Canada (CA)	Czech Republic (CZ)
Bulgaria (BG)	Israel Jews (IL-J)	Germany-East (DE-E)
Chile (CL)	Philippines (PH)	Germany-West (DE-W)
Hungary (HU)	United States (US)	Great Britain (GB-GBN)
Poland (PL)	Uruguay (UY)	Israel Arabs (IL-A)
South Africa (ZA)	Venezuela (VE)	Russia (RU)
		Slovakia (SK)
		Taiwan (TW)
Cluster 4	Cluster 5	Cluster 6
Denmark (DK)	Australia (AU)	Latvia (LV)
Finland (FI)	Korea (South) (KR)	Netherlands (NL)
France (FR)	New Zealand (NZ)	Slovenia (SI)
Ireland (IE)	Portugal (PT)	Spain (ES)
Japan (JP)		Sweden (SE)
Norway (NO)		Switzerland (CH)

Table 5. Dynamic clustering partition of the countries in K=6 clusters

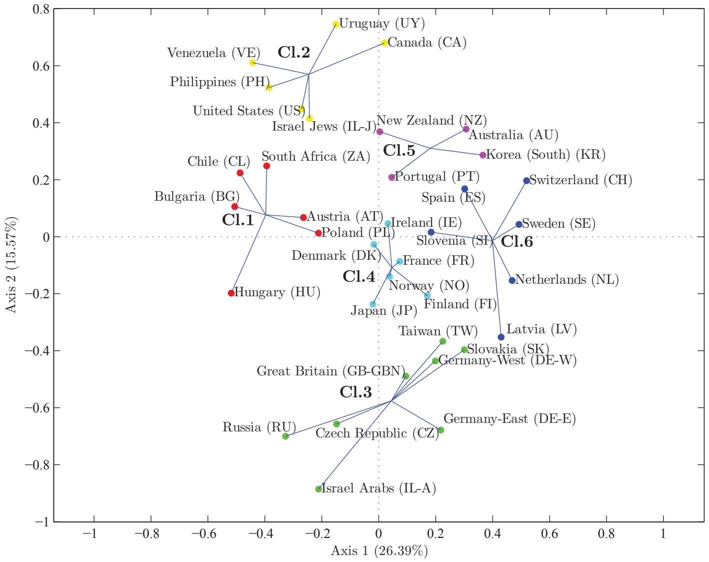


Figure 7. Country means representation on the factorial plane (axis 1 and 2) by CAT-PCA. Countries are connected according to their belonging to the clusters obtained by DCA on the distributions of the country - with K=6.

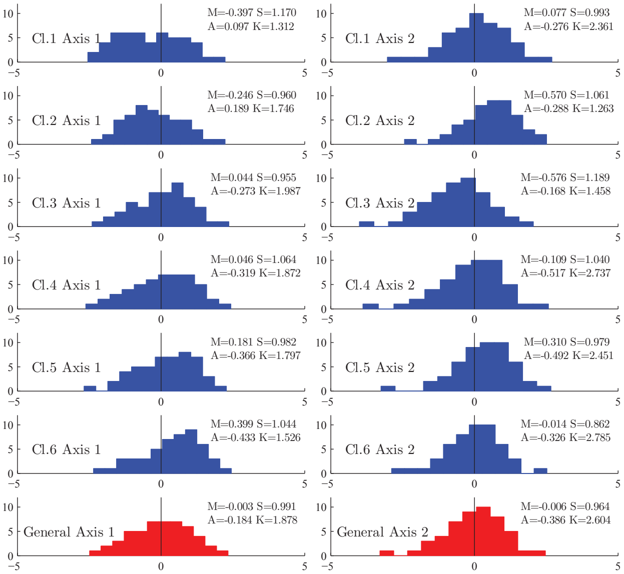


Figure 8. Histogram prototypes of the K=6 clusters by DCA (axis 1 and 2)

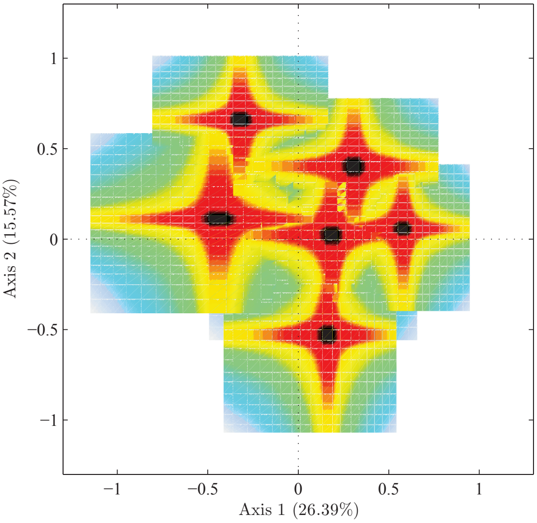


Figure 9. Histogram-valued prototypes of the K=6 clusters by DCA (first factorial plane) in a scale of color according to the density of the distributions

the clusters in terms of *average* histogram. They are also the elements, associated to the quantile distributions, which minimize the sum of the squares ℓ_2 Wasserstein distance, that is the loss function in the analysis. A different representation of the barycenters of the clusters would not be consistent with the criterion of the algorithm and, then, it would not guarantee the convergence of the criterion value to the a stationary point.

We observe that the final partition obtained by the DCA is different from the partition achieved by k-means algorithm on the means of the countries. It is evident an increasing of the internal variability of the clusters due to have taken into consideration not only the variability of the averages of the countries respondents, but the distributions of the counties. The main differences are especially for Cluster 6 along the fist axis. Now, it includes Latvia and Slovenia that, in the k-means representation, were in different clusters. Its interpretation is of a more strong cluster of no-nationalist countries. Cluster 5 of DCA, including two elements of Cluster 5 obtained by k-means and Portugal and New Zealand, it is placed in a more soft position with respect to national identity feeling. Opposed along the first axis, we find Cluster 1 of the countries where the respondents have expressed a strong national identity feeling. These countries are the same of Cluster 1 obtained by k-means algorithm. Cluster 3 obtained by DCA merges two clusters (Cluster 3 and Cluster 6) of the k-means partition and it defines better a cluster of countries the most unfavorable to the immigration, while Cluster 2 joins those countries where a positive feelings towards immigrants is prevalent (it is constituted by countries in Cluster 2 of the k-means partition, except for New Zealand that now migrates in Cluster 5 of DCA partition). Finally, Cluster 1 is characterized by those countries where the opinions about the national identity and the immigrations are mainly expressed by the intermediate categories of the scales of the answers (corresponding to moderate positions).

6. Conclusion

The main advantage of our approach versus the k-means on the centroids is to take into account the distributions of the respondents with respect to the quantified categorical ordinal variables; this allows to keep the variability and the skewness of them in the synthesis of the data (when they are grouped in subclasses). In our opinion it is important, especially in social researches, to do not ”compress” all the information through the ”means” of the responses but to enrich the data in the analysis by considering a more suitable synthesis of them, for instance, by histograms.

The introduction of the Wasserstein metric to compare distributions in the Multivariate Data Analysis of Complex data (including symbolic data) has open a new field of research with a recent development of new techniques (e.g. clustering, regression, forecasting model, factorial analysis). In perspective, we think to investigate the improvement of our strategy by introducing as space-dimensional reduction technique (the PCA step) in optimal scaling quantification, a factorial method which works directly on the distributions for each quantified variable of groups of data (like for the countries in the example). Other main challenges to face will concern the interpretation of the graphical representations of distributions on factorial planes.

References

Arroyo, J., Mate, C. (2009) Forecasting histogram time series with k-nearest neighbours methods, *International Journal of Forecasting* **25** (1), 192-207.

Billard, L., Diday, E. (2006) *Symbolic Data Analysis. Conceptual Statistics and Data Mining*. Wiley, Chichester.

Bock, H. H., Diday, E. (2000) *Analysis of Symbolic Data*, Springer-Verlag, Heidelberg.

Calinski T., Harabasz. J. (1974) A dendrite method for cluster analysis. *Communications in Statistics*, **3** (1), 1–27.

Dall'Aglio, G. (1956) Sugli estremi dei momenti delle funzioni di ripartizione doppia, *Ann. Scuola Normale Superiore di Pisa*, Cl. Sci **3** (1), 3374.

De Leeuw, J. (1990). Multivariate analysis with optimal scaling. In S. Das Gupta & J. Sethuraman (Eds.), *Progress in multivariate analysis*. Calcutta: Indian Statistical Institute

Dias, S., Brito, P. (2011) *A new linear regression model for histogram-valued variables*, in: 58th ISI World Statistics Congress, Dublin, Ireland,
URL http://isi2011.congressplanner.eu/pdfs/950662.pdf.

Diday, E. (1971). La méthode des Nueés dynamiques. *Revue de Statistique Applique*. **19** (2), 19-34

Diday, E., Noirhomme, M., (2008) *Symbolic Data Analysis and the SODAS Software*, Wiley.

Gifi, A. (1981) Nonlinear multivariate analysis. DSWO-Press, University of Leiden/NL, 452 S.

Gifi, A. (1990) Nonlinear multivariate analysis. New York: John Wiley & Sons, 579, ISBN 0-471-92620-5

Gini, C. (1914) *Di una misura della dissomiglianza tra due gruppi di quantit e delle sue applicazioni allo studio delle relazioni stratistiche,* Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti, Tomo LXXIV parte seconda.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst et al. (Eds.), The prediction of personal adjustment, New York: Social Science Research Council, 319348.

Hayashi, C. (1952) On the Prediction of Phenomena from Qualitative Data and the Quantification of Qualitative Data from the Mathematic-Statistical Point of View, in *Annals of the Institute of Statistical Mathematics*, **2**, 93–96.

Heiser, W.J., Meulman, J.J. (1994). Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships. In: M. Greenacre, J. Blasius (Eds.), *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*, New York: Academic Press, 179–209.

Irpino, A., Romano, E. (2007) Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation, in: M. Noirhomme-Fraiture, G. Venturini (Eds.), EGC, vol. RNTI-E-9 of *Revue des Nouvelles Technologies de lInformation*, Cépadués-Éditions, 99-110.

Irpino, A., Verde, R., Lechevallier, Y. (2006) *Dynamic clustering of histograms using Wasserstein metric*, in: COMPSTAT, 869876.

Irpino A., Verde, R. (2008) Dynamic clustering of interval data using a Wasserstein-based distance. *Pattern Recognition Letters*, **29**, 1648- 1658

Irpino, A., Verde, R. (2013) *Dimension reduction techniques for distributional symbolic data*. In: SIS 2013 Advances in Latent Variables - Methods, Models and Applications. June, 2013, Brescia, Italy.

*h*ttp://meetings.sis-statistica.org/in-dex.php/sis2013/ALV/pa-per/viewFile/2586/443

Kantorovich, L. (1940) On one effective method of solving certain classes of extremal problems, *Dokl. Akad. Nauk*, **28** 212215.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society, Series B*, **27**, 251-263.

Kruskal, J. B., Shepard, R. N. (1974). A nonmetric variety of linear factor analysis. *Psychometrika*, **39**, 123-157.

Mallows, C. L. (1972) A Note on Asymptotic Joint Normality, *The Annals of Mathematical Statistics*, **43** (2), 508515.

Meulman, J.J., Heiser, W.J. & SPSS (1999) SPSS Categories 10.0 Chicago: SPSS.

Meulman, J.J. (2003). Prediction and Classification in Nonlinear Data Analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, **68**, 493-517.

Meulman, J.J., Van der Kooij,. A.J., Heiser, W.J. (2004). Principal Components Analysis with Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data. In: D. Kaplan (ed.), *Handbook of Quantitative Methods in the Social Sciences*, Newbury Park, CA: Sage Publications, 49–70.

Salvemini, T. (1943) Sul calcolo degli indici di concordanza tra due caratteri quantitativi, in: *Atti della VI Riunione della Soc. Ital. di Statistica*, Roma.

Shepard, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology*, **3**, 287315.

Verde, R., Irpino, A. (2007) Dynamic Clustering of Histogram Data: Using the Right Metric, in: P. Brito, G. Cucumel, P. Bertrand, F. Carvalho (Eds.), *Selected Contributions in Data Analysis and Classification, Studies in Classification, Data Analysis, and Knowledge Organization*, chap. 12, Springer Berlin Heidelberg, Berlin, Heidelberg, 123-134.

Verde, R., Irpino, A. (2013) Multiple Linear Regression for Histogram Data using Least Squares of Quantile Functions: a Two-components model, *Revue des Nouvelles*

Technologies de l'Information, RNTI-E-**25**, 78–93.

Wasserstein, L. (1969) Markov processes over denumerable products of spaces describing large systems of automata, *Prob. Inf. Transmission* **5**, 47-52.

Winsberg, S., Ramsay, J. O. (1983). Monotone spline transformations for dimension reduction. *Psychometrika*, **48**, 575-595.

Young, F. W., Takane, Y., de Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, **43**, 279281.

Partial possibilistic regression path modeling for subjective measurement

Rosaria Romano
Department of Economics, Statistics and Finance, University of Calabria
E-mail: rosaria.romano@unical.it

Francesco Palumbo
Department of Political Sciences, University of Naples Federico II
E-mail: fpalumbo@unina.it

Summary: This work presents the use of the Partial Possibilistic Path Modeling in the context of subjective measurement, where ordinal data are collected from rating surveys to measure latent concepts. The method combines the principles of PLS path modeling to model the net of relations among the latent concepts, and the principles of possibilistic regression to model the vagueness of the human perception. Possibilistic regression de-fines the relation between variables through possibilistic linear functions and considers the error due to the vagueness of the human perception as reflected in the model via interval-valued parameters. A case study on the the motivational and emotional aspects of teaching is used to illustrate the method.

Keywords: Component-based SEM; Possibilistic regression; Interval valued data; Subjective data.

1. Introduction

Structural equation models (SEMs) include various statistical methodologies that aim to estimate a network of causal relationships among latent variables (LVs) defined by blocks of manifest variables (MVs) (Bollen, 1989; Kaplan 2000). SEMs undertake a multivariate analysis of multi-causal relationships among different independent phenomena. For example, in relational and social studies SEMs allow behavior and performance to be explained and predicted.

The research paradigm of SEMs grounds on psychometric (covariance-based, LISREL) and chemometric research tradition (variance-based, Partial Least Squares (PLS)).

LISREL, an acronym for linear structural relations, is a statistical software package used in structural equation modeling (Jöreskog, 1970); specialized literature refers to LISREL also as a methodology known as 'classical SEM'. This originates from the classical test theory, which entails application of a covariance-based structural equation model (CB-SEM). The PLS approach, also called 'component-based SEM' or 'PLS Path Modeling' (PLS-PM), is considered a soft-modeling approach constructing composite indexes and entails application of a variance-based structural equation model (VBSEM) (Wold, 1975). There is an important difference in theoretical background between the two approaches. CBSEM is considered a confirmatory method, which is guided by theory because it seeks to replicate the existing covariation among manifest variables. VBSEM is considered an exploratory method also based on some theoretical foundations, but its goal is to predict the behavior of relationships among constructs and to explore the underlying theoretical concepts. From a statistical point of view, CBSEM aims to reproduce the existing sample covariance matrix by using a global optimization criterion, whereas VBSEM formulates the causality dependencies between LVs in terms of linear conditional expectations and aims to maximize explained variance by solving separately any regression in the model.

The choice between the CBSEM or VBSEM approach is grounded on some research conditions: (i) conceptual background of the research problem under study; (ii) indicator-construct design; (iii) measurement scale; (iv) sample or population size under study.

PLS-PM as a soft modeling approach has to be preferred in those application fields where the traditional assumptions (related to the distributions, the measurement scale and the sample size) are not tenable (Tenenhaus et al., 2005). This is why PLS-PM is increasingly being used in empirical studies of many socio-economic phenomena. In particularly, PLS-PM is strictly related to subjective measurements, where data derived from surveys are collected to measure concepts like racism, happiness, corruption and customer satisfaction. These latent concepts are generally measured by scores defined over interval scales. Being latent variables, such scores are defined through the sum of (weighted) scores measured on manifest variables. However, the use of ordinal scales is largely preferred and very common to collect subjective measurements (Davino and Romano, 2013). The Likert scale (named after its inventor Rensis Likert), which is often also referred to as the rating scale, actually represents the sum of rating scale measured scores. Likert assumed that the frequency distributions over the measuring scales were symmetric and that the sum of the ratings could be reasonably approximated by a distribution defined over $\Re$. To justify the approximation of the rating scales to an interval scale, in his original proposal, Likert assumed that the number of items referring to a single scale was eight and that the measuring scale had seven ordered ratings. Generally, it frequently happens that answers to a questionnaire are given on ordinal Likert scales, assuming a *unique* common rating measurement scale.

In most research and applied works, PLS-PM and other statistical techniques are conventionally used for handling variables measured on ordinal scales. According to

Likert's original proposal, such a practice is consistent when the number of items is large enough and it is widely accepted. However, many contributions have been recently proposed in the PLS-PM framework aiming to treat the ordinal indicators in their own nature (Betzin and Henseler, 2005; Jakobowicz and Derquenne, 2007; Lauro et al., 2011; Russolillo, 2012).

Similarly to classical least squares regression, in PLS-PM the process of data analysis is represented by the simple equation: *data* = *model* + *error* (Judd and McClelland, 2009). Here *randomness* represents the main source of uncertainty, i.e. random measurement errors in collecting data. However, there are other sources of uncertainty besides randomness (Coppi, 2008). When human judgments are involved in the data analysis process, as in subjective measurement, *vagueness* is the major source of uncertainty (Zadeh, 1973). Some examples, for instance, are represented by sensory analysis where judges are the measurement tools, or consumer analysis where consumers express their preferences, or social science where individuals are the subjects of the survey. *Vagueness* characterizes phenomena which are vague in their own nature, which means they have no sharp definition. For instance, concepts such as satisfaction, trust, happiness and stress, well define the underlying phenomenon, without quantifying it. That said, this type of information codification is very common and more in line with the human way of thinking than any other type of codification.

The present paper proposes the use of Partial Possibilistic Regression Path Modeling (PPR-PM) (Romano and Palumbo, 2013) in the subjective measurement framework.

Following the PLS-PM approach, PPR-PM aims to explain at best the residual variance in any regression inside the model, but it is based on the use of possibilistic regression to model relations, which is geared to model vagueness rather than randomness.

Different approaches have been proposed to cope with vagueness in regression analysis. For sake of the simplicity they can be grouped into two broad categories: Fuzzy Least Square Regression (FLSR) and Possibilistic Regression (PR). Two papers can be considered seminal for each approach, while many others have proposed further developments.

Diamond's papers (Diamond 1988, Diamond 1990) introduce the FLSR approach (see also Coppi et *al.* 2006), which is closer to the traditional statistical approach. In fact, following the Least Squares line of thought, the aim is to minimize the distance between the observed and the estimated fuzzy data. This approach has been extended to the interval data analysis (Blanco-Fernandez *et al.*, 2011; Billard and Diday, 2000; Marino and Palumbo, 2003) and to symbolic data analysis (see Lima Neto and de Carvalho, 2010).

The paper by Tanaka *et al.* (1982) and that by Tanaka (1987) introduced the PR approach. We refer the reader to the book by Tanaka and Guo (1999) for an exhaustive overview of possibilistic data analysis. In the perspective of this paper, it is worth noting that in PR the error term is embedded in the interval parameters that model the vagueness in the relation among the variables.

PPR-PM is a flexible methodology for analyzing phenomena characterized by complex structures of relations among the variables and where the vagueness is the major source of uncertainty.

In PPR-PM the process of data analysis is represented by the equation: *data = possibilistic model*. Unlike the classical statistical paradigm, where uncertainty is considered an additional element to the deterministic relation among the variables, possibilistic regression considers uncertainty as being reflected inside the model via the parameters.

In the following, we will first introduce the PR and then present the basic algorithm of PPR-PM. A case study on a meta-cognitive questionnaire for teachers will be illustrated. The paper will end with the main conclusions and some open issues.

2. Methods

2.1. Possibilistic Regression

In a general framework, PR defines the relation between one dependent variable Y and a set of P predictors $X_1, X_p, \ldots, X_P$ through a linear function holding interval valued coefficients:

$$Y = \tilde{\omega}_1 X_1 + \ldots + \tilde{\omega}_p X_p + \ldots + \tilde{\omega}_P X_P \tag{1}$$

where $\tilde{\omega}_p$ denotes the generic interval coefficient in terms of midpoint and spread: $\tilde{\omega}_p = \{c_p; a_p\}$. There are no restrictive assumptions on the model. Unlike statistical regression, the deviations between data and linear models are assumed to depend on the imprecision of the parameters and not on measurement errors. This means that in PR there is no external error component but the spread of the coefficients embeds all uncertainty, such that PR minimizes the total spread of the interval coefficients:

$$\min_{a_p} \sum_{p=1}^{P} \left(\sum_{n=1}^{N} a_p |x_{np}| \right), \forall p = 1, \ldots, P \tag{2}$$

under the following linear constraints:

$$\begin{aligned} \sum_{p=1}^{P} c_p x_{np} + \sum_{p=1}^{P} a_p |x_{np}| \geq y_n, \forall n = 1, \ldots, N, \\ \sum_{p=1}^{P} c_p x_{np} - \sum_{p=1}^{P} a_p |x_{np}| \leq y_n, \forall n = 1, \ldots, N. \end{aligned} \tag{3}$$

satisfying the following conditions: *i)* $a_p \geq 0$, *ii)* $c_p \in R$, *iii)* $x_{n1} = 1$. Constraints in (3) guarantee the inclusion of the whole given data set in the estimated boundaries.

