Quaderni di STATISTICA

Liguori Editore

Volume 14 - 2012

Quaderni di STATISTICA

VOLUME 14 - 2012

LIGUORI EDITORE

Volume 14, anno 2012

ISSN 1594-3739 (printed edition)

Registration n. 5264, 6 December 2001, Court of Justice at Naples
ISBN-13 **978 - 88 - 207 - 5853-0**

Director: Gennaro Piccolo

– Ministero delle Politiche Agricole, Alimentari e Forestali (gestione ex Centro per la formazione in economia e politica dello sviluppo rurale)
– Dipartimento di Scienze Economiche e Statistiche, Università di Salerno
– Dipartimento TEOMESUS, Università di Napoli Federico II
– MIUR, PRIN 2008, Dipartimento di Scienze Statistiche, Università di Bologna
– MIUR, PRIN 2008 CUP E61J100000200001, 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

Foreword

This special issue of "Quaderni di Statistica" is dedicated to the topics discussed during the International Conference on "Methods and Models for Latent Variables" MMLV012, held in Naples on 17-19 May, 2012. It comprises papers which were peer-reviewed and selected for publication.

The Conference was initiated as part of the dissemination activities linked to the PRIN 2008 research project on: "Latent structure analysis: new boundaries in statistical methods and models", which was coordinated by professor Paola Monari and carried out by the research units working at the University of Bologna, Naples Federico II, Padua and Florence. The project was financed by MIUR.

The aim of the Conference was to provide insights on latent structure analysis and modelling and to give the opportunity to participants to meet and network.

The Conference included plenary sessions with the invited contributions of eminent statisticians worldwide: A. Agresti (Department of Statistics, University of Florida, USA), N. Balakrishnan (Department of Mathematics and Statistics, McMaster University, Canada), G. Celeux (Department de Mathématiques, Université Paris-Sud, France), J. Chen (Department of Statistics, University of British Columbia, Canada), F. Pesarin, (University of Padua, Italy), E. Ronchetti (Research Center for Statistics and Department of Economics, University of Geneva, Switzerland), Y. Wang (Department of Statistics and Applied Probability, University of California, USA).

I wish to express my sincere thanks to each of the authors for their papers and I hope that this issue of "Quaderni di Statistica" will stimulate further interest and attract more researchers to the many important and interesting problems that are still to be solved.

The Editor

Quaderni di Statistica
Vol. 14, 2012

Clustering of categorical data: a comparison of different approaches

Laura Anderlucci
Department of Statistical Sciences, University of Bologna
E-mail: laura.anderlucci3@unibo.it

Christian Hennig
Department of Statistical Science, University College London
E-mail: chrish@stats.ucl.ac.uk

Summary: In clustering, one may be interested in the classification of similar objects into groups, and one may be interested in finding observations that come from the same true homogeneous distribution. But do both of these aims lead to the same clustering? And how good are clustering methods designed to fulfil one of these aims in terms of the other? In order to address this, two approaches, namely a latent class model (mixture of multinomial distributions) and a partition around medoids one, are evaluated and compared by Adjusted Rand Index (which is expected to favour model-based clustering) and Average Silhouette Width index (which is expected to favour distance-based clustering) in a fairly wide simulation study. Visualization of the clustering outcomes is obtained with a special use of Multidimensional Scaling.

Keywords: Latent Class Analysis, Partition Around Medoids, Clustering.

1. Introduction

There are different ways to do cluster analysis of categorical data in the literature. Choice among them is strongly related to the aim of the researcher. Main approaches for clustering can be partitioned into model-based and distance-based methods: the former assume that objects belonging to the same class are similar in the sense that they come from the same homogeneous probability distribution, of which the parameters are unknown and need to be estimated; the latter evaluate distances among objects by a defined dissimilarity measure and, basing on it, allocate units to the closest group. Since clustering is defined as the classification of similar objects into groups, we wonder whether

observations assumed to come from the same probability distribution are as similar as observations in the same cluster would be if a distance-based method were applied and how good the latter methods are in finding the 'true' clustering, if it exists. In order to address this, two approaches, namely a latent class model (mixture of multinomial distributions) and a partition around medoids one, are evaluated and compared by Adjusted Rand Index (ARI, which is expected to favour model-base clustering) and Average Silhouette Width index (ASW, which is expected to favour distance-based clustering) in a fairly wide simulation study.

2. Latent class clustering

Goodman (1974) shows a relatively simple method for determining whether the observed relationships among the variables in a p-way contingency table can be explained by a K-class latent structure and for evaluating the local identifiability of the model parameters.

Starting from a p-way contingency table which cross-classifies a sample of n individuals with respect to p manifest polytomous variables, if there is some latent polytomous variable (with K categories), so that each of the n individuals is in one of the K latent classes with respect to this variable, and within the kth latent class, the manifest variables are mutually independent, then this K-class latent structure would serve as a simple explanation of the observed relationships among the variables. Here is the model:

$$f(x) = \sum_{k=1}^{K} p_k f(x, a_k) \tag{1}$$

with $\sum_{k=1}^{K} p_k = 1, f(x, a_k) = \prod_{j=1}^{p} \prod_{l=1}^{m_j} (a_k^{jl})^{x^{jl}}$ where $a_k = (a_k^{jl},\ l = 1, \ldots, m_j,\ j = 1, \ldots, p)$, given $\sum_{l=1}^{m_j} a_k^{jl} = 1$. $f(x, a_k)$ is the density function of a multivariate multinomial distribution for independent variables with parameters a_k.

Latent Class Analysis yields a probabilistic clustering approach. This means that although each object is assumed to belong to one class or cluster, it is taken into account that there is uncertainty about an object's class membership. An individual's posterior class-membership probabilities are computed from the estimated model parameters and his observed scores; units are thus assigned to the class with the highest posterior probability.

3. Partition around medoids

In the real world, a well defined true clustering often does not exist and the aim may rather be putting similar observations together. For this reason we chose to consider PAM which serves this aim.

The PAM algorithm (developed by L. Kaufman and P. J. Rousseeuw) is very similar to k-means; the procedure minimises, for a given dissimilarity measure d, the objective function

$$g(\mathbf{x}_1^*, \ldots, \mathbf{x}_K^*) = \sum_{i=1}^{n} \min_{h \in \{1,\ldots,K\}} d(\mathbf{x}_i, \mathbf{x}_h^*) \tag{2}$$

by the choice of K medoids $\mathbf{x}_1^*, \ldots, \mathbf{x}_K^*$ from $\mathbf{x}$. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers. Here, we used the *Manhattan distance* (or L_1 *distance*) as a dissimilarity measure, with categories coded by zero-one variables.

4. Simulation study: model-based vs. distance-based approaches

In order to investigate the similarity of results in terms of classification, a systematic simulation study has been carried out. In particular, we have examined the impact of the following aspects on the clustering:

- no. of latent classes (small/large);
- no. of observed variables (4/12) and no. of their categories (2/4/8);
- no. of units for each data set (small samples/big samples);
- entity of mixing proportions (extremely different/equal);
- expected cluster separation (clear/unclear).

From the combination of all these specific features we obtain 128 settings, called 'patterns'. For each pattern 2000 different data sets have been generated with the Latent Gold (LG) software and the true classification of units was recorded. Then we estimated the model according to a model-based approach with the same (commercial) software and with a distance-based method (using `pam` function, contained in the R-package `cluster`, dissimilarity measure = *manhattan*). We also estimated the model, again according to a maximum likelihood approach, with an open-source software (using an EM algorithm, implemented as a function `lcmixed` in the R-package `fpc`), with the aim of comparing results, precision and time with LG. We finally compared the obtained classifications according to different indexes that respectively measure the ability of recovering the true classification (ARI) and similarity/dissimilarity of observations (ASW).

The results of the simulation study show that performances (in terms of quality of clustering) of the two approaches highly depend on data features, even though the direction of the dependence is not always very clear. We wondered whether these characteristics may significantly affect the differences between the two approaches and, in this case, which are their directions. Hence, in order to improve our understanding of the problem and to individuate these determinants, we performed an analysis of variance on

the differences between the indexes we calculated for the LG and the PAM clustering outcomes.

Visualization of the clustering results is obtained with a special use of Multidimensional Scaling.

5. Conclusions

Overall, the simulations show that, in terms of recovering the 'true' clustering (according to a 'true' unknown model), the Latent Class Clustering generally behaves better, yielding better results in terms of ARI, even when the clusters are supposed to overlap. When clusters are expected to be separated, then a PAM approach would not make the results worse. PAM performances improve when the mixing proportions of the components of the mixture that generated the data are about the same, i.e. when the clusters have about the same size. The factors that are more important for distinguishing the methods are the number of latent classes, the sample size and the variation of the mixing proportion.