In a geometric view, where statistical units are represented as points in the $\Re^{P+1}$ space, the optimal solution ensures the inclusion of the whole given data set in the estimated boundaries with the minimum spread of parameters.

Wang and Tsaur (2000) provided a suitable interpretation of the regression interval. The basic idea was to find a representative value of the interval among the infinite values enclosed within the interval boundaries. Let $\underline{y}_n$ and $\overline{y}_n$ be the lower and the upper bound of the estimated value $\tilde{y}_n^*$. The authors proved that in models with symmetric coefficients the mean value of $\tilde{y}_n^*$ given by:

$$y_n^m = \frac{\underline{y}_n + \overline{y}_n}{2}$$

is equal to the value occurring with the higher possibility level and denoted by $\tilde{y}_n^1$. In other words, $\tilde{y}_n^1$ is the best representative value of the possibilistic interval and, more generally, the regression line $\tilde{Y}^1$ has the best ability to interpret the given data. Starting from these results the following quantities were defined:

- *Total Sum of Squares* (SST)
  a measure of the total variation of y_n in $\tilde{y}_n$

$$SST = \sum_{n=1}^{N} \left(y_n - \underline{y}_n \right)^2 + \sum_{n=1}^{N} \left(\overline{y}_n - y_n \right)^2 \tag{4}$$

- *Regression Sum of Squares* (SSR):
  a measure of the variation of $\tilde{y}_n^1$ in $\tilde{y}_n^*$

$$SSR = \sum_{n=1}^{N} \left(\tilde{y}_n^1 - \underline{y}_n \right)^2 + \sum_{n=1}^{N} \left(\overline{y}_n - \tilde{y}_n^1 \right)^2 \tag{5}$$

- *Error Sum of Squares* (SSE):
  an estimate of the difference when $\tilde{y}_n^1$ is used to estimate y_n

$$SSE = 2 \sum_{n=1}^{N} \left(\tilde{y}_n^1 - y_n \right)^2 \tag{6}$$

Thus, using 4 and 5, an index of confidence is built, which is similar to the traditional R^2 in statistics. The index is defined as: IC=SSR/SST, with $0 \leq IC \leq 1$, and gives a measure of the variation of Y between $\underline{Y}$ and $\overline{Y}$. The higher the IC, the better the $\tilde{Y}^1$ used to represent the given data. A high value of IC means that a well estimated PR is modeled and can support a better prediction.

2.2. Partial Possibilistic Regression Path Modeling

Partial possibilistic regression path modeling (PPR-PM) is a method to analyze phenomena whose description requires the analysis of a complex structure of relations among the variables inside the system, and where there is an additional source of complexity arising from the involvement of influential human beings. This is achieved by combining the principles of PLS-PM (Tenenhaus et al., 2005) and possibilistic regression (Tanaka and Guo, 1999). A combination of possibilistic regression and path-modeling was proposed by Palumbo and Romano (2008) and Palumbo *et al.* (2008) but in these proposals the use of possibilistic regression was limited to model the relationships among LVs. PPR-PM novelty consists in extending the approach to the whole path model. However, the extension of the possibilistic approach to the entire model may be an inappropriate choice in the context of subjective measures. Here, in fact, the items used to measure attitudes and preferences often have a skewed distribution. Moreover, the presence of outliers is very common in this context.

In this work, PPR-PM is thus further modified in order to obtain a more appropriate method for the analysis of subjective data. The main idea is to use quantile regression (Koenker and Basset, 1978; Davino et al, 2013) to model the relations between each LV and its respective block of indicators. This choice allows us on the one hand to have a robust measure of the latent variables, on the other to take into account the imprecision inherent in systems where human estimation is influential and the observations cannot be described accurately.

Let us assume P variables ($p = 1,\dots,P$) observed on N units ($n = 1,\dots,N$) and collected into a partitioned table $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_h, \dots, \mathbf{X}_H]$, where $\mathbf{X}_h$ is the generic block composed by P_h indicators. In the PLS-PM literature, it is used to distinguish the *structural model* (or inner model) linking the LVs, and the *measurement model* (or outer model) linking the LVs with their respective block of MVs. The measurement model can be *reflective* or *formative* according to the linkage between the LVs and the MVs (Tenenhaus *et al.* , 2005). In a *reflective model* the block of manifest variables related to a latent variable is assumed to measure a unique underlying concept. In a *formative model*, each manifest variable or each sub-block of manifest variables represents a different dimension of the underlying concept. Unlike the reflective model, the formative model does not assume homogeneity nor unidimensionality of the block. PPR-PM only focuses on the *reflective model*, which appears to be the appropriate model in the social studies.

In PPR-PM, an iterative procedure permits estimation of the latent variable scores and the outer weights, while path coefficients are obtained from possibilistic regressions between the estimated latent variables.

The algorithm computes the latent variables' scores alternating the *outer* and *inner* estimation till convergence. The procedure starts on centered (or standardized) MVs by choosing arbitrary weights w_{ph}. In the external estimation, the latent variable is estimated as a linear combination of its own MV:

$$\mathbf{v}_h \propto \sum_{p=1}^{P_h} w_{ph}\mathbf{x}_{ph} = \mathbf{X}_h\mathbf{w}_h \tag{7}$$

where $\mathbf{v}_h$ is the standardized outer estimation of the latent variable ξ_h and the symbol $\propto$

means that the left-hand side of the equation corresponds to the standardized right-hand side. In the internal estimation, the latent variable is estimated by considering its links with the other adjacent H' latent variables:

$$\vartheta_h \propto \sum_{h'=1}^{H'} e_{hh'} \mathbf{v}_{h'} \tag{8}$$

where ϑ_h is the standardized inner estimation of the latent variable ξ_h and the inner weights, according to the so called *centroid scheme* (Tenenhaus *et al.* , 2005), are equal to the sign of the correlation between the outer estimate $\mathbf{v}_h$ of the h-th latent variable and the outer estimate of the h' latent variable $\mathbf{v}_{h'}$ connected with $\mathbf{v}_h$.

These first two steps allow us to update the outer weights w_{ph}. The weight w_{ph} is the regression coefficient in the quantile regression of the p-th manifest variable of the $h-$th block $\mathbf{x}_{ph}$ on the inner estimate of the h-th latent variable ϑ_h:

$$\mathbf{x}_{ph} = w_{ph} \vartheta_h + \varepsilon_{ph} \tag{9}$$

The quantile regression is an extension of the classical estimation of the conditional mean to the estimation of a set of conditional quantiles (Koenker and Basset, 1978; Davino et al, 2013):

$$Q_\theta(\mathbf{x}_{ph}|\vartheta_h) = \vartheta_h w_{ph}(\theta) + \varepsilon_{ph} \tag{10}$$

where $0 < \theta < 1$, $Q_\theta(.|.)$ denotes the conditional quantile function for the θ-th quantile. In particular, PPR-PM considers only the case in which $\theta = 0.5$, i.e. the median is the single chosen quantile.

The algorithm iterates till convergence. After convergence, structural (or path) coefficients are estimated through possibilistic regression among the estimated LVs:

$$\xi_j = \tilde{\beta}_{0j} + \sum_{h:\xi_h \to \xi_j} \tilde{\beta}_{hj} \xi_h \tag{11}$$

where $\xi_j (j = 1, \ldots, J)$ is the generic endogenous (dependent) latent variable and $\tilde{\beta}_{hj}$ is the generic *interval path coefficient* in terms of midpoint and range $\tilde{\beta}_{hj} = \{c_{hj}; a_{hj}\}$, or equivalently $[\underline{\beta}_{hj}, \overline{\beta}_{hj}] = [c_{hj} \pm a_{hj}]$, interrelating the h-th exogenous (independent) variable to the j-th endogenous one. The higher the midpoint coefficient the higher the contribution to the prediction of the endogenouse LV, while the higher the spread coefficient the higher the imprecision in the relation among the considered LVs.

An important aspect to note is that in PPR-PM the model can be valitad using the same criteria defined in the PLS-PM framework. In particular, this applies to the assessment of the measurement model, which can be validated by means of the *communality index* (Tenenhaus *et al.*, 2005). However, this reasoning cannot be extended to the validation of the structural model, and even less to the global model. The reason is that traditionally the various indexes used in the PLS-PM to validate these parts of the model are based on the assessment of each individual structural equation measured by the R^2 fit index. In PPR-PM each individual structural equation is modeled by possibilistic regression which includes the error term in the parameters; thus no residuals are provided. The quality of the model is here measured by the IC index presented in section (2.1).

3. An empirical evidence: the MESI questionnaire

The case study presents research carried out in the administrative area of Naples, which has set itself the objective of investigating some dimensions that affect the quality of teaching in high schools. In particular, we examined the motivational and emotional aspects of teachers depending on the type of high school, their working position, gender and the socio-cultural context in which the teacher operates. The tool used to conduct this study was the questionnaire MESI (Motivation, Emotions, Strategies, Teaching) (Moè *et al.*, 2010), which consists of six scales that investigate job satisfaction, practices, teaching strategies, emotions, self-efficacy, and incrementally. The idea is that an effective teacher is a teacher with a high sense of self-efficacy, satisfied with his work and able to sustain himself through the activation of positive emotions in the workplace and in his personal life. The questionnaire was administered to 216 teachers working in high schools of the province of Naples. The high schools that joined the research were 15, which were divided into three different categories: Liceo (5), Technical Institute (6) and Professional Institute (4). In the following, the focus will be only on some of the scales composing the questionnaire: job satisfaction, emotions, and self-efficacy. The first scale (satisfaction) is used to assess how job satisfaction is perceived from the point of view of the teachers. It consists of five items on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). The second scale (emotions) is composed of two subscales that each measures what emotions teachers experience when they teach (teach-emotions) and what emotions they live in the role of teacher (role-emotions). The scale is composed of a total of 30 items, each of which is constituted by a specific positive or negative emotion, and for each the teacher's frequency in experiencing the emotion is recorded on a 5-point scale (1 =hardly ever, 5 = almost always). In this study, we will focus only on the positive emotions measured by 13 items, the same for both subscales. Finally, the third scale (self-efficacy) explores the perception of self-efficacy of teaching by presenting a number of situations. Originally, it consisted of 24 items to which the teacher must respond with a 9-point scale (1 = not at all, 9 = very much), how she/he feels able to deal with certain situations. However, a reduced subset of items is used in this study (9 items).

According to theoretical assumptions, we propose an empirical framework (see Figure 1) for analyzing the relationships among the subscales composing the MESI.

PPR-PM is adopted to check the research framework. Indicator reliability is assessed by looking at the standardized loadings in Table 1, where it is shown that all indicators

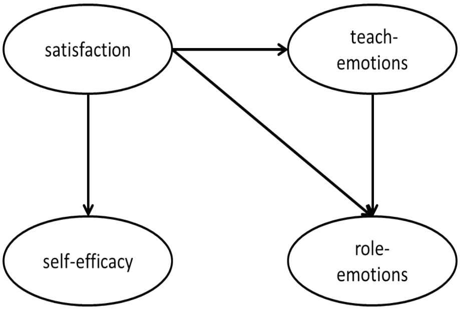


Figure 1. Structural model of the MESI questionnaire

are highly correlated with the respective constructs. To assess construct reliability, we calculate Dillon-Goldstein's ρ (DG.rho) and the communality indexes. As we show in Table 1, both the DG.rho and the communality values of all constructs are above the cut-off value of 0.7 and 0.5, respectively. That means in the first case that constructs are homogeneous and in the second case that they capture on average 64%, 61%, 46% and 50% of the variance of their indicators in relation to the amount of variance due to measurement error. Consistent with the communality, the satisfaction scale presents the highest loadings.

The results of the structural model are shown in Figure 2, where interval path coefficients are reported in terms of midpoints and spreads.

As can be seen, there is no relation between satisfaction and self-accuracy, since the path coefficient is equal to 0. Teach-emotions is positively related to satisfaction with a path coefficient equal to 0.69, which means that when a teacher is satisfied he/she feels more frequently positive emotions while teaching. Both satisfaction and teach-emotions are good predictors of role-emotions, with path coefficients equal to 0.39 and 0.22, respectively. In other words, when a teacher is satisfied he/she feels more frequently

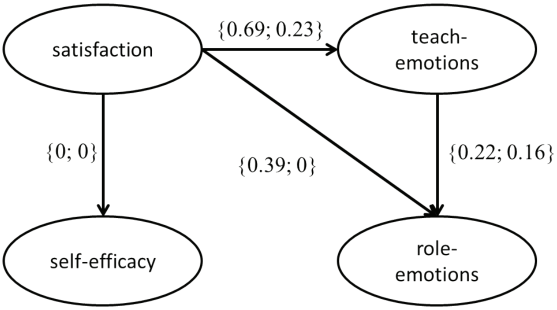


Figure 2. Structural model results of the MESI questionnaire

Table 1. Indicator and construct reliability.

LV	MV	standardized loadings	DG.rho	Communality
satisfaction	item1	0.818	0.899	0.641
	item2	0.743		
	item3	0.878		
	item4	0.834		
	item5	0.718		
self-efficacy	item1	0.713	0.934	0.608
	item2	0.862		
	item3	0.721		
	item4	0.858		
	item5	0.749		
	item6	0.821		
	item7	0.777		
	item8	0.816		
	item9	0.675		
teach-emotions	item1	0.726	0.917	0.463
	item2	0.762		
	item3	0.501		
	item4	0.568		
	item5	0.608		
	item6	0.562		
	item7	0.684		
	item8	0.680		
	item9	0.748		
	item10	0.796		
	item11	0.803		
	item12	0.743		
	item13	0.708		
role-emotions	item1	0.733	0.926	0.495
	item2	0.738		
	item3	0.487		
	item4	0.499		
	item5	0.615		
	item6	0.620		
	item7	0.781		
	item8	0.726		
	item9	0.802		
	item10	0.767		
	item11	0.813		
	item12	0.830		
	item13	0.686		

positive emotions also in his/her role as a teacher. In addition, the increase in positive emotions while teaching also increases positive emotions in the role of teacher. It is worth noting that some relations indicate a certain imprecision. This holds for the relationship between satisfaction and teach-emotions, whose path coefficient has a spread equal to 0.23, and the relationship between the latter and the role-emotion, whose path coefficient has a spread equal to 0.16.

In Table 2 the results of the PPR-PM are compared with those of the classical PLS-PM. In particular, the table shows the values of the path coefficients and of the goodness of fit indexes. As can be seen, PPR-PM results are consistent with the results obtained on the classical single valued parameters model. The weak relationship between satisfaction and self-efficacy highlighted by a path coefficient close to zero in the PPR-PM approach, is underlined by the low value of the R^2 index in PLS-PM. The coefficient between satisfaction and teach-emotions is very similar in the two approaches, but PPR-PM also provides information on the uncertainty of the relation. In other words, the spread of the coefficient shows that the variation in the opinions of the respondents with respect to these two scales is not sufficient to arrive at a precise measurement of the dependent relationship between the two scales. Finally, both approaches show that role-emotions depend on the satisfaction and teach-emotions, but the PPR-PM approach highlights the fact that there is a greater margin of imprecision in the second relation (higher spread).

Table 2. PLS-PM and PPR-PM structural model results.

Relations	PLS-PM path	R^2	PPR-PM path	IC
satisfaction > self-efficacy	0.21	0.05	{0.0; 0.0}	0.77
satisfaction > teach-emotions	0.60	0.37	{0.69; 0.23}	0.88
satisfaction > role-emotions	0.27	0.59	{0.39; 0.0}	0.80
teach-emotions > role emotions	0.56		{0.22; 0.16}	

4. Conclusion and Perspectives

The present work presented the use of PPR-PM in the context of subjective measurement. After discussing the methodological aspects, the work has focused on a case study and interpretation of the findings. It has been shown how the use of PPR-PM puts into the light the component of the uncertainty inherent in subjective evaluations, in addition to analyzing the relationship between latent concepts. On-going research concerns the possibility of considering all structural equations simultaneously. This means the interval path coefficients would be estimated by optimizing a single objective function based on the spreads of all the coefficients inside the structural model. Another line of research concerns the possibility of assessing the significance of the relationships through the use of non-parametric procedures such as those generally used in the classical approach.

References

Betzin, J., Henseler, J. (2005). Looking at the antecedents of perceived switching costs. A PLS path modeling approach with categorical indicators, in: T. Aluja *et al.* (eds.): *Proceedings of the PLS'05 International Symposium*, SPAD.

Blanco-Fernndez, A., Corral, N., Gonzlez-Rodrguez, G. (2011). Estimation of a flexible simple linear model for interval data based on set arithmetic, *Computational Statistics and Data Analysis*, **55**, 2568-2578.

Billard, L., Diday, E. (2000). Regression analysis for interval-valued data, In *Data Analysis, Classification and Related Methods*, Proc. of 7^{th} Conference IFCS, Eds. H.A.L. Kiers, J.P. Rasson, P.J.F. Groenen, M. Schader, 369-374.

Bollen, K.A. (1989). *Structural equations with latent variables*, Wiley, New York.

Coppi, R., D'Urso, P., Giordani, P., Santoro, A. (2006). Least squares estimation of a linear regression model with LR fuzzy *Computational Statistics & Data Analysis*, **51**, 267-286, `dx.doi.org/10.1016/j.csda.2006.04.036`.

Coppi, R. (2008). Management of uncertainty in Statistical Reasoning: The case of Regression Analysis, *International Journal of Approximate Reasoning*, **47**, 284-305.

Davino, C., Furno, M. and Vistocco, D. (2013). *Quantile Regression: Theory and Applications*, Wiley.

Davino, C., Romano, R. (2013). Assessing Multi-Item Scales for Subjective Measurement, in: C. Davino *et al.* (eds.): *Survey Data Collection and Integration*, Berlin Heidelberg, Springer, pp. 45-59.

Diamond, P. (1988). Fuzzy least squares, *Information Sciences*, **46**, 141-157, `dx.doi.org/10.1016/0020-0255(88)90047-3`.

Diamond, P. (1990). Least squares fitting of compact set-valued data, *Journal of Mathematical Analysis and Applications*, **147**, 531-544.

Jakobowicz, E., Derquenne, C. (2007). A modified PLS path modeling algorithm handling reflective categorical variables and a new model building strategy, *Computational Statistics and Data Analysis*, **51**, 3666-3678.

Jöreskog, K.G. (1970). A general method for analysis of covariance structures, *Biometrika*, **57**, 239-251.

Judd, C.M, McClelland, G.H. (2009). *Data Analysis: A Model Comparison Approach*, Routledge, New York.

Kaplan, D. (2000). *Structural Equation modeling: foundations and extensions*, Sage, Thousands Oaks, California.

Koenker, R., Basset, G.W. (1978). Regression Quantiles, *Econometrica*, **46**, 33-50.

Lauro, C., Nappo, D., Grassia, M.G. and Miele R. (2011). Method of Quantification for Qualitative Variables and their Use in the Structural Equations Models, in: B. Fichet *et al.* (eds.): *Classification and Multivariate Analysis for Complex Data Structures*, Berlin Heidelberg: Springer, pp. 325-333.

Lima Neto, E.A., de Carvalho, F.A.T. (2010). Constrained linear regression models for symbolic interval-valued variables, *Computational Statistics & Data Analysis*, **54**, 333-347.

Marino, M., Palumbo, F. (2002). Interval Arithmetic for the Evaluation of Imprecise Data Effects in Least Squares Linear Regression, *Statistica Applicata, Italian Journal of Applied Statistics*, **14**, 277-291.

Moè A., Pazzaglia F. and Friso G. (2010). *MESI, Motivazioni, Emozioni, Strategie e Insegnamento, Questionari metacognitivi per insegnanti*, Erickson, Trento.

Palumbo, F., Romano R. (2008). Possibilistic PLS path modeling: A new approach to the multigroup comparison, in: P. Brito (ed.): Compstat 2008, Berlin Heidelberg: Physica-Verlag, Springer, pp. 303-314.

Palumbo, F., Romano R. and Esposito Vinzi, V. (2008). Fuzzy PLS path modeling: A new tool for handling sensory data, in: C. Preisach, *et al.* (eds.): *Data Analysis, Machine Learning and Applications*, Berlin Heidelberg: Springer, pp. 689-696.

Romano R., Palumbo F. (2013). Partial Possibilistic Regression Path Modeling, in: T. Minerva, *et al.* (eds.): *Cladag 2013 Book of Abstracts*, CLEUP, pp. 409-412.