Performance of the two methods in terms of retrieving homogeneous groups is more difficult to evaluate and more considerations are needed. No method outperforms the other one always, so it is not easy to make general statements. What is more surprising is that LG, by trying to put together observations assuming they come from the same distribution, accomplished to get similar observations together and to separate objects that are very different in a way that is not much worse than a distance-based method (as PAM) usually does, and actually sometimes LG works even better. Of course this does not mean that PAM should not be used anymore, since there are still situations in which it works better than LG. Finally, both LG and PAM obtained values of ASW higher than the true clustering; of course this does not mean that they are better than the 'truth', but sometimes observations coming from different groups are more similar to each other than objects in the same class. The amount of quality they 'lose' by not finding the true class membership is regained in terms of similarity/dissimilarity, which is a good tradeoff.

References

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

Kaufman L., Rouseeuw P.J. (1990), *Finding groups in data*, Wiley, New York.

Hubert L., Arabie P. (1985), Comparing Partitions, *Journal of Classification*, 2, 193–218.

Quaderni di Statistica
Vol. 14, 2012

Multidimensional extensions of IRT models and their application to customer satisfaction evaluation

Federico Andreis Pier Alda Ferrari
Department of Economics, Business and Statistics - Università degli Studi di Milano

E-mail: pieralda.ferrari@unimi.it, federico.andreis@unimi.it

Summary: Multidimensional IRT models (MIRTM), developed in the fields of psychometrics and ability assessment, are here considered in connection with the problem of evaluating customer satisfaction. Different models, that allow us to take into account more complex and, possibly, more realistic latent constructs than those usually assumed, are presented and discussed. Eventually, these models are applied to a real dataset, MCMC techniques for the estimation are implemented and analogies and differences with results from previous analyses on the same survey in the literature are discussed.

Keywords: Binary responses, Compensatory models, MIRT.

1. Introduction

It has become increasingly clear in the last decades that, in complex settings where unobservable (latent) quantities are of interest, the usual IRT assumption of a single underlying component influencing the observable outcomes might be not realistic [5, Chap.3]. Multidimensional IRT models (MIRTM) arise from the fields of psychometrics and ability assessment (as do their ancestors, the unidimensional IRT models, UIRTM), their aim being to overcome this limitation. The rationale behind these techniques is to provide an instrument capable of describing the usually not trivial apparatus of skills that a person brings to a test, in order to obtain a diagnostic tool about several subscales simultaneously and a way to model the interaction between examinees and test items' characteristics. It is important to point out, though, that this approach is not intended to provide a measurement of the latent trait in the sense Rasch introduced [4] (i.e. objective measurement), rather than to investigate from a modeling point of view the complexity of such unobservable phenomenon, thus following the statistical logic of model tun-

ing in order to fit data, i.e. choosing the best tool available. Such distinction has been highlighted since the beginnings: the Rasch model is a theoretical ideal, a definition of measurement, as opposed to statistical modeling, which is a toolbox we reckon being useful to try and disentangle not trivial intertwinings within the underlying (often very articulated) reality. Along these lines, our interest is into evaluating the possibility of employing MIRTM in a field they haven't previously been applied to, specifically customer satisfaction, assessing interpretability of models' parameters and comparing this approach to current methodology.

2. Multidimensional IRT models

Different approaches to multi-dimensionality have been carried out in the literature, leading to the definition of two main classes of models, the 'compensatory' and 'non (or partially) compensatory' ones [5, Chap. 4] and to the introduction of the concepts of 'between-items' and 'within-items' dimensionality [1]. Compensatory models allow, through proper parameterization, latent traits' effects to compensate for each other, e.g. a high level of ability on one dimension can make up for a low level on another - think of an additive functional form for the parameters; non-compensatory (even though it would be more correct to say 'partially' compensatory) models do not admit such strong compensation - think of a multiplicative form. Between-items and within-items dimensionality embody assumptions regarding how latent traits are represented by items in a questionnaire, i.e. if each item is related to one, and only one, of the latent traits (between-items), or is to be linked to more than one at the same time (within-items).