Russolillo, G. (2012). Non-Metric Partial Least Squares, *Electronic Journal of Statistics*, **6**, 1641-1669.

Tanaka, H., (1987). Fuzzy data analysis by possibilistic linear models, *Fuzzy Sets and Systems*, **24**, 363-375.

Tanaka, H., Guo, P. (1999). *Possibilistic Data Analysis for Operations Research*, Physica-Verlag, Wurzburg.

Tanaka, H., Uejima, S., Asai, K. (1982). Linear regression analysis with fuzzy model, *IEEE Trans. Sys. Man Cyber.*, **12**, 903-907.

Tenenhaus, M, Esposito Vinzi, V, Chatelin, Y-M. and Lauro, C. (2005). PLS path modeling, *Computational Statistics and Data Analysis*, **48**, 159-205.

Wang, H.F., Tsaur, R.C. (2000). Insight of a fuzzy regression model, *Fuzzy Sets and Systems*, **112**, 355-369.

Wold, H. (1975). Modelling in complex situations with soft information, in: *Third World Congress of Econometric Society*, Toronto, Canada.

Zadeh, L.A. (1973). Outline of a new approach to the analysis of complex systems and decision processes, *IEEE Trans. Systems Man and Cybernet*, **1**, 28-44.

Non-iterative methods in probit regression models based on pairs of observations and linear approximation

Subir Ghosh
Department of Statistics, University of California, Riverside, USA
E-mail: subir.ghosh@ucr.edu

Haoyu Wang[1]
Department of Statistics, University of California, Riverside, USA

Summary: Numerous methods are available for iteratively solving the maximum likelihood estimating equations (MLEEs) for estimating the parameters of probit regression models. Without a closed form solution of MLEEs, the exact theoretical properties of estimators are unavailable and the asymptotic properties are only available. This paper introduces five non-iterative methods Mu, u = 1,...,5, for approximately solving MLEEs using the exact solutions for all possible pairs of observations and the exact solution obtained by linear approximations of two weight functions in MLEEs. The method M5 is based on only the exact solutions for all possible pairs of observations and there is no linear approximation is involved. The method M1 is based only on the linear approximations with no pairs of observations method is used. The methods Mu, u=2,3,4, are combinations of linear approximations and pairs of observations methods. The validity of linear approximations is argued for the dose-response studies. Even in the situations where linear approximations are not valid, the proposed method M5 prevails because it does not depend on it. The estimators of parameters based on the proposed methods permit us to establish their theoretical properties for evaluating the proposed methods. Although the estimators of parameters by the proposed methods are not exact solutions of MLEEs, their closed forms permit to establish the theoretical properties. A real-world data is used to demonstrate the closeness of solutions of the proposed new methods with that of the standard methods available in literature and statistical software. The simulation results demonstrate that the proposed method M3 performs far better than the available methods in the literature with respect to the estimated mean squared error.

Keywords: Binary Response; Estimating Equations; Linear Approximations; Maximum Likelihood; Mill's Ratio; Probit Model; Weight Functions.

[1] Current Address: Abbvie Bio-pharmaceuticals, North Chicago, Illinois, USA

1. Introduction

The logistic and probit regression models are two popular binary response models. They are widely used in dose response modeling as for regression and classification problems in bioinformatics, image analysis, disease detection and classification, pattern recognition and numerous other situations. This paper addresses the probit modeling while Ghosh and Banerjee (2010) addressed the logistic modeling. The probit model was introduced in the pioneering paper by Bliss (1935) for analyzing binary response data with an appendix by Fisher (1935) on the method of maximum likelihood estimation. As probit models evolved from the dosage-mortality curve of Bliss to the numerous modern statistics applications in the leading book by Finney (1952), the computational advancement together with the theoretical and methodological progresses in Statistics made the implementation more user friendly with respect to fitting probit models and comparing them with other possible alternative binary response models (Albert and Anderson, 1981; Dobson and Barnett, 2008; Morgan, 1992; Silvapulle, 1981; Wedderburn, 1974cf). However, some inherent impossibility remained present during the development process, as no exact solution can be obtained for the MLEEs. This paper addresses the *fundamental complexity* in the process leading to this particular inherent impossibility for obtaining an exact solution with the understanding that various optimization algorithms may be applied in order to maximize the likelihood. The *simplicity* in the process is found by demonstrating two situations where the exact solutions for the MLEEs can be obtained. The first situation arises when we consider only two observations resulted from any two groups out of N groups. We call them as a pair of observations. The total number of such pairs of observations is $\binom{N}{2}$. We start with a pair of observations where exact solutions of the MLEEs are available and then make repeated use of this for all possible pairs of observations. All possible pairs of observations were considered for logistic regression models in Ghosh and Banerjee (2010). In this paper we introduce the second situation as the local linear approximations of the weight functions which provide exact solutions of the MLEEs. In our new proposed methods, we also integrate all possible pairs of observations with local linear approximations for probit regression models. We further compare the new methods with standard methods, which are available for everyday implementation. Here presented methods do not need any initial values as they are based on exact solutions on special situations. Furthermore, the standard methods can use the solutions of the new methods as the initial values having the proper scientific justification for the choice. In this sense, the standard method can be blended with the proposed new methods.

The paper is organized as follows. In section 2 the model is defined and a simple situation with two groups providing an exact solution of the MLEEs while reviewing the probit regression model is presented. Section 3 introduces the local linear approximations of the weight functions, which are present in the estimating equations. Exact solutions of the estimating equations are given. Five new non-iterative methods are presented for approximately solving MLEEs. In the fourth section the performances of the proposed methods with the available standard methods for the Cornfield (1962) data are compared. A simulation study is also presented in order to compare the performances of the proposed methods with the standard methods. Tables 4 and 5 demonstrate that the numerical values of the estimated mean squared errors for two parameters are smallest for the method M3 in comparison to the other proposed methods. The detailed theory behind the new methods is also given. Section 5 provides the final discussions and conclusions.

2. Probit Regression Model

In a probit regression model the observations for a response variable Y are generated from a binary random variable which takes only one of two category of values on a unit: *yes* (1) with probability p_i and *no* (0) with probability $1 - p_i$. The observed value of Y in the first category is the number of *yes* values and in the second category number of *no* values on n units. In the study of the dependence of response variables Y on explanatory variables X in N groups, the response variable for the i^{th} group is denoted by Y_i and the corresponding explanatory variable by X_i for $i = 1, ..., \mathrm{N}$. For the i^{th} group with n_i units and x_i as the fixed value of the explanatory variable X_i, the observed value of response variable Y_i in the first category is y_i and the observed value in the second category is $n_i - \mathrm{y}_i$. The random variables $\mathrm{Y}_1, ... , \mathrm{Y}_{\mathrm{N}}$ are independent. We have

$$\mathrm{E}(\mathrm{Y}_i) = n_i p_i, \ \mathrm{Var}(\mathrm{Y}_i) = n_i p_i(1 - p_i), \text{ and } \mathrm{Cov}(\mathrm{Y}_i, \mathrm{Y}_{i'}) = 0, \text{ for } i \neq i'.$$

The N groups generate the observations as $\Big(y_i, n_i - y_i; x_i; i = 1, ..., \mathrm{N}\Big)$. In the probit regression model (Bliss, 1935; Finney, 1952; Dobson and Barnett, 2008; Morgan, 1992) the probability p_i is modeled by

$$p_i \overset{\text{def}}{=} \Phi(\alpha + \beta x_i), \tag{2.1}$$

where $\Phi(.)$ is the standard normal distribution function. The parameters α and β are unknown and target of the maximum likelihood estimation procedure. Maximum likelihood estimators provide asymptotically unbiased and efficient estimates. The likelihood function is given by

$$L(\alpha, \beta; y, n, x) \overset{\text{def}}{=} \prod_{i=1}^{\mathrm{N}} \binom{n_i}{y_i} \Big(\Phi(\alpha + \beta x_i)\Big)^{y_i} \Big(1 - \Phi(\alpha + \beta x_i)\Big)^{n_i - y_i}. \tag{2.2}$$

The MLEEs are now obtained by taking the derivatives of the log−likelihood function with respect to α and β and setting them equal to zero (Dobson and Barnett (2008), Morgan (1992)).

$$\begin{cases} \sum_{i=1}^{\mathrm{N}} y_i A(\alpha + \beta x_i) - \sum_{i=1}^{\mathrm{N}} n_i B(\alpha + \beta x_i) = 0, \\ \\ \sum_{i=1}^{\mathrm{N}} y_i x_i A(\alpha + \beta x_i) - \sum_{i=1}^{\mathrm{N}} n_i x_i B(\alpha + \beta x_i) = 0, \end{cases} \tag{2.3}$$

with the "weight" functions

$$A(u) \overset{\text{def}}{=} \frac{\phi(u)}{\Phi(u)(1 - \Phi(u))}, \qquad B(u) \overset{\text{def}}{=} \frac{\phi(u)}{1 - \Phi(u)}. \tag{2.4}$$

The function $B(u)$ is known in statistics and econometrics literature as the inverse of Mill's Ratio (Mills, 1926). The function $B(u)$ is also known in reliability analysis as well as in survival analysis (Dobson and Barnett, 2008, p. 189) as the hazard function for the special case of standard normal distribution.

The solutions for α and β of the MLEEs in (2.3) are denoted by $\hat{\alpha}$ and $\hat{\beta}$. The MLEEs in (2.3) provide the exact solutions for α and β when $\mathrm{N} = 2$ in the sense that the solutions have closed-form mathematical expressions. When $\mathrm{N} = 2$ and the observations are $\Big(y_i, n_i - y_i; x_i; i = 1, 2\Big)$, we get from (2.3)

$$\hat{p}_i \overset{\text{def}}{=} \frac{y_i}{n_i} = \Phi(\hat{\alpha} + \hat{\beta} x_i), \text{ with } \mathrm{E}(\hat{p}_i) = p_i, \quad \mathrm{V}ar(\hat{p}_i) = \frac{p_i(1 - p_i)}{n_i}, \quad i = 1, 2. \tag{2.5}$$

Exact solutions of the MLEEs in (2.3) are obtained from (2.5) as

$$\hat{\alpha} = \frac{x_1 \Phi^{-1}(\frac{y_2}{n_2}) - x_2 \Phi^{-1}(\frac{y_1}{n_1})}{x_1 - x_2}, \quad \hat{\beta} = \frac{\Phi^{-1}(\frac{y_1}{n_1}) - \Phi^{-1}(\frac{y_2}{n_2})}{x_1 - x_2}. \tag{2.6}$$

When $\mathrm{N} > 2$, the numerical solutions are obtained by using the available statistical packages in R, SAS and numerous others. The methods used in the available statistical packages are referred as the standard methods. Three criterion functions are chosen for the standard method (SM) to minimize: -2log(Likelihood), Deviance, and Chi-square using the *iterative re-weighted least squares* which are referred as SM(L), SM(D), and SM(χ^2). The likelihood function is given in (2.2). The expressions of Deviance (D) and Chi-square (χ^2) are given later in (3.5) of Section 3. The complexity of finding numerical solutions of the MLEEs in (2.3) when $\mathrm{N} > 2$ is circumvented in this paper by introducing linear approximations of the functions $A(u)$ and $B(u)$ and alternatively by harnessing the strength of exact solutions for α and β for all pairs of observations.

We now consider two pairs p and p' consisting each of two groups (i_1, i_2) and (i_3, i_4). This provides the estimates:

$$\frac{y_{i_k}}{n_{i_k}} \overset{\text{def}}{=} \hat{p}_{i_k}^{(p)}, \quad k = 1, 2, 3, 4,$$

$$\hat{\alpha}^{(p)} = \frac{x_{i_1}\Phi^{-1}(\hat{p}_{i_2}^{(p)}) - x_{i_2}\Phi^{-1}(\hat{p}_{i_1}^{(p)})}{x_{i_1} - x_{i_2}}, \quad \hat{\alpha}^{(p')} = \frac{x_{i_3}\Phi^{-1}(\hat{p}_{i_4}^{(p')}) - x_{i_4}\Phi^{-1}(\hat{p}_{i_3}^{(p')})}{x_{i_3} - x_{i_4}},$$

$$\hat{\beta}^{(p)} = \frac{\Phi^{-1}(\hat{p}_{i_1}^{(p)}) - \Phi^{-1}(\hat{p}_{i_2}^{(p)})}{x_{i_1} - x_{i_2}}, \quad \hat{\beta}^{(p')} = \frac{\Phi^{-1}(\hat{p}_{i_3}^{(p')}) - \Phi^{-1}(\hat{p}_{i_4}^{(p')})}{x_{i_3} - x_{i_4}}.$$

The pairs p and p' may or may not have a group in common. The $\binom{\text{N}}{2}$ estimates of both α and β can be calculated from $\binom{\text{N}}{2}$ pairs of observations.

We now circumvent the complexity from $A(u)$ and $B(u)$ by introducing some approximations of $A(u)$ and $B(u)$. These approximations provide simplicity in finding estimates for α and β when N > 2.

3. *Linear Approximations of* $A(u)$ *and* $B(u)$

We observe that $A(u) = B(-u) + B(u)$, $A(-u)$ = $A(u)$, and $A(u)$ is symmetric about 0. Moreover, $A(u) \geq 0$ and $B(u) \geq 0$. Figure 1 demonstrates the shapes of $A(u)$ and $B(u)$ within the range $-3 \leq u \leq 3$. We have $\Phi(u) = \frac{1}{2}$ for $u = 0$, $> \frac{1}{2}$ for $u > 0$, and $< \frac{1}{2}$ for $u < 0$. In a dose-response study, we are often interested only in $\Phi(u) \leq \frac{1}{2}$ and hence the values of u satisfying $u \leq 0$. When we observe the proportion killed or cure for a particular dose, the more than 50% is not good for the environment or living beings. For this reason, we approximate $A(u)$ and $B(u)$ "locally" for $u \leq 0$ instead of "globally" for all u. This could be a serious limitation in general but is an acceptable constraint for a dose-response study. For simplicity, we implement only the linear approximation in this paper. We approximate $A(u)$ and $B(u)$ for $u \leq 0$ by

$$A(u) \approx \gamma - \delta u, \quad B(u) \approx \eta + \theta u. \tag{3.1}$$

The sample proportions $\frac{y_i}{n_i}$ are ad hoc estimators of p_i, $i = 1, ..., $N, in (2.1). If the sample proportions are all smaller than $\frac{1}{2}$, then we find from (2.1) the support in favor of $u \leq 0$.

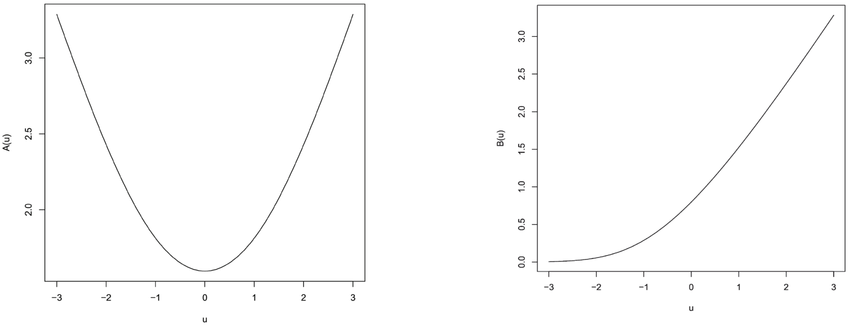


Figure 1. Plots of $A(u)$ and $B(u)$ against u

We define

$$\begin{cases} w_{1i} = (\delta y_i + \theta n_i), \qquad w_{2i} = (\gamma y_i - \eta n_i), \\ \mathbf{W} = \begin{bmatrix} \sum\limits_{i=1}^{\mathrm{N}} w_{1i} & \sum\limits_{i=1}^{\mathrm{N}} w_{1i} x_i \\ \sum\limits_{i=1}^{\mathrm{N}} w_{1i} x_i & \sum\limits_{i=1}^{\mathrm{N}} w_{1i} x_i^2 \end{bmatrix}, \quad \boldsymbol{\omega} = \begin{bmatrix} \sum\limits_{i=1}^{\mathrm{N}} w_{2i} \\ \sum\limits_{i=1}^{\mathrm{N}} w_{2i} x_i \end{bmatrix}, \quad \boldsymbol{\phi} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \\ |\mathbf{W}| = \sum\limits_{i=1}^{\mathrm{N}} w_{1i} x_i^2 \sum\limits_{i=1}^{\mathrm{N}} w_{1i} - (\sum\limits_{i=1}^{\mathrm{N}} w_{1i} x_i)^2. \end{cases} \tag{3.2}$$

From (2.3), (3.1) and (3.2), we get

$$\mathbf{W}\boldsymbol{\phi} = \boldsymbol{\omega}. \tag{3.3}$$

Estimators of α and β are calculated for given values of γ, δ, η, and θ by:

$$\begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix} \overset{\text{def}}{=} \hat{\boldsymbol{\phi}} \overset{\text{def}}{=} \mathbf{W}^{-1}\boldsymbol{\omega} = |\mathbf{W}|^{-1}\begin{bmatrix} \sum\limits_{i=1}^{N} w_{1i}x_i^2 \sum\limits_{i=1}^{N} w_{2i} - \sum\limits_{i=1}^{N} w_{1i}x_i \sum\limits_{i=1}^{N} w_{2i}x_i \\ \\ \sum\limits_{i=1}^{N} w_{1i} \sum\limits_{i=1}^{N} w_{2i}x_i - \sum\limits_{i=1}^{N} w_{1i}x_i \sum\limits_{i=1}^{N} w_{2i} \end{bmatrix}. \tag{3.4}$$

Approximate values of γ, δ, η, and θ are calculated in dependence of the given observations. In the following section, we present five methods for calculating possible values of γ, δ, η, and θ. These values are then treated as the known values in order to obtain estimates $\hat{\alpha}$ and $\hat{\beta}$ from (3.4).

We define

$$\begin{cases} \hat{p}_i \overset{\text{def}}{=} \Phi(\hat{\alpha} + \hat{\beta}x_i), \quad \hat{y}_i \overset{\text{def}}{=} n_i\hat{p}_i, \\ \Delta_y \overset{\text{def}}{=} \sum\limits_{i=1}^{N}|y_i - \hat{y}_i|, \quad \chi^2 \overset{\text{def}}{=} \sum\limits_{i=1}^{N} \frac{(y_i-\hat{y}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}, \\ D \overset{\text{def}}{=} 2\sum\limits_{i=1}^{N}\left[y_i log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) log\left(\frac{n_i-y_i}{n_i-\hat{y}_i}\right)\right], \\ A_i \overset{\text{def}}{=} A(\hat{\alpha} + \hat{\beta}x_i), \quad \hat{A}_i \overset{\text{def}}{=} \gamma - \delta(\hat{\alpha} + \hat{\beta}x_i), \quad \Delta_A \overset{\text{def}}{=} \sum\limits_{i=1}^{N}|A_i - \hat{A}_i|, \\ B_i \overset{\text{def}}{=} B(\hat{\alpha} + \hat{\beta}x_i), \quad \hat{B}_i \overset{\text{def}}{=} \eta + \theta(\hat{\alpha} + \hat{\beta}x_i), \quad \Delta_B \overset{\text{def}}{=} \sum\limits_{i=1}^{N}|B_i - \hat{B}_i|. \end{cases} \tag{3.5}$$

The values of γ and δ that give a smaller value of Δ_A represent a better linear approximation of A in (3.1). Similarly the values of η and θ in Δ_B for having a better linear approximation of B in (3.1). The values of $\hat{\alpha}$ and $\hat{\beta}$ that give *overall* smaller values of Δ_y, χ^2, -2log L, and D represent a better fit of the probit regression model to the observations.

3.1. Determining γ, δ, η, and θ

We now present five methods for determining values of γ, δ, η, and θ in (3.1).

Method 1: M1

We define u_{1i} satisfying $\Phi(u_{1i}) = \frac{y_i}{n_i}$, $i = 1, ..., \text{N}$. Assuming $u_{1i} \leq 0$, $i = 1, ..., \text{N}$, we fit a least squares line through the points $(u_{1i}, A(u_{1i}))$ to determine the values $\tilde{\gamma}_1$ and $-\tilde{\delta}_1$ as the estimated intercept and slope of the line. We also fit a least squares line through the points $(u_{1i}, B(u_{1i}))$ to determine the values $\tilde{\eta}_1$ and $\tilde{\theta}_1$ as the estimated intercept and slope of the line. With the numbers $\tilde{\gamma}_1$, $\tilde{\delta}_1$, $\tilde{\eta}_1$, and $\tilde{\theta}_1$ we now get with formula (3.4) estimates $\hat{\alpha}_1$ and $\hat{\beta}_1$ of the parameters α and β.

Method 2: M2
We choose a pair of observations for two values of i in $\Big(y_i, n_i - y_i; x_i; i = 1, ..., \mathrm{N}\Big)$. We call this pair p where $p = 1, ..., \binom{\mathrm{N}}{2}$. We now get from (2.6) the values of $\hat{\alpha}$ and $\hat{\beta}$ for the p^{th} pair, denoted by $\hat{\alpha}^{(p)}$ and $\hat{\beta}^{(p)}$. We define $u_{2i}^{(p)} = \hat{\alpha}^{(p)} + \hat{\beta}^{(p)} x_i, i = 1, ..., \mathrm{N}$. Assuming $u_{2i}^{(p)} \leq 0,\ i = 1, ..., \mathrm{N}$, we fit a least squares line through the points $(u_{2i}^{(p)}, A(u_{2i}^{(p)}))$ to determine the values $\tilde{\gamma}_2^{(p)}$ and $-\tilde{\delta}_2^{(p)}$ as the estimated intercept and slope of the line. We also fit a least squares line through the points $(u_{2i}^{(p)}, B(u_{2i}^{(p)}))$ to determine the values $\tilde{\eta}_2^{(p)}$ and $\tilde{\theta}_2^{(p)}$ as the estimated intercept and slope of the line. The determined values $\tilde{\gamma}_2^{(p)}$, $\tilde{\delta}_2^{(p)}$, $\tilde{\eta}_2^{(p)}$, and $\tilde{\theta}_2^{(p)}$ are now applied in order to obtain the vector $\hat{\boldsymbol{\phi}}^{(p)}$ using equation (3.4). We select the best pair that gives overall small values of Δ_y, χ^2, -2log L, and D in (3.5). The values $\hat{\alpha}_2$ and $\hat{\beta}_2$ for the best pair are now taken as the estimates for α and β.

Method 3: M3
We follow Method 2 to obtain the determined values $\tilde{\gamma}_2^{(p)}$, $\tilde{\delta}_2^{(p)}$, $\tilde{\eta}_2^{(p)}$, and $\tilde{\theta}_2^{(p)}$ for the pair p of observations, $p = 1, ..., \binom{\mathrm{N}}{2}$. We denote

$$\tilde{\gamma}_3 = \frac{1}{\binom{\mathrm{N}}{2}} \sum_{p=1}^{\binom{\mathrm{N}}{2}} \tilde{\gamma}_2^{(p)},\ \tilde{\delta}_3 = \frac{1}{\binom{\mathrm{N}}{2}} \sum_{p=1}^{\binom{\mathrm{N}}{2}} \tilde{\delta}_2^{(p)},\ \tilde{\eta}_3 = \frac{1}{\binom{\mathrm{N}}{2}} \sum_{p=1}^{\binom{\mathrm{N}}{2}} \tilde{\eta}_2^{(p)},\ \tilde{\theta}_3 = \frac{1}{\binom{\mathrm{N}}{2}} \sum_{p=1}^{\binom{\mathrm{N}}{2}} \tilde{\theta}_2^{(p)}.$$

With the values $\tilde{\gamma}_3$, $\tilde{\delta}_3$, $\tilde{\eta}_3$, and $\tilde{\theta}_3$ we get using formula (3.4) $\hat{\alpha}_3$ and $\hat{\beta}_3$ as the estimates for α and β.

Method 4: M4
We follow Method 2 to find the vectors $\hat{\boldsymbol{\phi}}^{(p)}, p = 1, ..., \binom{\mathrm{N}}{2}$ from (3.4). The elements of the vectors $\hat{\boldsymbol{\phi}}^{(p)}, p = 1, ..., \binom{\mathrm{N}}{2}$, from (3.4) are denoted by $\hat{\alpha}_{(p)}$ and $\hat{\beta}_{(p)}$. We denote

$$\hat{\alpha}_4 = \frac{1}{\binom{\mathrm{N}}{2}} \sum_{p=1}^{\binom{\mathrm{N}}{2}} \hat{\alpha}_{(p)},\ \hat{\beta}_4 = \frac{1}{\binom{\mathrm{N}}{2}} \sum_{p=1}^{\binom{\mathrm{N}}{2}} \hat{\beta}_{(p)}.$$

The $\hat{\alpha}_4$ is chosen as the estimate for α and $\hat{\beta}_4$ is chosen as the estimate for β.

Method 5: M5
We follow Method 2 to obtain from (2.6) the values of $\hat{\alpha}^{(p)}$ and $\hat{\beta}^{(p)}, p = 1, ..., \binom{\mathrm{N}}{2}$. We denote

$$\hat{\alpha}_5 = \frac{1}{\binom{\mathrm{N}}{2}} \sum_{p=1}^{\binom{\mathrm{N}}{2}} \hat{\alpha}^{(p)},\ \hat{\beta}_5 = \frac{1}{\binom{\mathrm{N}}{2}} \sum_{p=1}^{\binom{\mathrm{N}}{2}} \hat{\beta}^{(p)}.$$

The $\hat{\alpha}_5$ is chosen as the estimate for α and $\hat{\beta}_5$ is chosen as the estimate for β.

The above five methods are notably different in their arriving at the estimates of α and β. The method M5 makes use of the exact solutions in (2.6) for pairs of observations but not the linear approximations in (3.1). The method M1 does exactly the opposite. The methods M2, M3, and M4 blends the linear approximations with the exact solutions for pairs of observations in three different ways.

4. Empirical Evidence with Real and Simulated Data

This section compares the methods M1$-$M5 with a real data before presenting the simulation results.

4.1. Real Data

The Cornfield data given in Cornfield (1962) were obtained from blood pressure measurements of male subjects within the age-group 40 to 59. The male subjects were divided into eight groups representing distinct blood pressure ranges (N $= 8$). For the i^{th} group, the number of male subjects (n_i), the number of subjects with heart disease (y_i), and the number of subjects without any heart disease ($n_i - y_i$) were observed. The mid-values of blood pressure ranges were taken as the values of $x_i, i = 1, ..., 8$ for the analysis. The mid-values of blood pressure ranges 117$-$126, 127$-$136, 137$-$146, 147$-$156, 157$-$166, and 167$-$186 are 121.5, 131.5, 141.5, 151.5, 161.5, and 176.5 respectively, with their first consecutive difference between 121.5 and 131.5 as 10 and the last difference between 161.6 and 176.5 as 15. The mid-value for $<$ 117 is taken as 121.5 - 10 = 111.5 and for $\geq$ 187 is taken as 176.5 + 15 = 191.5. The collected data are displayed in Table 1. Note that the sample proportions, $\frac{y_i}{n_i}, i = 1, ..., 8$, for the Cornfield data are all smaller than $\frac{1}{2}$, such that one might assume $u \leq 0$.

The here presented estimation methods M1,...,M5 and the standard methods SM(L), SM(D), and SM(χ^2) are used for estimating parameters α and β defining the probability of heart diseases in dependence on the blood pressure. The resulting estimates and the values of the optimality criteria $\Delta_y, -2log\text{L}, \text{D}$, and χ^2 are presented in Table 2. The method M1 gives the smallest, M5 the second smallest, and SM(χ^2) the largest value of Δ_y. The Δ_y values of M4, SM(L), and SM(D) are very close to each other. The values for M2 and M3 are also very close. The methods M1, M5, and SM(χ^2)

Table 1. The Cornfield Data

i	1	2	3	4	5	6	7	8
n_i	156	252	284	271	139	85	99	43
y_i	3	17	12	16	12	8	16	8
$n_i - y_i$	153	235	272	255	127	77	83	35
x_i	111.5	121.5	131.5	141.5	151.5	161.5	176.5	191.5
$\frac{y_i}{n_i}$	0.019	0.067	0.042	0.059	0.086	0.094	0.162	0.186

Table 2. The methods M1, M2, M3, M4 and M5 versus SM(L), SM(D), and SM(χ^2)

	$(\hat{\alpha}, \hat{\beta})$	Δ_y	$-2log$L	D	χ^2
M1	(-3.24837, 0.01223)	16.45	38.99	6.28	7.14
M2	(-3.11237, 0.01143)	17.93	38.83	6.13	6.50
M3	(-3.11827, 0.01145)	17.85	38.84	6.14	6.56
M4	(-3.17931, 0.01184)	17.20	38.83	6.13	6.71
M5	(-3.36549, 0.01322)	16.77	38.98	6.28	7.06
SM(L)	(-3.19699, 0.01205)	17.17	38.76	6.06	6.51
SM(D)	(-3.19699, 0.01205)	17.17	38.76	6.06	6.51
SM(χ^2)	(-3.07957, 0.01136)	18.63	38.94	6.24	6.31

Table 3. The values Δ_A and Δ_B for the methods M1, M2, and M3

	Δ_A	Δ_B
M1	0.12307	0.09250
M2	0.10474	0.07852
M3	0.12241	0.08265

give large, SM(L) and SM(D) small values for $-2log$L and D. The values for methods M2, M3, and M4 fall in between the large and small values. The methods M1 and M5 give large, SM(χ^2) the smallest, and M4 in between values of χ^2. The values for the methods M2, M3, SM(L), and SM(D) are closer to the smallest value of SM(χ^2) than the largest value of M1. The values of Δ_A and Δ_B presented in Table 3 demonstrate the goodness of linear approximations of $A(u)$ and $B(u)$. They are given only for the pertinent methods M1, M2, and M3. The linear approximations for method M2 perform better than the other two methods.

4.2. Simulation

In the Cornfield data, the “true” values of the parameters α and β defining the probability of heart diseases in dependence on the blood pressure are unknown. Consequently, we have compared the closeness of estimates of α and β obtained by our methods M1, M2, M3, M4, and M5 to the values obtained from SM(L), SM(D), and SM(χ^2). We now perform a simulation study with the given values of $\alpha = -3.2$ and $\beta = 0.01$ and the same number of groups with sample sizes n_i and explanatory variables x_i as in Table 1. Then we use the program R to randomly generate y_i values for the probit regression model in the setup of Cornfield data. For this simulated data, the estimates $\hat{\alpha}$ of α and $\hat{\beta}$ of β are obtained by the proposed methods as well as the standard methods. We generate 100,000 samples of this simulated data in order to get insight in the behavior of the different estimators for the parameters α and β. The summary statistics of the 100,000 values of $(\hat{\alpha} + 3.2)$ are then calculated. Similarly for the 100,000 values of $(\hat{\beta} - 0.01)$ are also calculated.

Table 4 presents the values of $\widehat{\text{Bias}}(\hat{\alpha})$, $\widehat{\text{Var}}(\hat{\alpha})$, and $\widehat{\text{MSE}}(\hat{\alpha})$ for the proposed methods M1$-$M5 as well as SM(L), SM(D), and SM(χ^2). Table 5 presents the similar values of $\widehat{\text{Bias}}(\hat{\beta})$, $\widehat{\text{Var}}(\hat{\beta})$, and $\widehat{\text{MSE}}(\hat{\beta})$. On the one hand the method M3 gives the smallest values of $\widehat{\text{MSE}}(\hat{\alpha})$, $\widehat{\text{MSE}}(\hat{\beta})$, $\widehat{\text{Var}}(\hat{\alpha})$, and $\widehat{\text{Var}}(\hat{\beta})$ in comparison with M1, M2, M4, M5, SM(L), SM(D), and SM(χ^2). On the other hand the method M3 gives the larger values of $\widehat{\text{Bias}}(\hat{\alpha})$ and $\widehat{\text{Bias}}(\hat{\beta})$ in comparison to SM(L), SM(D), and SM(χ^2). Again, on the one hand the smaller values of $\widehat{\text{MSE}}(\hat{\alpha})$ and $\widehat{\text{MSE}}(\hat{\beta})$ are provided by the methods M2 and M4 than the methods SM(L) and SM(D) while on the other hand the smaller value of $\widehat{\text{MSE}}(\hat{\alpha})$ is provided by the method SM(χ^2) than the methods M2 and M4. The method M4 gives a value of $\widehat{\text{MSE}}(\hat{\beta})$ smaller than the methods SM(L), SM(D), and SM(χ^2). The methods M1 and M5 perform poorly in comparison to all the other methods. Overall, the method M3 performs better than the standard methods SM(L), SM(D), and SM(χ^2) with respect to $\widehat{\text{MSE}}$ and $\widehat{\text{Var}}$ but the standard methods perform better over M3 with respect to $\widehat{\text{Bias}}$. The criterion MSE has turned out to be

Table 4. The values of $\widehat{Bias}(\hat{\alpha})$, $\widehat{Var}(\hat{\alpha})$, and $\widehat{MSE}(\hat{\alpha})$ for the proposed methods M1−M5, SM(L), SM(D), and SM(χ^2)

	$\widehat{\text{Bias}}(\hat{\alpha})$	$\sqrt{\widehat{\text{Var}}(\hat{\alpha})}$	$\widehat{\text{MSE}}(\hat{\alpha})$
M1	-0.06209	0.51349	0.26753
M2	0.00960	0.43139	0.18619
M3	0.06204	0.40639	0.16900
M4	-0.02041	0.42729	0.18299
M5	-0.02227	0.48936	0.23997
SM(L)	-0.00198	0.44081	0.19432
SM(D)	-0.00198	0.44081	0.19432
SM(χ^2)	0.02049	0.42494	0.18099

Table 5. The values of $\widehat{Bias}(\hat{\beta})$, $\widehat{Var}(\hat{\beta})$, and $\widehat{MSE}(\hat{\beta})$ for the proposed methods M1−M5, SM(L), SM(D), and SM(χ^2)

	$\widehat{\text{Bias}}(\hat{\beta})$	$\sqrt{\widehat{\text{Var}}(\hat{\beta})}$	$\widehat{\text{MSE}}(\hat{\beta})x10^{-5}$
M1	0.00017	0.00340	1.15889
M2	-0.00010	0.00295	0.87125
M3	-0.00067	0.00275	0.80114
M4	-0.00009	0.00288	0.83025
M5	-0.00003	0.00338	1.14253
SM(L)	-0.00003	0.00301	0.90610
SM(D)	-0.00003	0.00301	0.90610
SM(χ^2)	-0.00002	0.00290	0.84104

important and meaningful to focus on because a major reduction in variance is achieved at the cost of allowing a small bias. The method M3 is precisely achieving this for estimating the parameters α and β. In the regression model parameter estimation by the method of ridge regression, the similar situations arise as noted on page 5 in the paper by Marquardt and Snee (1975). See also the article by Hoerl and Kennard (2000). As in the ridge regression, the MSE provides a meaningful comparison of the biased estimates in this simulation study by considering both bias and variance in a combined form.

5. Discussions and Conclusions

Exact solutions of MLEEs are demonstrated for the probit regression model for pairs of observations as well as considering local linear approximations of the weight functions $A(u)$ and $B(u)$ for the MLEEs. The proposed methods M1, M2, M3, M4, and M5 make use of exact solutions of either pairs of observations or the linear approximations. The standard methods are iterative methods and require the choice of initial values for the unknown parameters α and β. The estimates provided by the proposed methods give reasonable choices of initial values for the standard methods. The Cornfield data demonstrate that the performance of proposed methods are comparable to the standard methods. The methods M2, M3, and M4 perform in the Cornfield example better than the methods M1 and M5 with respect to the criterion functions $-2log\mathrm{L}$, D, and χ^2, whereas the methods M1 and M5 perform best with respect to the criterion function Δ_y. Simulation results also demonstrate that the proposed method M3 is better than the standard methods as well as the other proposed methods with respect to $\widehat{\mathrm{MSE}}$. The attractions of the proposed methods are their simplicity, that they are non-iterative, and that they make use of pairs of observations or linear approximations with exact solutions. Both for the real and simulated data, the number of groups is $\mathrm{N} = 8$. In the simulation, one could consider many other possible values of N to evaluate the performance of proposed methods relative to the standard methods.

Appendix A. Standard Errors of $\hat{\alpha}$ and $\hat{\beta}$

It follows from (2.4) that

$$B(-u) = \frac{\phi(u)}{\Phi(u)}, A(u) = A(-u) = B(u) + B(-u). \qquad (A.1)$$

The proposed local linear approximations of $A(u)$ and $B(u)$ in (3.1) are valid only for $u \leq 0$. Therefore, the local approximation $B(u) = \eta + \theta u$ in (3.1) cannot be used for calculating the local approximation $A(u) = \gamma - \delta u$ in (3.1) by the formula in $(A.1)$.

Appendix A.1. Methods M1, M2, and M3

We now denote $\hat{\alpha}$ and $\hat{\beta}$ in (3.4) by

$$\hat{\alpha} \overset{\text{def}}{=} h_1(y_1, ..., y_\mathrm{N}),\ \hat{\beta} \overset{\text{def}}{=} h_2(y_1, ..., y_\mathrm{N}). \qquad (A.1.1)$$

The first order Taylor series approximations of $\hat{\alpha}$ and $\hat{\beta}$ in $(A.1.1)$ about $(n_1p_1, ..., n_\mathrm{N}p_\mathrm{N})$ are

$$\begin{cases} \hat{\alpha} \simeq a_0^* + \frac{a_1}{n_1}(y_1 - n_1p_1) + ... + \frac{a_\mathrm{N}}{n_\mathrm{N}}(y_\mathrm{N} - n_\mathrm{N}p_\mathrm{N}) \ = \ a_0 + \frac{a_1}{n_1}y_1 + ... + \frac{a_\mathrm{N}}{n_\mathrm{N}}y_\mathrm{N}, \\ \hat{\beta} \simeq b_0 + \frac{b_1}{n_1}y_1 + ... + \frac{b_\mathrm{N}}{n_\mathrm{N}}y_\mathrm{N}, \end{cases} \qquad (A.1.2)$$

where, for $i = 1, ..., \mathrm{N}$,
$a_0^* = h_1(n_1p_1, ..., n_{\mathrm{N}}p_{\mathrm{N}})$, $a_0 = a_0^* - a_1p_1 - ... - a_Np_N$,
$b_0^* = h_2(n_1p_1, ..., n_{\mathrm{N}}p_{\mathrm{N}})$, $b_0 = b_0^* - b_1p_1 - ... - b_Np_N$,
$\frac{a_i}{n_i} = \frac{\partial}{\partial y_i} h_1(y_1, ..., y_{\mathrm{N}})\Big|_{(n_1p_1,...,n_{\mathrm{N}}p_{\mathrm{N}})}$, $\frac{b_i}{n_i} = \frac{\partial}{\partial y_i} h_2(y_1, ..., y_{\mathrm{N}})\Big|_{(n_1p_1,...,n_{\mathrm{N}}p_{\mathrm{N}})}$.

The first order approximations of expectations, variances, and standard errors (SEs) of $\hat{\alpha}$ and $\hat{\beta}$ are

$$\begin{cases} \mathrm{E}(\hat{\alpha}) \simeq a_0 + \sum_{i=1}^{\mathrm{N}} a_i p_i, \quad \mathrm{E}(\hat{\beta}) \simeq b_0 + \sum_{i=1}^{\mathrm{N}} b_i p_i, \\ \mathrm{Var}(\hat{\alpha}) \simeq \sum_{i=1}^{\mathrm{N}} \frac{a_i^2}{n_i} p_i(1-p_i), \mathrm{Var}(\hat{\beta}) \simeq \sum_{i=1}^{\mathrm{N}} \frac{b_i^2}{n_i} p_i(1-p_i), \\ \hat{p}_i \overset{\mathrm{def}}{=} \Phi(\hat{\alpha} + \hat{\beta} x_i), i = 1, ..., \mathrm{N}, \\ \widehat{\mathrm{Var}}(\hat{\alpha}) \overset{\mathrm{def}}{=} \sum_{i=1}^{\mathrm{N}} \frac{a_i^2}{n_i} \hat{p}_i(1-\hat{p}_i), \widehat{\mathrm{Var}}(\hat{\beta}) \overset{\mathrm{def}}{=} \sum_{i=1}^{\mathrm{N}} \frac{b_i^2}{n_i} \hat{p}_i(1-\hat{p}_i), \\ \mathrm{SE}(\hat{\alpha}) \simeq \sqrt{\widehat{\mathrm{Var}}(\hat{\alpha})}, \ \mathrm{SE}(\hat{\beta}) \simeq \sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}. \end{cases} \qquad (A.1.3)$$

An expansion of $\Phi(u)$ is given by

$$\Phi(u) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \left(u - \frac{u^3}{3 \cdot 2} + \frac{u^5}{5 \cdot 2! \cdot 2^2} - \frac{u^7}{7 \cdot 3! \cdot 2^3} + ... \right).$$

A first order approximation of this expansion provides:

$$\Phi(u) \simeq \frac{1}{2} + \frac{1}{\sqrt{2\pi}} u \Rightarrow p_i = \Phi(\alpha + \beta x_i) \simeq \frac{1}{2} + \frac{1}{\sqrt{2\pi}} (\alpha + \beta x_i),$$

such that with the approximation of the expectation in $(A.1.3)$ follows:

$$\begin{cases} \mathrm{E}(\hat{\alpha}) \simeq \left(a_0 + \frac{1}{2} \sum\limits_{i=1}^{\mathrm{N}} a_i \right) + \frac{1}{\sqrt{2\pi}} \left(\sum\limits_{i=1}^{\mathrm{N}} a_i \right) \alpha + \frac{1}{\sqrt{2\pi}} \left(\sum\limits_{i=1}^{\mathrm{N}} a_i x_i \right) \beta, \\ \mathrm{E}(\hat{\beta}) \simeq \left(b_0 + \frac{1}{2} \sum\limits_{i=1}^{\mathrm{N}} b_i \right) + \frac{1}{\sqrt{2\pi}} \left(\sum\limits_{i=1}^{\mathrm{N}} b_i \right) \alpha + \frac{1}{\sqrt{2\pi}} \left(\sum\limits_{i=1}^{\mathrm{N}} b_i x_i \right) \beta. \end{cases}$$

We have the theorem below:
Theorem 1. If the conditions below hold for $\hat{\alpha}$ and $\hat{\beta}$ in $(A.1.2)$

$$a_0 + \frac{1}{2} \sum_{i=1}^{\mathrm{N}} a_i = 0, \sum_{i=1}^{\mathrm{N}} a_i = \sqrt{2\pi}, b_0 = \sum_{i=1}^{\mathrm{N}} b_i = 0, \sum_{i=1}^{\mathrm{N}} a_i x_i = 0, \sum_{i=1}^{\mathrm{N}} b_i x_i = \sqrt{2\pi},$$

then $\mathrm{E}(\hat{\alpha}) \simeq \alpha$ and $\mathrm{E}(\hat{\beta}) \simeq \beta$.
Theorem 1 is applicable for the methods M1, M2, and M3 which are based on formula (3.4). Consequently, the a_i's and b_i's in $(A.1.2)$ depend on the values of γ, δ, η and θ. The values of γ, δ, η, and θ satisfying the conditions of Theorem 1 reduce the bias in $\hat{\alpha}$ and $\hat{\beta}$. The determined values of γ, δ, η, and θ for the Cornfield data are given in Table 6 for the methods M1, M2, and M3.
The values of γ and η

$$\gamma = A(0) = \frac{4}{\sqrt{2\pi}}, \quad \eta = B(0) = \frac{2}{\sqrt{2\pi}},$$

are not the best compared to their values obtained by our methods for the local approximation (3.1).