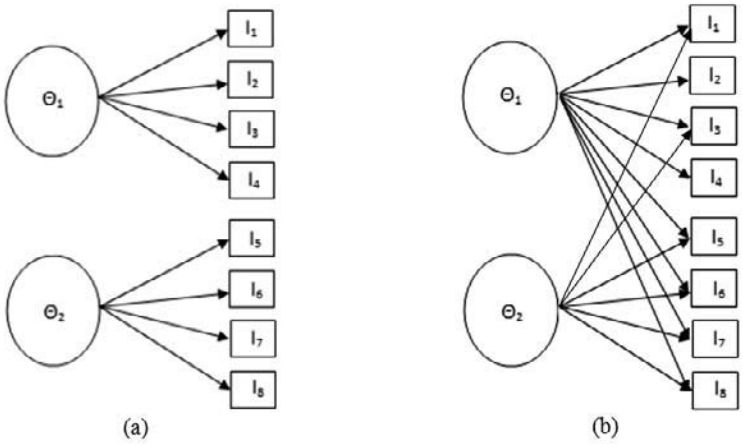


Figure 1. Between-items (a) and within-items (b) dimensionality with two latent traits.

Multidimensional extensions of classic IRT models have been introduced; they permit to work with dichotomous or polytomous test responses, but also to include covariates [5, Chap. 4].
Since MIRTM require a greater number of parameters to be estimated than unidimensional IRT models, estimation issues arise and are addressed. In the literature, MCMC methods are advocated as a helpful tool to obtain accurate results, but implementation of the algorithms and assessment of convergence to the desired posterior distribution require careful assessment [6].

3. Customer satisfaction assessment with MIRTM

Recently, unidimensional IRT models have been applied to the field of customer satisfaction (CS) evaluation, via a convenient re-interpretation of the role of their parameters [2]. Since it is legitimate to think about the existence of more than one latent factor in CS-related investigations too, the aim of this work is to evaluate the possibility of identifying and applying suitable MIRTM to this context.
In order to better appreciate the meaning and additional contribution of the extension to more than one latent trait, we first review the basic one-dimensional dichotomous model, whose expression is given by:

$$P(X_{ij} = x|\theta_i, \beta_j) = \frac{e^{x(\theta_i - \beta_j)}}{1 + e^{\theta_i - \beta_j}} \tag{1}$$

where $x = 0, 1$ if the customer is, respectively, unsatisied or satisfied, θ_i is *satisfaction* for the $i-$th respondent and $(-\beta_j)$ the quality of the $j-$th item. This model is algebraically equivalent to that introduced by Rasch [4] in order to evaluate ability tests, θ_i being the person ability and β_j the item difficulty.

The multidimensional extension we consider is called *multidimensional 2 parameter logistic* (M2PL) model, and has the following expression:

$$P(X_{ij} = x|\boldsymbol{\theta}_i, \mathbf{a}_j, d_j) = \frac{e^{x(\mathbf{a}_j'\boldsymbol{\theta}_i + d_j)}}{1 + e^{\mathbf{a}_j'\boldsymbol{\theta}_i + d_j}} \tag{2}$$

where both $\mathbf{a}_j$ and $\boldsymbol{\theta}_i$ are $m-$dimensional vectors, so that the sum in the exponent of e can be rewritten as $\mathbf{a}_j'\boldsymbol{\theta}_i + d_j = \sum_{t=1}^{m} a_{jt}\theta_{it} + d_j$. This model assumes that m latent traits characterize the satisfaction, with single scores θ_{it} whose relevance is weighted by the a_{jt} parameters, known as items' discrimination related to the $t-$th dimension, and here intended to describe the relevance of the $j-$th item to the $t-$th trait. The M2PL belongs to the class of compensatory models, since it adopts a parameterization that is a linear combination of the satisfaction parameters (one for each presumed dimension of the latent construct) for each individual (customer/user), i.e. a high satisfaction value θ_{it} on one latent dimension can compensate for a low θ_{is} on another.