Table 6. The determined values of γ, δ, η, and θ by M1, M2, and M3 for the Cornfield data

	γ	δ	η	θ
M1	1.20091	0.59886	0.52265	0.24495
M2	1.23449	0.57080	0.54820	0.26628
M3	1.24603	0.56829	0.55142	0.26743

Appendix A.2. Methods M4 and M5

In this subsection we consider the pairs of observations p, $p = 1, ..., \binom{\mathrm{N}}{2}$, for the methods M4 and M5. The $\hat{\alpha}_{(p)}$'s for the method M4 are correlated with each other and the same is true for $\hat{\beta}_{(p)}$'s. Denoting the a_i and b_i in $(A.1.2)$ for the pair p by $a_i^{(p)}$ and $b_i^{(p)}$, $i = 1, ..., \mathrm{N}$, we get for two pairs p and p'

$$\begin{cases} \mathrm{Cov}(\hat{\alpha}_{(p)}, \hat{\alpha}_{(p')}) \simeq \sum_{i=1}^{\mathrm{N}} \frac{a_i^{(p)} a_i^{(p')}}{n_i} p_i(1-p_i), \\ \mathrm{Cov}(\hat{\beta}_{(p)}, \hat{\beta}_{(p')}) \simeq \sum_{i=1}^{\mathrm{N}} \frac{b_i^{(p)} b_i^{(p')}}{n_i} p_i(1-p_i). \end{cases}$$

We now present the approximate standard errors of $\hat{\alpha}_4$ and $\hat{\beta}_4$ for the method M4:

$$\begin{cases} \hat{p}_i^{(4)} \overset{\mathrm{def}}{=} \Phi(\hat{\alpha}_4 + \hat{\beta}_4 x_i), i = 1, ..., \mathrm{N}, \\ \widehat{\mathrm{Cov}}(\hat{\alpha}_{(p)}, \hat{\alpha}_{(p')}) \overset{\mathrm{def}}{=} \sum_{i=1}^{\mathrm{N}} \frac{a_i^{(p)} a_i^{(p')}}{n_i} \hat{p}_i^{(4)}(1-\hat{p}_i^{(4)}), \\ \widehat{\mathrm{Cov}}(\hat{\beta}_{(p)}, \hat{\beta}_{(p')}) \overset{\mathrm{def}}{=} \sum_{i=1}^{\mathrm{N}} \frac{b_i^{(p)} b_i^{(p')}}{n_i} \hat{p}_i^{(4)}(1-\hat{p}_i^{(4)}), \\ \left[\binom{\mathrm{N}}{2}\right]^2 \widehat{\mathrm{Var}}(\hat{\alpha}_4) \simeq \sum_{p=1}^{\binom{\mathrm{N}}{2}} \sum_{p'=1}^{\binom{\mathrm{N}}{2}} \widehat{\mathrm{Cov}}(\hat{\alpha}_{(p)}, \hat{\alpha}_{(p')}), \\ \left[\binom{\mathrm{N}}{2}\right]^2 \widehat{\mathrm{Var}}(\hat{\beta}_4) \simeq \sum_{p=1}^{\binom{\mathrm{N}}{2}} \sum_{p'=1}^{\binom{\mathrm{N}}{2}} \widehat{\mathrm{Cov}}(\hat{\beta}_{(p)}, \hat{\beta}_{(p')}), \\ \mathrm{SE}(\hat{\alpha}_4) = \sqrt{\widehat{\mathrm{Var}}(\hat{\alpha}_4)}, \quad \mathrm{SE}(\hat{\beta}_4) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_4)}. \end{cases}$$

The $\hat{\alpha}^{(p)}$'s for the method M5 are uncorrelated with each other if the pairs do not have any groups in common and are correlated with each other if the pairs have one group in common. The same is true for the $\hat{\beta}^{(p)}$'s. If a pair p consists of two groups i_1 and i_2 giving $\hat{p}_{i_1}^{(p)}$ and $\hat{p}_{i_2}^{(p)}$ and another pair p' consists of two groups i_1 and i_3 giving $\hat{p}_{i_1}^{(p')}$ and $\hat{p}_{i_3}^{(p')}$, then we get from (2.6)

$$\begin{cases} \mathrm{Cov}(\hat{\alpha}^{(p)}, \hat{\alpha}^{(p')}) \simeq \frac{x_{i_1}^2 \mathrm{Cov}\left(\Phi^{-1}(\hat{p}_{i_2}^{(p)}), \Phi^{-1}(\hat{p}_{i_3}^{(p')})\right)}{(x_{i_1}-x_{i_2})(x_{i_1}-x_{i_3})}, \\ \mathrm{Cov}(\hat{\beta}^{(p)}, \hat{\beta}^{(p')}) \simeq \frac{\mathrm{Cov}\left(\Phi^{-1}(\hat{p}_{i_1}^{(p)}), \Phi^{-1}(\hat{p}_{i_1}^{(p')})\right)}{(x_{i_1}-x_{i_2})(x_{i_1}-x_{i_3})}. \end{cases}$$

If a pair p consists of two groups i_1 and i_2 and another pair p' consists of two groups i_3 and i_4, then we get from (2.6)

$$\mathrm{Cov}(\hat{\alpha}^{(p)}, \hat{\alpha}^{(p')}) = \mathrm{Cov}(\hat{\beta}^{(p)}, \hat{\beta}^{(p')}) = 0.$$

We now present the approximate standard errors of $\hat{\alpha}_5$ and $\hat{\beta}_5$ for the method M5:

$$\begin{cases} \left[\binom{\mathrm{N}}{2}\right]^2 \operatorname{Var}(\hat{\alpha}_5) \simeq \sum_{p=1}^{\binom{\mathrm{N}}{2}} \sum_{p'=1}^{\binom{\mathrm{N}}{2}} \operatorname{Cov}\left(\hat{\alpha}^{(p)}, \hat{\alpha}^{(p')}\right), \\ \left[\binom{N}{2}\right]^2 \operatorname{Var}(\hat{\beta}_5) \simeq \sum_{p=1}^{\binom{\mathrm{N}}{2}} \sum_{p'=1}^{\binom{\mathrm{N}}{2}} \operatorname{Cov}\left(\hat{\beta}^{(p)}, \hat{\beta}^{(p')}\right), \\ \hat{p}_i^{(5)} \overset{\text{def}}{=} \Phi(\hat{\alpha}_5 + \hat{\beta}_5 x_i), i = 1, ..., \mathrm{N}, \\ \widehat{\operatorname{Var}}(\hat{\alpha}_5) \overset{\text{def}}{=} \left[\operatorname{Var}(\hat{\alpha}_5) \text{ with } p_i = \hat{p}_i^{(5)}, i = 1, ..., \mathrm{N}\right], \\ \widehat{\operatorname{Var}}(\hat{\beta}_5) \overset{\text{def}}{=} \left[\operatorname{Var}(\hat{\beta}_5) \text{ with } p_i = \hat{p}_i^{(5)}, i = 1, ..., \mathrm{N}\right], \\ \operatorname{SE}(\hat{\alpha}_5) = \sqrt{\widehat{\operatorname{Var}}(\hat{\alpha}_5)}, \quad \operatorname{SE}(\hat{\beta}_5) = \sqrt{\widehat{\operatorname{Var}}(\hat{\beta}_5)}. \end{cases}$$

Acknowledgements: The authors thank two reviewers whole heartedly for their careful readings of the previous version of this paper and their thoughtful constructive comments.

References

Albert, A., Anderson, J. A. (1981), Probit and logistic discriminant functions, *Communications in Statistics− Theory and Methods*, Ser. A, **10**, 641−657.

Bliss, C. I., 1935, The calculation of the dose-mortality curve, *Annals of Applied Biology,* **22**, 134−167.

Cornfield, J., 1962, Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure - a discriminant function analysis, *Federation Proceedings,* **21(4)**, Pt 2, 58−61.

Dobson, A. J., Barnett, A. J., 2008, *An Introduction to Generalized Linear Models,* 3^{rd} ed., Chapman & Hall/CRC. Florida.

Finney, D. J., 1952, *Probit Analysis,* 2nd ed., Cambridge University Press, New York.

Fisher, R. A., 1935, Appendix to Bliss, C. I. : The case of zero survivors, *Annals of Applied Biology,* **22**, 164−165.

Ghosh, S., Banerjee, H., 2010, Methods of finding the initial values of parameters in the maximum likelihood estimating equations for a logistic regression model and comparison of their final solutions using different criterion functions, *Journal of the Korean Statistical Society,* **39**, 471−477.

Hoerl, A. E., Kennard, R. W., 2000, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics,* **42**, 80−86.

Marquardt, D. W., Snee, R. D., 1975, Ridge regression in practice, *The American Statistician,* **29**, 3−20.

Mills, J. P., 1926, Table of the ratio: area to bounding ordinate, for any portion of normal curve, *Biometrika,* **18**, 395−400.

Morgan, B.J.T., 1992, *Analysis of Quantal Response Data,* Chapman & Hall/CRC. Florida.

Silvapulle, M. J., 1981, On the existence of maximum likelihood estimators for the binomial response models, *Journal of the Royal Statistical Society,* Ser. B, **43**, 310−313.

Wedderburn, R. W. M., 1974, Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika,* **61**, 439−447.

A model-based approach to validate a measuring instrument for assessing the quality of care

Federica Cugnata
Dipartimento di Economia e Statistica "Cognetti de Mrtiis", Università degli Studi di Torino
E-mail: federica.cugnata@unito.it

Chiara Guglielmetti
Dipartimento di Economia, Management e Metodi Quantitativi, Università degli Studi di Milano
E-mail: chiara.guglielmetti@unimi.it

Silvia Salini
Dipartimento di Economia, Management e Metodi Quantitativi, Università degli Studi di Milano
E-mail: silvia.salini@unimi.it

Summary: The Functional Assessment of Chronic Illness Therapy - Treatment Satisfaction - Patient Satisfaction (FACIT-TS-PS) is a measure for assessing the quality of care and satisfaction in chronically ill patients. Validity of the multi-dimensional structure and reliability of the FACIT-TS-PS were investigated in a sample of 431 chronically ill patients, using Confirmative Factor Analysis (CFA) and CUB models. Integrated use of CUB models and CFA resulted in a satisfactory structure, leading to confirmation of the original, reliable, seven-factor structure, even with a reduction in items from 25 to 15. The FACIT-TS-PS appears to be a practical instrument that is reliable and has good construct validity.

Keywords: CUB models; Confirmatory Factor Analysis; Validation; chronic illness; quality of care

1. Introduction

The first aim of this study was to validate an Italian version of the Functional Assessment of Chronic Illness Therapy - Treatment Satisfaction - Patient Satisfaction (FACIT-

TS-PS), which measures quality of care and patient satisfaction in patients with different types of chronic illness. A second, methodological, aim concerned the proposal of a new approach to validation, including a sequential use of Confirmative Factor Analysis (CFA) and CUB models. CFA is a type of Structural Equation Models that deals with the measurement model, that is the relationships between observed measures or indicators and latent variables or factors (Brown, 2006). Since the early 80s, CFA has become one of the most commonly used statistical procedures in applied research (Cole, 1987; Floyd and Widaman, 1995). Piccolo (2003) proposed CUB models, mainly motivated by psychological arguments. In these models, the answers to ordinal response items in a questionnaire are interpreted as the result of a cognitive process in which judgement is intrinsically continuous but is expressed in a discrete way within a prefixed scale of m categories. The rationale for this approach stems from interpretation of the final choices of respondents because of a complex mechanism whose main components are the *feeling* of the subject toward the item and an intrinsic *uncertainty* in choosing the ordinal value of the response (Iannario and Piccolo, 2012). The paper is organized as follows. Section 2 introduces the methods. Section 3 presents the instrument. Section 4 reports results, and Section 5 draws conclusions.

2. Methods

First a Confirmative Factor Analysis (CFA) is performed using AMOS software (Arbuckle, 2005). Differently from Explorative Factor Analysis, CFA has an hypothesis-driven nature: it must be based on an evidence-based theory which defines a model, a hypothesis about (a) the number of factors, (b) whether the factors are correlated or uncorrelated and (c) how items are associated with the factor (Santor et al., 2011). CFA is almost always used during the process of a scale validation or a translation into a different language or culture (Lai, Crane and Cella, 2006; Al-Shair et al., 2012). In this context CFA is used in validating the dimensional structure of a measure (factors) and the patterns of item-factor relationship (factor loadings). When the latent structure is multifactorial, CFA allows to understand how a test must be scored using sub-scales or using a total score. Recently CFA has been used in scale validation in addition to other statistical methods (e.g., Rasch Analysis - Krgeloh, 2012).
The 25 items are the observed variables of the model; the seven factors extracted are the latent variables. The maximum likelihood method is selected to test the model. To assess the fit of the model, the comparative fit index (CFI; Bentler, 1990)) and the root mean square error of approximation (RMSEA) are used (Browne and Cudeck, 1993). Next, the saturation coefficients among items and the latent variables are examined. To show that the items contribute to the model, they first must be saturated only on the expected factor and their coefficient of saturation must be significant.

In order to understand if a more parsimonious version of the instrument could be developed, the CUB model is applied. CUB models are a class of statistical models introduced by Piccolo (2003) for the specific purpose of interpreting and fitting ordinal responses. They are interpreted as a result of two components: *feeling* and *uncertainty*. The first component is expressed by a shifted binomial random variable. The second component is expressed by a uniform random variable. The two components are combined linearly in a mixture distribution. Let R be a random variable that assumes m possible categories, $r = 1, 2, 3, \ldots, m$. Formally, the probability distribution of the CUB model is given by:

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

Since the distribution is well defined when parameters $\pi \in (0,1]$ and $\xi \in [0,1]$, the parametric space is the (left open) unit square:

$$\Omega(\pi,\xi) = \{(\pi,\xi):\; 0 < \pi \le 1,\;\; 0 \le \xi \le 1\}\,.$$

Iannario (2010) proved that such a model is identifiable for any $m > 3$.

From an interpretive point of view, $(1-\xi)$ may be understood as a measure of the *feeling* of the respondent toward the item, whereas $(1-\pi)$ reflects *uncertainty* in the final judgement.

To improve the performance of this structure, an extension of the CUB model with covariates was proposed (Iannario, 2007; Piccolo and D'Elia, 2008). If q covariates are introduced for explaining *feeling*, the probability distribution of the CUB model, now indicated by CUB$(0, q)$ is:

$$Pr(R_i = r \mid \boldsymbol{w}_i) = \pi \binom{m-1}{r-1} \xi_i^{m-r}(1-\xi_i)^{r-1} + (1-\pi)\left(\frac{1}{m}\right)\,, \quad r = 1, 2, \ldots, m; \quad (2)$$

where

$$\xi_i = \frac{1}{1 + e^{-(\gamma_0 + \gamma_1\, w_{1i} + \cdots + \gamma_q\, w_{1q})}}\,,$$

for any $i = 1, 2, \ldots, n$, the symbols $w_{i1}, w_{i2}, \ldots, w_{iq}$ denote the observation on the covariates of the i-th subject selected to explain ξ_i.

For a positive increasing w_{ik}, $k = 1, 2, ..., q$ (all other things being equal), we see that *feeling* $(1-\xi)$ decreases for $\gamma_k > 0$, and it increases for $\gamma_k < 0$.

Asymptotic statistical inference for CUB models has been developed by Piccolo (2006); an effective EM procedure for maximum likelihood estimators has been implemented, and a program in R is freely available (Iannario and Piccolo, 2012).

CUB models are also a potentially useful tool to measure importance of ordinal variables on an overall satisfaction variable in a customer/user/patient satisfaction survey (Cugnata and Salini, 2013).

We proposed to use CUB models to validate the questionnaire. That is, we proposed to use CUB models to select only the significant items for each dimension and to obtain a more parsimonious version of the instrument. We estimated a CUB model with overall satisfaction as a dependent variable and satisfaction with items as covariates to explain *feeling*, and we used a stepwise strategy using the p-values of Wald or likelihood ratio (LR) tests to select only the significant covariates. We used a stepwise algorithm to select the best CUB(0,q) model with covariates to explain feeling:

Step 1. Start with a model with no predictors, the CUB(0,0) model.

Step 2. For every variable eligible for inclusion, estimate the CUB model with a covariate and calculate its corresponding significance based on LR or Wald.

Step 3. Choose the variable with the smallest significance value. If this value is less than a probability threshold, then add it to the model; otherwise stop the stepwise algorithm.

Step 4. Update the current model by adding a new variable. Calculate LR or Wald statistic for each variable in the current model and then calculate its corresponding significance. Choose the variable with the smallest significance.

Step 5. At each step after adding a variable, try to eliminate any variable that is not significant at some level.

Step 6. Continue until every remaining variable is significant at cut-off level and every excluded variable is insignificant, or until the variable to be added is the same as the last deleted variable.

3. The instrument

The FACIT-TS-PS is part of the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System, which is a comprehensive, extensive set of self-report instruments for the assessment of health-related quality of life (QOL) in patients with cancer or other chronic illnesses. (Cella, 1997). The measurement system, under development since 1987, began with the creation of a generic CORE questionnaire called the Functional Assessment of Cancer Therapy-General (FACT-G) targeted to the management of chronic illness. ”FACIT” (Functional Assessment of Chronic Illness Therapy) was adopted as the formal name of the measurement system in 1997 to portray the expansion of the familiar ”FACT” (Functional Assessment of Cancer Therapy) questionnaires into other chronic illnesses and conditions. Most FACIT measures have undergone a standard scale development and validation methodology, which takes place in four phases: item generation, item-reduction, scale construction, and psychometric evaluation (Webster et al, 2003).
This specific version FACIT-TS-PS was developed to assess patients perception of quality of care and related satisfaction in health care services for the chronically ill. The FACIT-TS-PS has not been published in Italy yet and, to our knowledge, there are no

validation articles for the English version, except an unpublished draft (authored by Elizabeth Hahn), which shows means and dimensions of the instrument on a sample of 51 patients (HIV and cancer) and a conference presentation.

The FACIT-TS-PS is a 25-items instrument, subdivided into seven core quality-of-care domains:
A. *Explanations* (four items) received about their illness
B. *Interpersonal* (three items) relations with health care personnel (physicians and nurses)
C. *Comprehensive* (three items) care in term of ability of multi-professional team, as a whole, to be responsible for all aspects of the disease, including the impact on personal life, relationships and work
D. *Technical quality* (three items) competence of physicians
E. *Decision making* (five items) patients involvement in care decisions
F. *Nurse* (three items) competence of nurses
G. *Trust* (four items) in physicians

Patients were required to evaluate the items on a four-point scale (0 = No, not at all; 1 = Yes, but not as much as I wanted; 2 = Yes, almost as much as I wanted; 3 = Yes, as much as I wanted).
In addition, the FACIT-TS-PS includes a three-item overall measure. The first (referred to as the recommendation item) asked patients if they would recommend the hospital to others, the second (referred to as the repeat choice item) asked patients if they would choose the same clinic or office again. Both of these items were on a three-point scale (with possible response categories of yes, maybe, and no). The third item (referred to as the satisfaction item) asked patients to rate their overall evaluation of care on a five-point scale (with response categories of poor, fair, good, very good, and excellent).
Compared to the original version of the instrument, the Italian version introduced two differences: it asked all patients to limit their answers to the last six months of service or health care at the clinic instead of letting them choose visits to rate or rating their experience in general; it skipped the repeated choice item, as some of the clinics or wards involved were the unique centre of reference for treating certain diseases.

4. Results

4.1. Descriptive statistics

Table 1 presents basic characteristics of respondents. Half of the respondents were male, and more than a quarter were older than 60. About 60% of respondents had a low educational level and almost half had a full-time job.
Table 2 shows the frequency distribution of the two response variables *Recommendation* and *Satisfaction*. The first has three response categories, the second five. The association between these two variables is quite strong as indicated by the Goodman-Kruskal

gamma coefficient equal to 0.8.

Table 1. Basic characteristics of respondents (n = 431)

Characteristic	Value	Frequencies (%)
Illness	Cardiology	53 (12.3%)
	Oncology	24 (5.57%)
	Endocrinology	73 (16.94%)
	Neurology	79 (18.33%)
	Immunology	33 (7.66%)
	Haematology	110 (25.52%)
	Nephrology	27 (6.26%)
	Other chronic diseases	32 (7.42%)
Sex	Male	217 (50.47%)
	Female	213 (49.53%)
Age (years)	$\leq$30	17 (3.94%)
	31 to 40	47 (10.9%)
	41 to 50	105 (24.36%)
	51 to 60	141 (32.71%)
	>60	121 (28.07%)
Educational level	Lower secondary education or less	248 (59.62%)
	Upper secondary education	118 (28.37%)
	Higher education	50 (12.02%)
Employment	Full-time	192 (44.96%)
	Housewife/retired	160 (37.47%)
	Part-time	44 (10.30%)
	Student/unemployed	31 (7.26%)

Table 2. Overall responses n = 431

	Recommendation			
Satisfaction	No	Maybe	Yes	Total
Poor	2 (0.46%)	0(0%)	0 (0%)	2 (0.46%)
Fair	2 (0.46%)	5 (1.16%)	6(1.39%)	13 (3.02%)
Good	0 (0%)	27 (6.26%)	61 (14.15%)	88 (20.42%)
Very Good	1 (0.23%)	10 (2.32%)	211 (48.96%)	222 (51.51%)
Excellent	1 (0.23%)	2 (0.46%)	103 (23.90%)	106 (24.59%)
Total	6 (1.39%)	44 (10.21%)	381 (88.4%)	431 (100%)

Table 3 reports the frequency distributions for each item (in the Appendix A the full labels for the items). Individual items are measured on a four-point scale. As can be seen, all variables have a distribution concentrated in the highest categories of the scale.

Table 3. Item responses n = 431

Item	Rating frequencies (%)			
	No, not at all	Yes, but not as much as I wanted	Yes, almost as much as I wanted	Yes, as much as I wanted
A1	3 (0.7%)	24 (5.57%)	146 (33.87%)	258 (59.86%)
A2	8 (1.86%)	23 (5.34%)	130 (30.16%)	270 (62.65%)
A3	21 (4.87%)	49 (11.37%)	112 (25.99%)	249 (57.77%)
A4	9 (2.09%)	35 (8.12%)	84 (19.49%)	303 (70.3%)
B1	10 (2.32%)	38 (8.82%)	116 (26.91%)	267 (61.95%)
B2	15 (3.48%)	42 (9.74%)	129 (29.93%)	245 (56.84%)
B3	9 (2.09%)	33 (7.66%)	112 (25.99%)	277 (64.27%)
C1	12 (2.78%)	44 (10.21%)	129 (29.93%)	246 (57.08%)
C2	14 (3.25%)	28 (6.5%)	131 (30.39%)	258 (59.86%)
C3	7 (1.62%)	51 (11.83%)	140 (32.48%)	233 (54.06%)
D1	3 (0.7%)	13 (3.02%)	73 (16.94%)	342 (79.35%)
D2	6 (1.39%)	9 (2.09%)	87 (20.19%)	329 (76.33%)
D3	4 (0.93%)	20 (4.64%)	101 (23.43%)	306 (71.00%)
E1	41 (9.51%)	56 (12.99%)	114 (26.45%)	220 (51.04%)
E2	46 (10.67%)	37 (8.58%)	141 (32.71%)	207 (48.03%)
E3	34 (7.89%)	40 (9.28%)	134 (31.09%)	223 (51.74%)
E4	21 (4.87%)	49 (11.37%)	132 (30.63%)	229 (53.13%)
E5	13 (3.02%)	40 (9.28%)	138 (32.02%)	240 (55.68%)
F1	10 (2.32%)	33 (7.66%)	133 (30.86%)	255 (59.16%)
F2	10 (2.32%)	28 (6.5%)	111 (25.75%)	282 (65.43%)
F3	13 (3.02%)	28 (6.5%)	122 (28.31%)	268 (62.18%)
G1	2 (0.46%)	22 (5.1%)	135 (31.32%)	272 (63.11%)
G2	10 (2.32%)	23 (5.34%)	72 (16.71%)	326 (75.64%)
G3	2 (0.46%)	13 (3.02%)	79 (18.33%)	337 (78.19%)
G4	2 (0.46%)	16 (3.71%)	82 (19.03%)	331 (76.80%)

4.2. Confirmatory factor analysis

The instrument, FACIT-TS-PS has not been validated yet in either English or Italian. Validation of the Italian version presents some problems. An attempt was made to validate it through a CFA, see Figure 1.

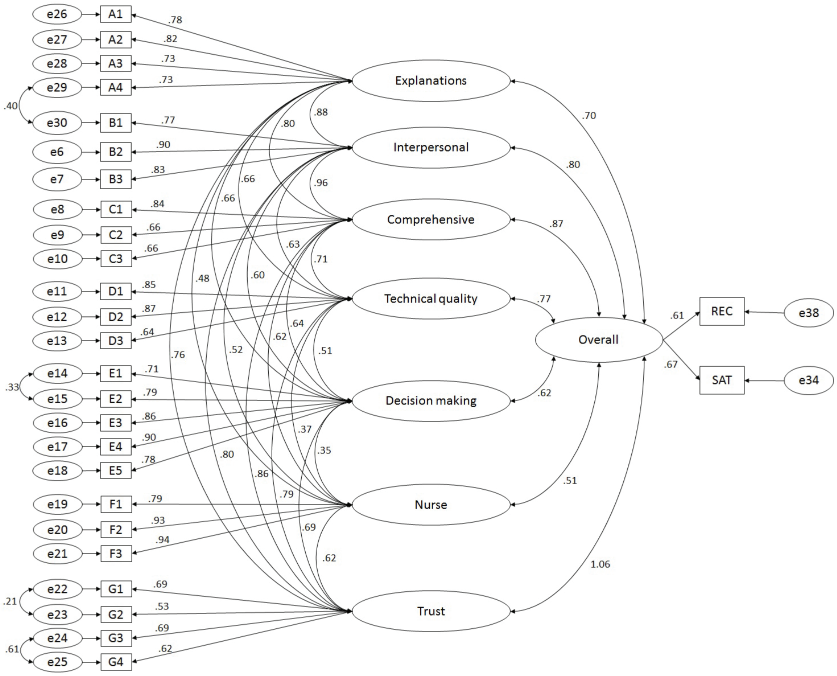


Figure 1. Complete model

A first step was to draw the path diagram for the hypothesized model. The results confirmed internal consistency of the items (in fact, the Cronbach's Alpha results were all close to 0.80). The CFI was equal to 0.897 and the RMSEA to 0.081. CFI values close to 1 indicate a very good fit, and the rule of thumb for RMSEA is that a value of the RMSEA of about .05 or less would indicate a close fit of the model in relation to the degrees of freedom. A value of about 0.08 or less for the RMSEA would indicate a reasonable error of approximation (Browne and Cudeck, 1993). We verified that all the standardized coefficients of the items were at least greater than 0.5, so all the items are considered. However, some modification indexes between the errors were very high; the modification index for a parameter is an estimate of the amount by which the discrepancy function would decrease if the analysis were repeated with the constraints on that parameter removed. We then added a covariance parameter in cases of very high modification index. In particular, as seen in Figure 1, we added covariance between items E3 and E4 (Decision Making); between items G1 and G2 (Trust) and also G3 and G4 (Trust); and between the item A4 (Explanation) and B1 (Interpersonal). The new model, presenting CFI equal to 0,926 and RMSEA equal to 0,069, was thus acceptable. It seemed, however, that some items were redundant and the model might be acceptable even with a smaller number of items. For this reason we decided to proceed with the

CUB model approach, in order to select only the significant items and to obtain a more parsimonious version of the instrument.

4.3. CUB approach

We estimated a series of CUB(0,q) models with overall measure as a dependent variable and satisfaction with items as covariates to explain feeling. We used the stepwise algorithm to select the best CUB(0,q) model with covariates to explain feeling. For each dimension, we selected the significant items to explain *feelings* of *recommendation* and *satisfaction*. We compared the significant items of *satisfaction* and of *recommendation*, and we considered all significant items of *satisfaction*, of *recommendation*, or of both.

We applied a CUB(0,25) model containing all the initial items and a CUB(0,15) model using the significant items with *recommendation* and *satisfaction* as dependent variables.

Table 4. Recommendation

	CUB(0,25)	CUB(0,15)
log-likelihood	-109	-115
AIC	272	263
BIC	381	332

Table 5. Satisfaction

	CUB(0,25)	CUB(0,15)
log-likelihood	-431	-434
AIC	916	903
BIC	1026	972

The maximized log-likelihood for the `CUB(0,25)` model for *recommendation* is -109, which is higher than the value for the `CUB(0,15)` model, which is -115. Moreover, for the first model $AIC = 272$ and $BIC = 381$ and for the second model $AIC = 263$ and $BIC = 332$. According to these criteria, the second model is preferable to the first, in fact the likelihood ratio test is not significant: $2(\ell_{25} - \ell_{15}) = 12$ and $\chi^2_{g=10} = 18.31$ at the significance level of $\alpha = 0.05$. Table 4 reports these results.

The maximized log-likelihood for the `CUB(0,25)` model for *satisfaction* is -431, which is higher than the value for the `CUB(0,15)` model, which is -434. Moreover, for the first model $AIC = 916$ and $BIC = 1026$ and for the second model $AIC = 903$ and $BIC = 972$, so according to these criteria, the second model is preferable to the first, in fact the likelihood ratio test is not significant: $2(\ell_{25} - \ell_{15}) = 6$ and $\chi^2_{g=10} = 18.31$. Table 5 reports these results.

The fact that the model with 25 items is, in fact, equal to the model with 15 (according to the likelihood ratio test) is an important result that emphasizes the need for validation of the instrument. A future effort toward validation would be to submit a shortened version of the questionnaire to patients and to analyse the results.

Figure 2 shows the map Feeling vs. Importance. The feeling value for each item is represented by the feeling $(1 - \xi\ )$, obtained by estimating as many CUB(0,0) models as there are items. Importance has been obtained using two CUB(0,15) models for *recommendation* and *satisfaction*. The coefficients $\gamma_1, \gamma_2, ..., \gamma_q$ reflect the importance of each item in predicting the feeling of the *recommendation* and of the *satisfaction*. Placement of items on the maps shows that the most important items for *recommendation* are G3 (*trust*), D2, (*technical quality*) and E4, (*decision making*) and the most important items for *satisfaction* are G3, again, and C3, (*comprehensive*).

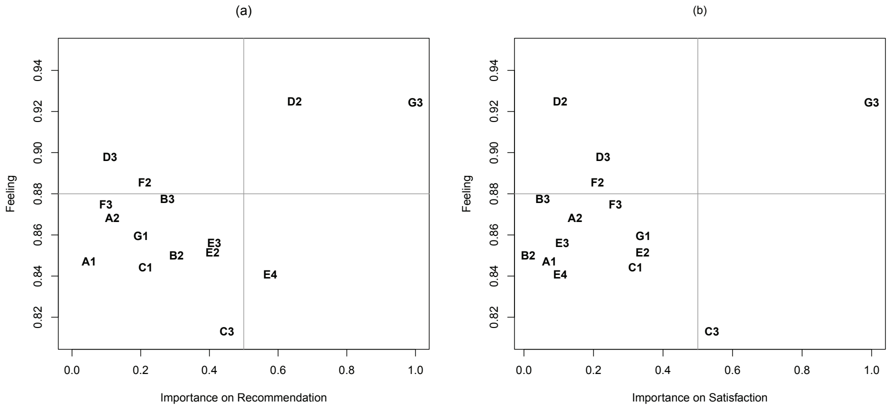


Figure 2. (a) CUB map Feeling vs. Importance on Recommendation (b) CUB map Feeling vs. Importance on Satisfaction

Finally, a CFA was done using only the 15 most significant items. Figure 3 shows the reduced model and Table 6 reports correlations among the dimensions and the overall measure that are not readable in the plot. The *trust* dimension is confirmed as the most relevant, as it was in the CUB model results.

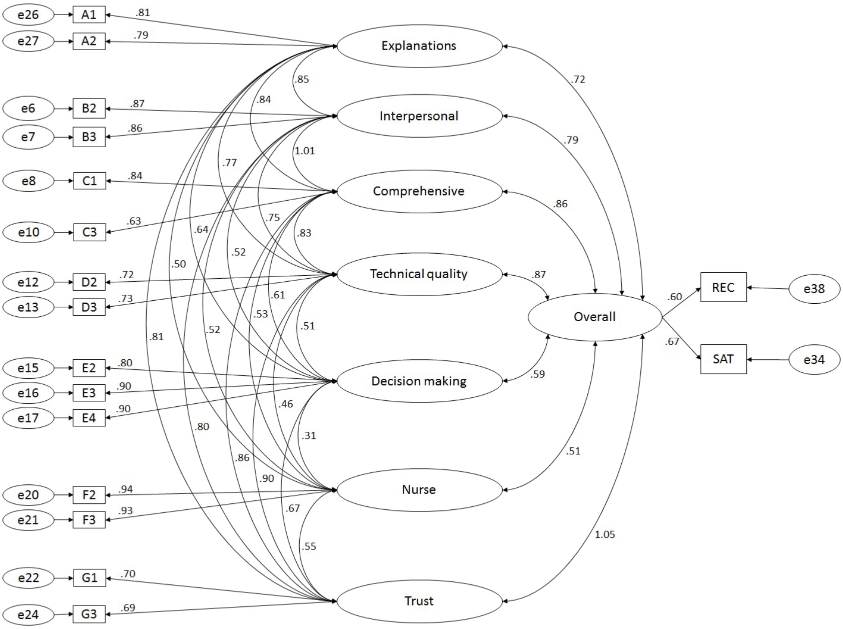


Figure 3. Reduced model

Table 6. CFA reduced model standardized weights

Dimensions	Overall
Trust	0.999
Technical Quality	0.871
Comprehensive Care	0.858
Interpersonal	0.793
Explanations	0.722
Decision Making	0.589
Nurse	0.505

Table 7 summarizes the fit of factor analysis for the complete and reduced models. Both CFI and RMSEA get acceptable values for the reduced model. Therefore, the reduced model can be accepted, and it confirms the original, reliable, seven-factor structure.

Table 7. CFA results

	Complete Model	Reduced Model
CFI	0.926	0.970
RMSEA	0.069	0.059

5. Conclusions

The Italian version of the Functional Assessment of Chronic Illness Therapy - Treatment Satisfaction - Patient Satisfaction (FACIT-TS-PS) can be considered a reliable instrument with good construct validity and a more practical structure than the original version, which included 25 items. The item reduction from 25 to 15 is particularly worthwhile, considering that the time required for completion of the questionnaire is a crucial variable, closely related to data validity, especially in a target population such as ill or older people.

The classic approach to validation pointed out some problems with the questionnaire in measuring the latent construct. Use of the CUB models, applied here for the first time to validate psychometric tests, enabled reduction of the questionnaire.

Concerning the observed very high level of patient satisfaction, the results are in agreement with previous research on patient satisfaction and quality of care (Fitzpatrick and Hopkins, 1983; vanCampen, et al., 1995; Jenkinson, et al., 2002; Gutteling, et al., 2008). Patients generally indicate that they are highly satisfied with care and that satisfaction is associated with willingness to recommend to others the hospital in which they received treatment. However, the observed discrepancy between declared overall satisfaction and the latent measure obtained from a combination of the satisfaction with individual dimensions shows that many respondents who indicated that they were satisfied with their health care on the declared satisfaction measure also indicated problems with some aspects of their experience as chronically ill patients. This seems to suggest that patient satisfaction scores, and the related issue of willingness to recommend a hospital to others, present a partial and optimistic picture. The evidence presented here suggests that satisfaction with patient care and willingness to recommend a medical facility do not imply that all aspects of that care were successfully delivered, confirming results from other industries, such as civil aviation, in which satisfaction scores may be high but customers complain about specific aspects of the service (Bethune and Huler, 1998). This result has important practical implications because health care services often use a single-item overall measure to report a high level of quality for the care they deliver.

Analysis of the importance that individual dimensions of patient experience have on overall satisfaction measures shows an interesting result: trust has the highest association with overall satisfaction measures, confirming other empirical studies (Dugan, Zheng, and Mishra, 2001). The literature demonstrates that trust in and satisfaction with health care services are closely related: trusting patients are likely to be more satisfied, and previous good experiences with a doctor are likely to foster trust. However, trust is concerned with much more than merely assessing a health care service; it is inextricably linked to the quality of the patient-doctor relationship, doctor characteristics and motivation (Dugan, Zheng, and Mishra, 2001), and it is less vulnerable than satisfaction to rapid revisions based on a single positive or negative experience (Murray and Holmes 1997). One study found that trust is better than satisfaction at predicting which patients continue their care with the same doctor and have good compliance with medical treatments (Thom, et al., 1999). Further analysis that considers demographic features of the patients (especially age and education) will allow better understanding of this almost-complete overlap between trust and satisfaction, as well as better understanding of differences in the other factors in FACIT-TS-PS.

Appendix A.

FACIT-TS-PS questions:

A1. Did your doctor(s) give explanations that you could understand?
A2. Did your doctor(s) explain the possible benefits of your treatment?
A3. Did your doctor(s) explain the possible side effects or risks of your treatment?
A4. Did you have an opportunity to ask questions?
B1. Did you get to say the things that were important to you?
B2. Did your doctor(s) seem to understand what was important to you?
B3. Did your doctor(s) show genuine concern for you?
C1. Did your doctor(s) seem to understand your needs?
C2. Did you feel that the treatment staff worked together towards the same goal?
C3. Were you able to talk to your doctor(s) when you needed to?
D1. Did you feel your doctor(s) had experience treating your illness?
D2. Did you feel your doctor(s) knew about the latest medical developments for your illness?
D3. Was the treatment staff thorough in examining and treating you?
E1. Did your doctor(s) discuss other treatments, for example, alternative medicine or new treatments?
E2. Were you encouraged to participate in decisions about your health care?
E3. Did you have enough time to make decisions about your health care?
E4. Did you have enough information to make decisions about your health care?
E5. Did your doctor(s) seem to respect your opinions?
F1. Did your nurse(s) give explanations that you could understand?
F2. Did your nurse(s) show genuine concern for you?
F3. Did your nurse(s) seem to understand your needs?
G1. Did you feel that the treatment staff answered your questions honestly?
G2. Did the treatment staff respect your privacy?
G3. Did you have confidence in your doctor(s)?
G4. Did you trust your doctor(s)' suggestions for treatment?

Acknowledgements: The authors thank the financial support of the project MIUR PRIN MISURA - Multivariate models for risk assessment.

References

Al-Shair, K., Muellerova, H., Yorke, J., Rennard, S. I., Wouters, E. F., Hanania, N. A., and Vestbo, J. (2012). Examining fatigue in COPD: development, validity and reliability of a modified version of FACIT-F scale. *Health and quality of life outcomes*, **10**(1), 100.

Arbuckle, J. (2005). *Amos 6.0* , AMOS Development Corporation, Spring House, PA.

Bentler, P. M. (1990). *Comparative fit indexes in structural models.*, Psychological bulletin, **107**(2), 238.

Bethune, G., Huler, S. (1998). *From worst to first: Behind the scenes of continental's remarkable comeback*, Wiley New York.

Browne, M. W., Cudeck, R. (1993). *Alternative ways of assessing model fit.*, Sage Focus Editions, **154**, 136-136.

Cella D. (1997). *Manual of the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System.* Center on Outcomes, Research and Education (CORE), Evanston Northwestern Healthcare and Northwestern University, Evanston IL, Version 4.

Chin, W.W. (2001). *PLS-Graph users guide*, CT Bauer College of Business, University of Houston, USA.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology*, **55**(4), 584

Cugnata, F., Salini, S. (2013). Model-based approach for importance-performance analysis, *Quality & Quantity*, 1-12.