The a_{jt} parameters may be estimated from the data or be fixed according to particular assumptions. If all are assumed to be equal to a fixed value $a_{jt} = a^*$, then $\mathbf{a}_j'\theta_i + d_j = a^* \sum_{t=1}^m \theta_{it} + d_j$, and we assume that the discriminating power of all the items is the same, also across all dimensions; this means that different items have different quality, but the same relevance to each trait. If $a^* = 1$, we obtain the initial model, with the positions $\theta_i = \sum_{t=1}^m \theta_{it}, d_j = -\beta_j$, thus obtaining a mere reparameterization of the θ_i parameters from the unidimensional model, not taking into account dimensionality, i.e. collapsing the model back to one latent trait only. Another possibility is to fix the a_{jt} either to be zero or different from zero: this way, it becomes possible to embody assumptions about between- or within-items dimensionality. If the researcher believes a certain item (or group of items) to be dependent on one latent trait only, for example the $s-$th, this can be implemented in the model by letting $a_{jt} = 0, t \neq s$, and a_{is} either to be estimated or fixed to some non-zero value; this is the structure in Fig.1 - (a).

4. Application

Eventually, the models we discuss are applied to real data, presented in [3, Chap. 2]; the dataset consists of responses (on a 1-5 Likert scale, that we dichotomize for our purposes) to a questionnaire concerning satisfaction for a large firm's services. Analogies and differences with previous IRT-oriented analyses on the same survey [3, Chap. 14] are investigated. The results are then compared and discussed.

References

Adams R.J., Wilson M., Wang W. (1997), The Multidimensional Random Coefficients Multinomial Logit Model, *Applied Psychological Measurement*, 21, 1-23 .

De Battisti F., Nicolini G., Salini S. (2005), The Rasch Model to Measure Service Quality, *The ICFAI Journal of Services Marketing*, III-3, 58-80.

Kennet R.S., Salini S. (2012), *Modern Analysis of Customer Satisfaction Surveys: with applications using R*, Wiley, New York.

Rasch G. (1960), *Probabilistic models for some intelligence and attainment tests*, Danish Institute for Educational Research, Copenhagen.

Reckase M.D. (2009), *Multidimensional Item Response Theory*, Springer, New York.

Sinharay S. (2004), Experiences With Markov Chain Monte Carlo Convergence Assessment in Two Psychometric Examples, *Journal of Educational and Behavioral Statistics*, 29, 461-488 .

Quaderni di Statistica
Vol. 14, 2012

Advances on inferential methods for heterogeneity comparisons

Rosa Arboretti
Department of Territory and AgriForestal Systems, University of Padua
E-mail: rosa.arboretti@unipd.it

Stefano Bonnini
Department of Economics and Management, University of Ferrara
E-mail: stefano.bonnini@unife.it

Summary: The purpose of the present work consists in the study of a non-parametric procedure for a two sample test for heterogeneity comparisons in presence of categorical variables. In particular a comparison between the approximated permutation solution proposed by Arboretti et al. (2009) with a new proposal based on a bootstrap resampling method is performed. Some results of the simulation study for analyzing the performances of the compared methods are shown. An application related to a customer satisfaction survey is also illustrated.

Keywords: Heterogeneity, Permutation test, Categorical Variables.

1. Introduction

A typical distributional aspect for categorical variables is represented by the so called heterogeneity. If the probability distribution $p_1, \dots, p_K$ of a categorical variable that can take K different categories were known, an index for measuring the degree of heterogeneity should satisfy the following properties: (a) it reaches its minimum when $p_k = 1$ for a given k and $p_h = 0\ \forall h \neq k$; (b) it assumes greater values if there are categories k and h such that $p_k > p_h$, all other probabilities remaining unchanged, then $Het(\dots, p_k, \dots, p_h, \dots) \leq Het(\dots, p_k - \epsilon, \dots, p_h + \epsilon, \dots)$ whenever $0 < \epsilon \leq (p_k - p_h)/2$; (c) it reaches its maximum in case of uniform distribution: $p_k = 1/K, k = 1, \dots, K$. Some of the most commonly used indexes for measuring heterogeneity are Gini's index defined as $G = \sum_k p_k(1 - p_k)$, Shannon's entropy