Dugan, E., Zheng, B., Mishra, A.K. (2001). Trust in physicians and medical institutions: what is it, can it be measured, and does it matter?, *Milbank Quarterly*, **79**(4), 613–639.

Fitzpatrick, R., Hopkins, A. (1983). Problems in the conceptual framework of patient satisfaction research: an empirical exploration, *Soc Health Illness*, **5**, 297–311.

Floyd, F. J., Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, **7**(3), 286.

Gutteling, J., Darlington, A.S., Janssen, H., Duivenvoorden, H., Busschbach, J. (2008). Effectiveness of health-related quality-of-life measurement in clinical practice: a prospective, randomized controlled trial in patients with chronic liver disease and their physicians, *Quality of Life Research*, **17**(2), 195–205.

Iannario, M. (2007). A statistical approach for modelling urban audit perception surveys, *Quaderni di Statistica*, **9**, 149172.

Iannario, M. (2010). On the identiability of a mixture model for ordinal data, *METRON*, **68** (1), 8794.

Iannario, M., Piccolo, D. (2012). Cub models: Statistical methods and empirical evidence, in: R. Kenett, S. Salini (eds.): *Modern Analysis of Customer Surveys: with Applications using R*, Statistics in Practice. Wiley

Jenkinson, C., Coulter, A., Bruster, S., Richards, N., Chandola, T. (2002). Patients experiences and satisfaction with health care: results of a questionnaire study of specific aspects of care, *Quality and Safety in Health Care*, **11**(4), 335–339.

Krgeloh, C. U., Kersten, P., Billington, D. R., Hsu, P. H. C., Shepherd, D., Landon, J., Feng, X. J. (2012). Validation of the WHOQOL-BREF quality of life questionnaire for general use in New Zealand: confirmatory factor analysis and Rasch analysis. *Quality of Life Research*, 1-7.

Lai, J. S., Crane, P. K., Cella, D. (2006). Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue.*Quality of Life Research*, **15**(7), 1179-1190.

Murray, S.L., Holmes, J.G. (1997). A leap of faith? positive illusions in romantic relationships. *Personality and Social Psychology Bulletin*, **23**(6), 586–604.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 86–104.

Piccolo, D. (2006). Observed information matrix for mub model, *Quaderni di Statistica*, **8**, 33 – 78.

Piccolo, D., D’Elia, A. (2008). A new approach for modelling consumers’ preferences, *Food Quality and Preference*, **19**(3), 247 – 259.

Santor, D. A., Haggerty, J. L., Lvesque, J. F., Burge, F., Beaulieu, M. D., Gass, D., Pineault, R. (2011). An Overview of Confirmatory Factor Analysis and Item Response Analysis Applied to Instruments to Evaluate Primary Healthcare. *Healthcare Policy*, **7**(Special Issue), 79

Thom, D.H., Ribisl, K.M., Stewart, A.L., Luke, D.A., et al. (1999). Further validation and reliability testing of the trust in physician scale, *Medical care **37**(5)*, 510–517.

vanCampen, C., Sixma, H., Friele, R.D., Kerssens, J.J., Peters, L. (1995). Quality of care and patient satisfaction: a review of measuring instruments, *Medical Care Research and Review*, **52**(1), 109–133.

Webster, K., Cella, D., Yost, K. (2003). The Functional Assessment of Chronic Illness Therapy (FACIT) measurement system: properties, applications, and interpretation.*Health and Quality of Life Outcomes*, **1**(1), 79.

QdS – Vol. 15, 2013

Forecasting aggregated Euro area inflation rate with space-time models

M. Simona Andreano
Faculty of Economics, Mercatorum University
E-mail: s.andreano@unimercatorum.it

Paolo Postiglione
Department of Economic Studies, University of Chieti-Pescara
E-mail:postigli@unich.it

Summary: Economic variables are typically observed over time and across different but likely correlated areas. When interested in forecasting the aggregate across the various areas, a question that naturally arises is whether gains in efficiency can be obtained using a direct approach or an indirect approach. This issue has been recently considered in Giacomini and Granger (2004), where it is shown that stationary space-time $AR(1,1)$ models are relatively more efficient than traditional $ARMAs$ and $VARs$ models in terms of forecasting accuracy. We extend these findings by considering a more general and realistic non-stationary context, where cointegration constraints in time are allowed to exist. A concrete application with monthly inflation rate for Euro-zone economies is presented.

Keywords: Space-time models, Aggregation, Forecasting, Inflation rate.

1. Introduction

The European Monetary Union (EMU) has stimulated the need for Euro area macroeconomic studies and forecasts of area wide aggregates. European integration means that political and business decisions increasingly depend on aggregate European real economic activity, so it is of increasing interest to consider the problem of forecasting real activity measures for the Euro area as a whole. This had induced a revival of the discussion of aggregation of time series variables over the last years. Forecasting Euro area aggregates is largely a new topic and there is considerable uncertainty about the best to approach this task.

When forecasting a contemporaneously temporally aggregated variable of interest, there are different possibilities to proceed.

The analyst might ask whether it will be more efficient to forecast the aggregate series directly or to model the individual components separately, and then aggregate the forecasts.

The theoretical literature shows that aggregating component forecasts is at least as accurate as directly forecasting the aggregate when the data generating process (DGP) is known (Lütkepohl, 1984, 1987; Granger, 1987; Garderen *et al.* 2000; Giacomini and Granger, 2004). However in practice the DGP used for forecasting is unknown and the parameters have to be estimated from data. Usually also the process orders and other characteristics are specified from the observed time series and, hence, are uncertain. Clements and Hendry (1998) and Lütkepohl (2006) discuss the implication of these complications. If estimation and specification uncertainty are taken into account, standard theoretical results may not be true, and it turns out that forecasts based on disaggregated multiple time series may not be better and may even be inferior to forecasting an aggregate directly.

Therefore it is largely an empirical question whether aggregating forecasts of disaggregates improves forecast accuracy of the aggregate of interest The literature that tries to answer the question is fairly large, but does not provide clear guidelines. An overview of the theoretical relative efficiency of forecasting the aggregate variable directly or indirectly from the univariate components is given in Giacomini and Granger (2004), Lütkepohl (1987, 2006), and Wei and Abraham (1981).

Giacomini and Granger (2004), hereafter GG, show that aggregate forecasts from a space-time autoregressive model improve forecast performance and offer a solution to the curse of dimensionality that arises when forecasting with VARs. Ignoring spatial correlation, even when it is weak, leads to highly inaccurate forecasts. Arbia *et al.* (2011) provide many Monte Carlo simulations starting from the GG findings.

In the present paper we extend the space-time model proposed in GG by considering a more general and realistic non-stationary context. Our empirical analysis compares the forecast performances of the space-time AR model with different direct and indirect approaches, when the cointegration constraints in time are allowed to exist.

Although the main focus of this paper is on comparing forecasting models, our findings might be of interest also to macroeconomists and policy maker in the Euro-zone. Forecasting price developments in the Euro area is essential in the light of the second pillar of the European Central Bank's (ECB) monetary policy strategy. Moreover the inflation forecasts play a determinant role in the uncertainty surrounding the estimated effects of alternative monetary rules on unemployment dynamics in the Euro area.

The aggregation of forecasts of inflation is receiving increasing attention by staff at central banks in the Eurosystem. Hubrich (2005), Altavilla and Ciccarelli (2009), Arratibel *et al.* (2009), Bermingham and D'Agostino (2011) are only any of these studies. However no one of them deals the problem from a spatial perspective.

Hubrich (2005) analyses whether the accuracy of forecasting aggregate Euro area inflation can be improved by aggregating forecasts of sub-indices of the Harmonised Indices of Consumer Prices (HICP) as opposed to forecasting the aggregate HICP directly. Various models are employed and the results indicate that aggregating forecasts by component does not necessarily help forecast year-on-year inflation twelve months ahead.

Altavilla and Ciccarelli (2009) explore the role that inflation forecasts play in the uncertainty surrounding the estimated effects of alternative monetary rules on unemployment dynamics in the Euro area. They use the inflation forecasts of eight competing models that are included in a Bayesian VAR and analyse the size and the timing of these effects. Combining inflation forecasts from many models not only yields more accurate forecasts than those of any specific model, but also reduces the uncertainty associated with the real effects of policy decisions.

The paper of Arratibel *et al.* (2009) is a systematic study of the predictive power of monetary aggregates for future inflation for the cross section of New EU Member States. It provides stylized facts on monetary versus non monetary (economic and fiscal) determinants of inflation in these countries as well as formal econometric evidence on the forecast performance of a large set of monetary and nonmonetary indicators. The forecast evaluation results suggest that, as has been found for other countries before, it is difficult to find models that significantly outperform a simple benchmark, especially at short forecast horizons. Nevertheless, monetary indicators are found to contain useful information for predicting inflation at longer (3-year) horizons.

The paper of Bermingham and D'Agostino (2011) is in line with that of Hubrich (2005) and try to determine whether it is better to forecast a series directly or instead construct forecasts of its components and then sum these component forecasts. The authors analyse two price datasets, one for the United States and one for the Euro area and provide a guide to model choice. They consider multiple levels of aggregation for each dataset and different models: an autoregressive model, a factor augmented autoregressive model, a large Bayesian VAR and a time-varying model with stochastic volatility. In contrast to other paper, they find that once the appropriate model has been found, forecast aggregation can significantly improve forecast performance.

The remainder of the paper is organized as follows. The space-time AR model is discussed in Section 2, and the Section 3 introduces the forecasting models that will be used in the empirical analysis. Results are presented in Section 4, and Section 5 concludes.

2. The space-time autoregressive model

The space-time autoregressive (space-time AR) models were originally proposed by Cliff and Ord (1973) and Cliff *et al.* (1975) and generalized by Pfeifer and Deutsch (1980). Only recently there has been a renewed interest for models of spatial dependence in traditional economics, perhaps due to the increasing availability of highly disaggregated and spatially referenced data. Recent discussions and applications of the space-time AR model in econometrics can be found in Elhorst (2001, 2012) and Szulc (2000).

The space-time AR model class expresses the observation at time t and location i as a weighted linear combination of previous observations lagged in both space and time. The basic mechanism for these models is a hierarchical spatial ordering of the neighbours of each site and sequence of weighting matrices $\mathbf{W}^s$. Matrix $\mathbf{W}^s$ has elements w_{ij}^s that are non-zero if and only if sites i and j are s-th order neighbours. Therefore the spatial lag of order s can be defined as a weighted average of all observations in a given neighbouring set (see, e.g., Anselin, 1988, pp. 22–26):

$$L^s x_t = \sum_{j \in J_s} w_{ij}^s x_j \quad s = 1, 2, \ldots$$

where J_s is the set of neighbours of order s of region i.

The choice of the weights w_{ij}^s is a crucial issue in spatial econometrics. The traditional unidimensional measure adopted in spatial econometrics is based on the geographical distance, but other measures reflecting various notions of social or economic distance can be considered.

While theory will in the best practice cases drive the structure of $\mathbf{W}$, it nevertheless is true that there are a number of degrees of freedom in the exact W specification. Harris *et al.* (2011) review some alternative approaches to construct $\mathbf{W}$.

According to the predefined criterion, the weights w_{ij}^s are assumed generally to satisfy:
a) $w_{ij}^s \geq 0 \quad \forall i, j$
b) $w_{ii}^s = 0 \quad \forall i$
c) $\sum_{j \in J_s} w_{ij}^s = 1$

As in GG we consider the simplest form of a space-time model for the conditional mean of the zero-mean variable xit, a space-time $AR(1;1)$, which ignores dependence beyond the first temporal and spatial lags and where the first order refers to the temporal and the second to the spatial lag:

$$x_{it} = \phi x_{it-1} + \psi \sum_{j=1}^{k} w_{ij} x_{it-1} + \epsilon_{it} \quad i = 1, \ldots, k \quad t = 1, \ldots, T \tag{1}$$

In (1) the w_{ij} represent the elements of the $\mathbf{W}^s$ matrix with s=1, where the suffix 1 was for convenience omitted and k is the number of the spatial unit, therefore in our case of the countries.

Typically, isotropy is assumed, so that only distance between i and j is relevant, not the direction i to j.

When $\psi = 0$ equation (1) represents a time-stationary process if the condition $|\phi| < 1$ holds. However, when the stationary region for the two-parameter model is given by

(Pfeifer and Deutsch, 1980):

$$|\phi| + |\psi| < 1 \tag{2}$$

In matrix notation, equation (1) can be rewritten as:

$$\boldsymbol{x}_t = \phi \boldsymbol{x}_{t-1} + \psi W \boldsymbol{x}_{t-1} + \epsilon_t \quad t = 1, \ldots, T \tag{3}$$

where $\boldsymbol{x}_t$ is a vector of dimension k and, as outlined in GG, the space-time $AR(1;1)$ model can be seen as a special case of a $VAR(1)$ model, with the autoregressive coefficient matrix $\boldsymbol{\Gamma}$ restricted to equal $\phi \boldsymbol{I} + \psi \boldsymbol{W}$.

In the present paper we extend this model by considering a more general and realistic non-stationary context, where cointegration constraints are allowed to exist.

In case of a space-time Error Correction Model, we obtain:

$$\nabla \boldsymbol{x}_t = (\phi \boldsymbol{I} + \psi \boldsymbol{W}) \nabla \boldsymbol{x}_{t-1} + \gamma(\alpha_1 \boldsymbol{I} - \alpha_2 \boldsymbol{W}) \boldsymbol{x}_{t-1} + \epsilon_t \quad t = 1, \ldots, T \tag{4}$$

From this equation it can be seen that a spatial unit is not only influenced by its local conditions, but also by those of its neighbours, dependent on the structure of the spatial weight matrix **W**. Furthermore, the impact of these conditions is not necessarily uniform across spatial units. In our case we get $k \times k$ different spatial longrun parameter estimates. It is clear that the amount of output might be a problem of this model. Even for small values of k, it may already be rather difficult to report the estimation results compactly.

It should be stressed that, while most previous studies in the analysis of space-time data are oriented toward spatial cross-section analysis, in this paper we shift the emphasis to time series modelling techniques. In accordance with Elhorst (2001), we observe that serial dynamic effects are usually more important than spatial dynamic effects. One explanation might be that serial dependence is measured between only two observations, whereas spatial dependence must be averaged over multiple observations, which automatically restricts it. Moreover, in contrast to stationarity in time, stationarity in space is quite difficult to impose, as evidenced by Griffith and Lagona (1998).

To this end we simplify the model in (4) as follows:

$$\nabla \boldsymbol{x}_t = (\phi \boldsymbol{I} + \psi \boldsymbol{W}) \nabla \boldsymbol{x}_{t-1} + \gamma(\alpha \boldsymbol{x}_{t-1}) + \epsilon_t \quad t = 1, \ldots, T \tag{5}$$

where the cointegration constrains are defined only over time and are given, as usual, by $\alpha \boldsymbol{x}_{t-1}$. The extension of the model for lags greater than one is straightforward.

3. Forecasting spatial aggregated series

Contemporaneous aggregated time series variables can be forecasted in different ways. For example, one may directly use the aggregated series, construct a time series model for its data DGP and use that for forecasting. Alternatively, one may construct a time series model for the DGP of the disaggregated data and forecast the disaggregated series.

In the present paper we follow GG and suppose that the goal is to forecast $y_t = \sum_{i=1}^{k} x_{it}$ the aggregate of the same variable across *k* regions (or countries) related by spatial dependence [1]. With respect to GG we introduce the presence of non-stationarity. To this end the comparison of the forecasts are obtained through the following different methods:

a) The aggregate y_t can be forecasted directly by fitting a univariate $ARIMA$ model to the series y_t.
b) Univariate $ARIMA$ of each variable $x_{it}, i = 1, ..., k$ can be obtained and then aggregated.
c) The vector y_t may be forecasted by fitting a multivariate $VECM$ model to the series x_{it}. A forecast for y_t is obtained by aggregating the resulting forecasts for each x_{it}.
d) A space-time AR (with error correction) model can be used to forecast each variable x_{it}. The forecasts for each component are then aggregated.

If the goal is to compare the forecast of the aggregated series obtained through different methods, it is useful to introduce a loss function or an evaluation criterion for the forecast performance. Given such a criterion, optimal forecast may be constructed.

Here we use the minimum Mean Squared Error (MSE) forecast, where:

$$MSE = E\left(y_{t+h} - \hat{y}_t(h)\right)^2 \tag{6}$$

and $\hat{y}_t(h)$ is the *h*-step ahead forecast of future value y_{t+h}. The reader is referred to Granger (1969) and Granger and Newbold (1977) for a discussion of other forecast evaluation criteria.

GG demonstrated that under some strict assumptions about the DGP and the orders of the $ARMA$ representations of the variables of interest in each forecasting situation, the following result holds:

$$MSE^4\left(\hat{y}_t(1)\right) \leq MSE^3\left(\hat{y}_t(1)\right) \tag{7}$$

where MSE^3 refers to the VAR and MSE^4 to the space-time AR model and are restrict to the one-step-ahead forecasts of y_t, denoted by $y_{t-1}(1)$.

In case of estimation and specification uncertainty forecasts comparisons give ambiguous results. Furthermore the presence in our models of non-stationarity worse the reference framework. To this end we follow a pure empirical approach in the evaluation of forecasts of the four models.

In the following section we compare the forecasts obtained with the previously presented methods. However in our application y_t, the aggregated variables across k regions

[1] This assumption will be relaxed later in the paper.

related by spatial dependence, is computed as a weighted average of the single series:

$$y_t \equiv \sum_{i=1}^{k} v_i x_{it} \tag{8}$$

where the weights v_i sum one. As the VAR processes are closed with respect to linear transformations (Lütkepohl, 1987), the theoretical results demonstrated for the aggregation through the sum continue to be valid also in our case.

4. Data and empirical analysis

In this section we investigate the behaviour of the forecasts of the year-on-year inflation rates in % in Euro zone obtained through the four different scenarios proposed in Section 3.

The data used in the present application are the Harmonised Indices of Consumer Prices (HICPs) of the first 12 European countries that entered in the Euro-zone: Austria, Belgium, Finland, France, German, Greece, Ireland, Italy, Luxembourg, Netherland, Portugal and Spain. The data employed are of monthly frequency starting in 2001M1 until 2011M12. The sample is split into estimation and a forecast period, leaving the last three months for evaluate the forecasts 1-2 and 3 steps ahead.

Harmonised indices of consumer prices give comparable measures of inflation for the countries and country groups they are produced.

They are economic indicators that measure the change over time of the prices of consumer goods and services acquired by households.

In particular, HICPs provide the official measure of consumer price inflation in the Euro area for the purposes of monetary policy and the assessment of inflation convergence as required under the Maastricht criteria.

The HICP country group aggregates for the Euro area are calculated by Eurostat using the HICPs provided by the Member States. The Euro area aggregate is compiled as a weighted average of the countries comprising the Euro area. The country weights are derived from national accounts data for Household Final Monetary Consumption Expenditure (HFMCE), naturally expressed in Euro. The weight of a country is its share of HFMCE in the total of the country group.

The aggregated Euro index is constructed by Eurostat updating the countries entering the Euro area. However, for technical reason, we maintain fix the aggregation over only the first 12 pioneering countries.

The twelve national HICP price index in logarithm are presented in Figure 1 and the year-on-year inflation rates in % are depicted in Figure 2. In Figure 3 we have the corresponding aggregated series.

The performance of the four different aggregation methods is evaluated for the year-on-year inflation rates (in %) series.