$S = -\sum_k p_k log(p_k)$, the generalized index proposed by Renyi $R_\theta = \frac{1}{1-\theta} log(\sum_k p_k^\theta)$. From the inferential point of view the problem under study consists in comparing the heterogeneity of two or more populations. An approximate nonparametric solution to this problem has been discussed within the permutation context in Arboretti et al. (2009). Let us suppose that two categorical variables $X_j, j = 1, 2$ can assume K different categories and denote with $P_j = \{p_{j1}, \ldots, p_{jK}\}$ the underlying distributions. The problem under study can be expressed as $H_0 : Het(P_1) = Het(P_2)$ against the alternative $H_1 : Het(P_1) > (\neq) Het(P_2)$, where $Het(P_j)$ denote the heterogeneity of the related distribution. The parameters p_{jk} are unknown but they can be estimated with the relative sample frequencies $\widehat{p}_{jk} = f_{jk}/n_j$. If the probabilities were known they could be arranged in non-increasing order within each population, $p_{j(1)}, \ldots, p_{j(K)}$, and we could say that the two populations present the same heterogeneity degree when $p_{1(k)} = p_{2(k)} \forall k \in \{1, \ldots, K\}$. Hence, when the two populations are equally heterogeneous, the data of two samples for which the respective categories are arranged according to the rule of the *Pareto diagram*, i.e. in non-increasing order respect to the corresponding probabilities, the data are exchangeable between samples and so the permutation testing principle is exactly applicable (see Pesarin and Salmaso, 2010). Instead of unknown population parameters p_{jk} we can only use the sampling estimates $\widehat{p}_{jk}$ and obtain a *data driven ordering* within population, affected by sampling variability and correspondent to the true ordering only asymptotically (Arboretti et al., 2009).
The conditional Monte Carlo solution for the testing problem proposed by Arboretti et al. (2009) is based on the following steps: (i) Consider the data-set $\vec{X} = (\vec{X_1}, \vec{X_2})$ and calculate the corresponding $2 \times K$ contingency table with absolute frequencies $f_{jk}, j = 1, 2, k = 1, \ldots, K$; (ii) order the table according to decreasing frequencies $f_{j(k)}, j = 1, 2, k = 1, \ldots, K$; (iii) consider the transformation of the original variables $X_j, j = 1, 2$ into the new variables $Y_j, j = 1, 2$ according to the ordered table, i.e. such that $Y_j = k$ with absolute frequency $f_{j(k)}$; (iv) compute the observed value of a suitable test statistic based on the difference between the observed heterogeneity indexes of the compared samples, $T_{H,oss} = H(\vec{Y_1}) - H(\vec{Y_2})$; (v) consider B random permutations $\vec{Y_{(b)}}^* = (\vec{Y_{1,b}}^*, \vec{Y_{2,b}}^*), b = 1, \ldots, B$ of the transformed dataset $\vec{Y}$; (vi) for each permutation compute the corresponding value of the test statistic $T^*_{H,b} = H(\vec{Y^*_{1,b}}) - H(\vec{Y^*_{2,b}})$; (vii) compute the p-value of the test $\lambda_H = \sum_{b=1}^{B} I(T^*_{H,b} > T_{H,oss})/(B + 1)$. In the present work a similar procedure based on the bootstrap method, i.e. on the resampling of data with replacement instead of permutations, is proposed.

2. Simulation study

Let us consider a simulation study where data are generated as $X = 1 + Int[K \cdot U^\delta]$, where $\delta > 0$ is a real parameter, U a random variable uniformly distributed in the open interval (0, 1), and $Int[\cdot]$ denotes the integer part of $[\cdot]$. The support of X is discrete and consists in the first K positive integers. The situation of maximum heterogeneity

can be simulated by making $\delta = 1$. By increasing the values of δ the distribution of X moves away from maximum heterogeneity, toward that of maximum homogeneity (or equivalently minimum heterogeneity).
In Table 1 the estimated powers of the simulation study are shown. The test statistics based on the indexes of Gini (T_G), of Shannon (T_S), of Renyi when the parameter tends to ∞ (T_{R_∞}) and when the parameter is equal to 3 (T_{R_3}) are considered for both the permutation procedure proposed by Arboretti et al. (2009), denoted by *"perm"*, and the new bootstrap procedure, denoted by *"boot"*.

Table 1. Rejection rates, $K = 4$ *categories,* $n_1 = n_2 = 40$*,* $B = 1000$ *permutations,* $CMC = 1000$ *datasets.*

Test	$\delta_1 = 1.5, \delta_2 = 1.5$		$\delta_1 = 1.5, \delta_2 = 2.0$		$\delta_1 = 1.5, \delta_2 = 2.5$	
	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$
T_G perm	0.048	0.090	0.222	0.360	0.462	0.582
T_S perm	0.045	0.086	0.172	0.347	0.434	0.560
T_{R_∞} perm	0.048	0.104	0.219	0.366	0.451	0.582
T_{R_3} perm	0.053	0.097	0.238	0.379	0.456	0.599
T_G boot	0.033	0.084	0.191	0.368	0.440	0.593
T_S boot	0.028	0.074	0.151	0.326	0.407	0.538
T_{R_∞} boot	0.067	0.132	0.272	0.410	0.495	0.632
T_{R_3} boot	0.056	0.109	0.266	0.413	0.505	0.626