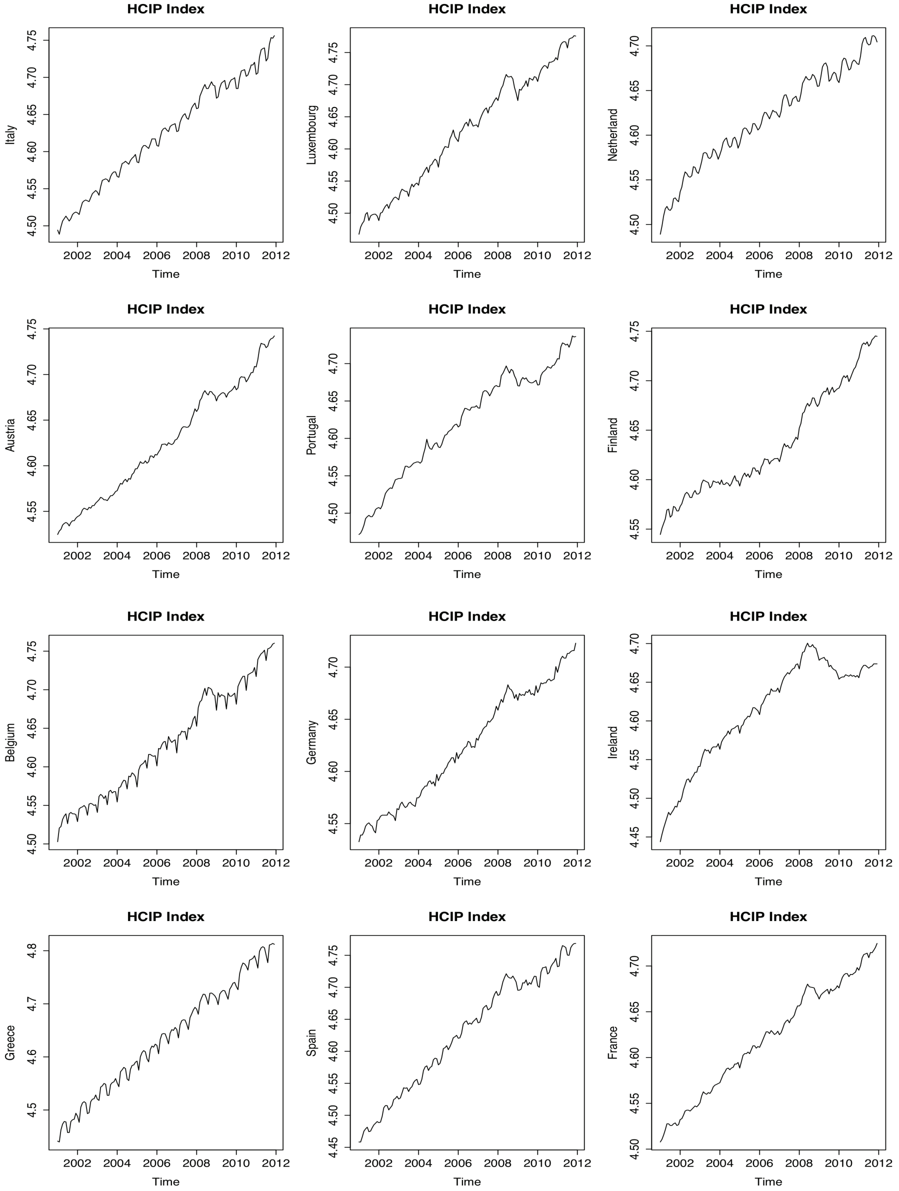


Figure 1. HICP price index in logarithm

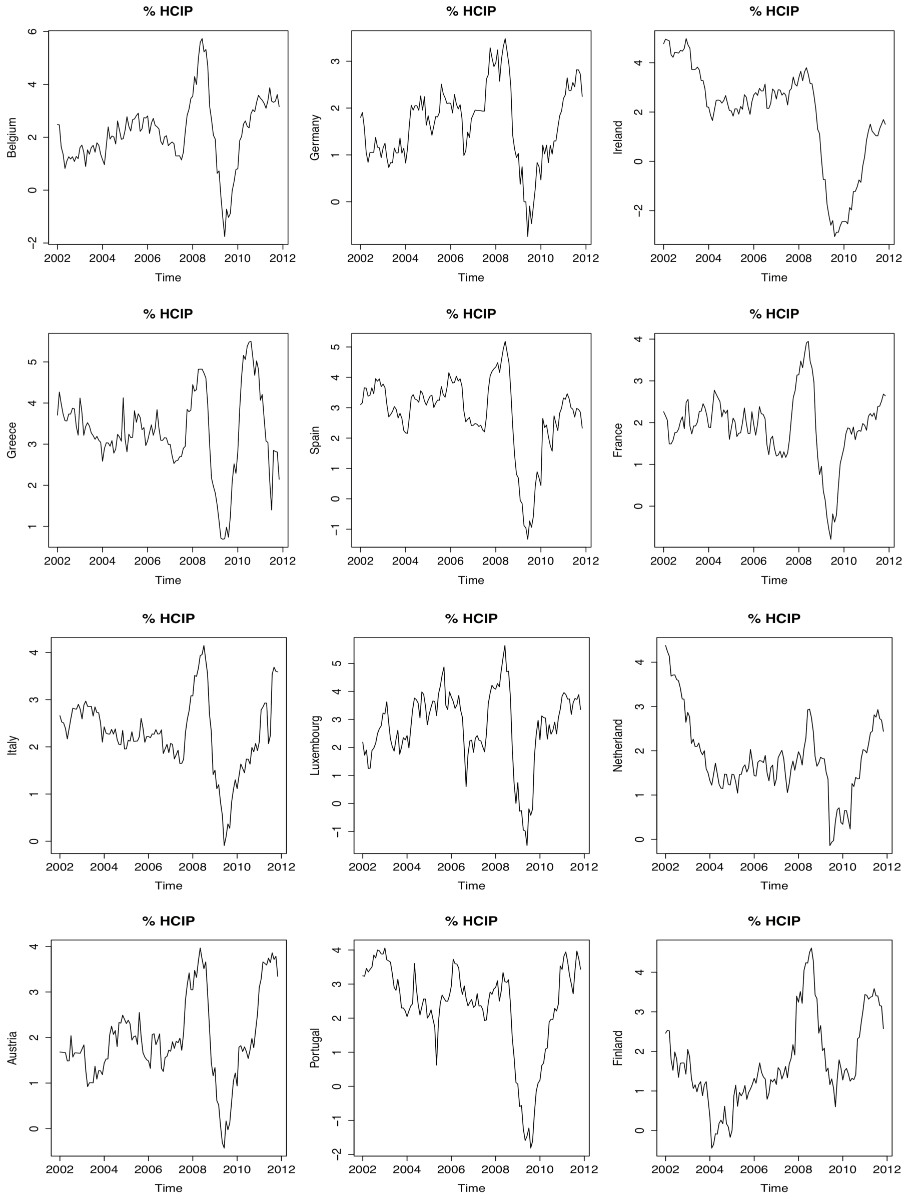


Figure 2. HICP year-on-year inflation rates in %

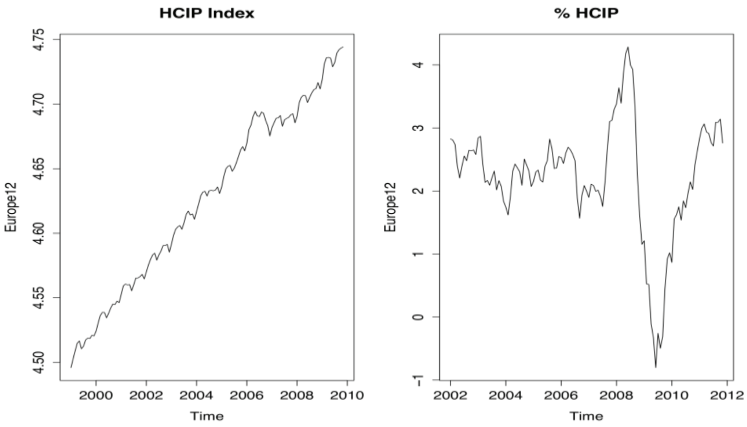


Figure 3. HICP price index in logarithm and % for EU12

The maximum forecast horizon is fixed at h = 3. Empirical and theoretical considerations have guided our choice, as forecasts with $h > 3$ have no practical interest and generally show high uncertainty.

We start the empirical analysis by performing standard Dickey-Fuller tests (DF and ADF) to investigate the order of integration of the series. The results, not reported here to save space, indicate that all series are $I(1)$ and make it possible to apply Johansen's maximum likelihood cointegration analysis. Table 1 reports the results of the trace and maximal eigenvalue (max λ) tests that verify the presence of r cointegration relationships. The model used assumes that the process has a linear trend, but that there is no trend in the cointegrating relations, as it seems appropriate for our series. Depending on the test, data support the existence of r = 10 (trace test) or r = 6 (max eigenvalue test) cointegration relationships among the twelve variables; this implies the existence of two or six common stochastic trends. In view of a parsimony criterion in the identification of the $VECM$, we choose r = 6. The cointegration analysis shows a high concordance in the long-run behaviour of the inflation rates in EU area.

We selected the order of the system by estimating $VECM$ of different lengths (from 1 to 6) and picking the one with the smallest Akaike Information Criterion (AIC). A $VECM$ of order 4 was estimated, also based on different diagnostic checking tests. For the univariate $ARIMA$ we fitted models with $p \leq 6$ and $q \leq 6$.

In the estimation of the $VECM(4)$ a number of coefficients were not significantly different from zero. We thus applied a top-down procedure to eliminate all zero coefficients, in a way to obtain a parsimonious final model and gain in efficiency.

For the estimation of the space-time AR model, firstly we had to check for the presence of spatial dependence between the 12 time series. To this end, we applied

Table 1. Johansen cointegration tests.

H_0	Trace test	$p - value$	$\max \lambda$	$p - value$
$r \leq 0$	680.9	0.0000*	128.9	0.0000*
$r \leq 1$	551.9	0.0000*	117.2	0.0000*
$r \leq 2$	434.7	0.0000*	101.7	0.0000*
$r \leq 3$	333.0	0.0000*	79.4	0.0001*
$r \leq 4$	253.6	0.0000*	67.9	0.0007*
$r \leq 5$	185.6	0.0000*	53.4	0.0073*
$r \leq 6$	132.2	0.0000*	39.3	0.0607
$r \leq 7$	92.9	0.00003*	31.8	0.0866
$r \leq 8$	61.1	0.0018*	27.5	0.0517
$r \leq 9$	33.6	0.0173*	21.1	0.0501
$r \leq 10$	12.5	0.1344	9.4	0.2540
$r \leq 11$	3.1	0.0784	3.1	0.0784

* denotes rejection of the hypothesis at least at the 0.05 level

Table 2. Moran's I test

Year	Moran's I test	$p - value$
2002	1.7052	0.04408
2006	2.5944	0.00474
2010	1.3614	0.08670

the Moran's I test for each year of the sample. In order to apply this test we need to define the spatial weight matrix. The choice of the weights is a crucial issue in spatial econometrics.

Our data are defined for the whole economy, while in general spatial models make use of lower aggregation levels. This suggests limiting the presence of spatial dependence to lag one.

In the European geographical framework, the presence of islands does not allow the definition of a simple binary contiguity weighting matrix. In this paper, the spatial weight matrix $\mathbf{W}$, is identified in terms of a row-standardized binary matrix, based on the l-nearest neighbouring regions, where each single region has the same number *(l)* of neighbours. With $l = 3$, Greece is connected to Italy, Ireland is linked with continental Europe and so on (see also Le Gallo and Dall'erba 2006).

In Table 2 we report some results of the Morans I test, for initial, median and final years.

The use of economic distance in the definition of the structure of the $\mathbf{W}$ matrix would be more realistic, however it poses problems in a time series framework. The assumption that $\mathbf{W}$ is necessarily a fixed matrix, is an issue when spatial econometrics is extended to time series or panel data modelling: $\mathbf{W}$ could evolves trough time and interacting with the regression variables (Corrado and Fingleton, 2012). Moreover, the

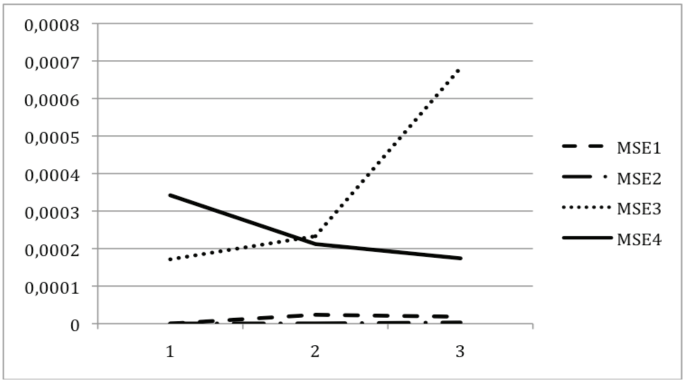


Figure 4. MSE for the different methods

sample considered in our application covers a period of significant economic changes, an assumption in contrast with the hypothesis of constant economic distance. To this end we prefer to deal with the traditional distance-based unidimensional measures used in spatial econometrics (Anselin, 1988).

The Morans I test confirms the presence of spatial dependence and this result favours the estimation of the space-time AR model. However, we note that the spatial dependence is stronger in the middle of the sample, while at the beginning and at the end of the period this dependence dimmed.

We fit a space-time $AR(4;1)$, where the first order refers to the temporal lag and the second to the spatial one. In accordance with the space-time model in (5), we introduce the six cointegration relationships.

Again, considering the large number of zeros in the matrix $\mathbf{W}$, we applied a parsimony procedure also to the space-time model. The final spatial model includes only 72 parameters out of the about 250 of the $VECM$.

The estimated MSE of the different methods are shown in Figure 4, and computed recursively for 1, 2 and 3 steps ahead forecasts, with:

$$M\hat{S}E(h) = \sum_{i=1}^{h} \frac{(y_{t+i} - \hat{y}_t(i))^2}{h} \qquad h = 1, 2, 3 \tag{9}$$

As MSEs are generally highly affected from the presence of extreme values, we report in Table 3 other evaluation criteria [2] which are likely less responsive.

The performances of the Mean Absolute Error MAE, the Mean Absolute Percentage Error $MAPE$ and the Root Mean Square Error $RMSE$ are all similar to those of the MSE.

[2] We report only the results for $h = 3$ to save space.

Table 3. Forecasting Evaluation Criteria for h=3

Model	*MAE*	*MAPE*	*RMSE*
AR-uni	0.000325	0.01125	0.000429
AR-multi	0.000143	0.00513	0.000173
VECM	0.002333	0.08286	0.002610
sp-AR	0.001248	0.04305	0.001319

Table 4. Diebold-Mariano test

h	H_0	*p-value*
h=1	$e_1 = e_2$	0.2578
	$e_3 = e_4$	0.2129
	others	0.0000
h=2	$e_3 = e_4$	0.1851
	others	0.0.0000
h=3	All	0.0000

The relative performance forecast of the space-time AR model versus the $VECM$, i.e. MSE^4 versus MSE^3, is evident in Figure 4. The aggregate forecast of the space-time AR model is overall more efficient, and the efficiency increases with h. However, the observed differences are very small. Therefore, we applied the Diebold and Mariano (1995) test to verify whether the (loss squared) difference between two forecast errors is statistically significant. The null hypothesis is that two alternative methods are - on average - equally accurate. We applied such a test over all pairs of forecasting methods.

The results are synthesized in Table 4, where $e_i = e_j$ is the null that the forecasting errors of methods *i* and *j* are equal. For horizon $h = 3$ the errors are all significantly different.

The poor performance of method (3), i.e. the $VECM$ model, is a surprising result. However, we note that the forecasts are computed from the reduced $VECM$ model, whose dimension was obtained through a general-to-specific search. A similar outcome was found by GG who highlighted the negative results of this selection procedure. In contrast with GG, we find that univariate methods perform better than multivariate ones. Different reasons might explain this finding. First, spatial dependence in the final period is weak and, in the last months, country-specific behaviours dominate the dynamic of each inflation rates. Second, the poolability condition (GG, 2004) that spatial influence is relatively uniform across regions can justify this result. In this case, the interrelationships between countries are less relevant.

Finally, while GG consider only simulated stationary series, our series are non stationary. The existence of cointegration relationships implies the estimation of a number of parameters that can increase forecasts uncertainty.

5. Conclusions

The aim of this article was to perform a comparison of the relative efficiency of different forecasting methods for aggregated data that are both temporally and spatially correlated. We extended the findings of GG by considering a more general and realistic non-stationary context, where co-integration constraints in time are allowed to exist. A concrete application on monthly inflation rate for Euro-zone economies is considered. The empirical analysis shows that there is not a method that dominates over the others.

Our results highlight a better forecast performance of the univariate methods with respect to the multivariate ones. It seems that the presence of non-stationarity and the estimation of the cointegration relationships increase significantly the uncertainty of the forecasts.

However the space-time model performs better than the traditional $VECM$ model, and the number of estimated parameters decreased dramatically from more than 250 to only 72.

In contrast with Bronars and Jansen (1987) we found that the weak spatial dependence in our forecast interval worsened the performance of the space-time AR model. In the central period of the sample, where the Moran's I test shows a higher spatial dependence, the fitting capability of the space model is significantly better.

The results obtained in our paper suggest further analysis. First of all a global stationarity (spatial and temporal) test should be performed and applied on the data.

Secondly, spatial analysis is made usually at a finer level of disaggregation, as NUTS2 for European regions. For these data the propagation on the space is more significant, that at country level. However in this case the order in the multivariate models will drastically increase and it is interesting to view the performance of the different methods. Arbia *et al.* (2011) present some simulation results over different scenarios.

Finally, Arbia *et al.* (2011) evidenced that standard MSE could be inappropriate to evaluate forecasts when aggregating with space-time series. In this case the outcome of the aggregation procedure is not merely a time series of data, but it is a new spacetime series. In these conditions a forecasting strategy has to be judged not only in terms of the standard MSE measure, but also in terms of the spatial characteristics of the forecasting errors. A forecasting method that provides an accurate estimate in terms of MSE may well be rejected if it provides forecasting errors that are concentrated in a systematic way in some definite portions of space, displaying a positive spatial correlation. The authors propose therefore to incorporate spatial analysis into forecasting evaluation criterions.

References

Altavilla, C., Ciccarelli, M. (2009).The Effects of Monetary Policy on Unemployment Dynamics under Model Uncertainty Evidence from the US and the Euro Area, *Journal of Money, Credit and Banking*, **41(7)**, 1265–1300.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, The Netherlands.

Arratibel, O., Kamps, C. and Leiner-Killinger, N. (2009). Inflation Forecasting in the new EU Member States, *Working Paper European Central Bank*, n. 1015.

Arbia, G., Bee, M. and Espa, G. (2011). Aggregation of Regional Economic Time Series with Different Spatial Correlation Structures, *Geographical Analysis*, **47**, 78–103.

Bermingham, C., DAgostino, A. (2011). Understanding and Forecasting Aggregate and Disaggregate Price Dynamics, *Working Paper European Central Bank*, n. 1365.

Bronars, S.G., Jansen, D.W. (1987). The Geographic Distribution of Unemployment rates in the U.S., *Journal of Econometrics*, **36**, 251–279.

Clements, M.P., Hendry, D. F. (1998). *Forecasting Economic Time Series*, Cambridge University Press, Cambridge.

Cliff, A.D., Ord, J.K. (1973). *Spatial Autocorrelation*, Pion, London.

Cliff ,A.D., Haggett, P., Ord, J.K., Bassett, K. and Davies, R. (1975). *Elements of Spatial Structure*, Cambridge University Press, Cambridge.

Corrado, L., Fingleton, B. (2012). Where Is The Economics In Spatial Econometrics?, *Journal of Regional Science*, **52(2)**, 210–239.

Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy, *Journal of Business and Economic Statistics*, **13**, 253–263.

Elhorst, J.P. (2001). Dynamic models in space and time, *Geographical Analysis*, **33**, 119–140.

Elhorst, J.P (2012). Dynamic spatial panels: models, methods and inference, *Journal of Geographical systems*, **14**, 5–28.

Garderen, K.J. van Lee, K. and Pesaran, M.H. (2000). Cross-sectional aggregation of non-linear models, *Journal of Econometrics*, **95**, 285–331.

Giacomini, R., Granger, C.W.J. (2004). Aggregation of SpaceTime Processes, *Journal of Econometrics*, **118**, 7–26.

Granger, C.W.J. (1969). Prediction with a generalized cost of error function, *Operations Research Quarterly*, **20**, 199–207.

Granger, C.W.J. (1987). Implications of aggregation with common factors, *Econometric Theory*, **3**, 208–222.

Granger, C.W.J., Newbold, P. (1977). *Forecasting Economic Time Series*, Academic Press, New York.

Griffith, D. A., Lagona, F. (1998). On the Quality of Likeliood-Based Estimators in Spatial Autoregressive Models When the Data Dependence Structure Is Misspecified, *Journal of Statistical Planning and Inference*, **69**, 153–74.

Harris, R., Moffat ,J. and Kravtsova, V. (2011). In Search of W, *Spatial Economic Analysis*, **6(3)**, 249–270.

Hubrich, K. (2005). Forecasting euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy? *International Journal of Forecasting*, **21**, 119–136.

Le Gallo, J., Dallerba, S. (2006). Evaluating the Temporal and Spatial Heterogeneity of the European Convergence Process 1980-1999, *Journal of Regional Science*, **46(2)**, 269–288.

Lütkepohl, H. (1984). Linear transformations of vector ARMA processes, *Journal of Econometrics*, **26**,283–293.

Lütkepohl, H. (1987), *Forecasting aggregated vector ARMA processes*, Springer, Heidelberg.

Lütkepohl, H. (2006). Forecasting with VARMA models, in: Elliott G., Granger C.W.J., Timmermann A. (eds), Handbook of Economic Forecasting, 1, Elsevier, Amsterdam, 287–325.

Pfeifer, P.E., Deutsch, S.J. (1980). A three-stage iterative procedure for space-time modeling, *Technometrics*, **22**, 35–47.

Szulc, E. (2000). Modelling the space time structure of the economic process on the example of unemployment, in: Zielinski Z. (ed.), *Dynamic Econometric Models*, **4**, Torun, Poland.

Wei, W.W.S., Abraham, B. (1981). Forecasting contemporal time series aggregates, *Communications in StatisticsTheory and Methods*, **A 10**, 1335–1344.