Under the null hypothesis of equality in heterogeneity, here represented by $\delta_1 = \delta_2 = 1.5$, almost all the tests respect the nominal α level hence they are well approximated and the exception is represented by the tests based on Renyi's indexes, with powers slightly greater than α. As expected power increases as the difference between heterogeneity in the two populations increases. The performances of the bootstrap tests are similar to the ones of permutation tests but the ones based on Renyi's indexes are the most powerful among all the considered testing procedures.

3. Real case study

To measure the satisfaction of the customers of a gym center in a small town in the province of Rovigo (Italy), a customer satisfaction survey was performed interviewing a sample of 77 people. The customers were classified according to the main goal, i.e. achievement either of *fitness* or of *health*. From Table 2 it emerges a generalized high satisfaction for both the groups and the heterogeneity test can be applied to compare the homogeneity of the responses of the groups. Formally we are interested in testing the inequality in heterogeneity (two-tailed-test) at the significance level $\alpha = 0.05$

Table 2. Observed contingency tables of the satisfaction levels of the customers of the gym center for 3 *specific aspects.*

	Goal achievement		Cleanliness		Price	
Satisfaction level	Fitness	Health	Fitness	Health	Fitness	Health
Very unsatisfied	1	0	1	0	0	1
Quite unsatisfied	1	1	1	1	1	1
Quite satisfied	22	3	20	3	18	6
Very satisfied	35	14	37	14	40	10

In Table 3 the p-values of the considered permutation and bootstrap tests are shown. It is evident that the null hypothesis of equality in heterogeneity should not be rejected at the specified significance level.

Table 3. P-values of the two sample two-tailed heterogeneity test.

	permutation tests				bootstrap tests			
Aspect	T_G	T_S	T_{R_∞}	T_{R_3}	T_G	T_S	T_{R_∞}	T_{R_3}
Goal Achievement	0.206	0.568	0.176	0.206	0.202	0.546	0.154	0.152
Cleanliness	0.270	0.682	0.232	0.270	0.258	0.590	0.218	0.220
Price	0.080	0.060	0.316	0.168	0.114	0.052	0.364	0.160

Acknowledgements: The authors would like to thank the University of Padua (CPDA092350/09) and the Italian Ministry of University and Research MIUR project PRIN2008 - CUP number C91J10000000000001 (2008WKHJPK/002) for providing the financial support for this research.

References

Arboretti Giancristofaro R., Bonnini S., Pesarin F. (2009), A permutation approach for testing heterogeneity in two sample categorical variables, *Statistics and Computing*, 19, 209–216.

Pesarin F., Salmaso L. (2010), *Permutation tests for complex data: theory, applications and software*, Wiley, Chichester.

Quaderni di Statistica
Vol. 14, 2012

Mixtures of equispaced Normal distributions and their use for testing symmetry in univariate data

Silvia Bacci
Department of Economics, Finance, and Statistics, University of Perugia
E-mail: silvia.bacci@stat.unipg.it

Francesco Bartolucci
Department of Economics, Finance, and Statistics, University of Perugia
E-mail: bart@stat.unipg.it

Summary: Given a random sample of observations, mixtures of Normal densities are often used to estimate the unknown continuous distribution from which the data come. Here we propose the use of this semiparametric framework for testing symmetry around an unknown value. The null hypothesis of symmetry may be formulated in terms of a Normal mixture model with equispaced support points and weights constrained to be equal one another around the centre of symmetry. The resulting model is nested in a more general unconstrained one, with the same number of mixture components and free weights. Therefore, symmetry may be tested against skewness through a likelihood ratio statistic. The behavior of the proposed mixture-based test is illustrated through a Monte Carlo simulation study, where we compare our test with the traditional one based on the third standardized moment.

Keywords: Density estimation, Expectation-Maximization algorithm, Skewness.

1. Introduction

Given a random sample of observations from a continuous distribution with density $f(x)$, a problem which may be useful to consider is testing for the symmetry of $f(x)$ about some unknown value. Several procedures have been proposed in the literature to deal with this issue: for a review see Hollander (2006).

Aim of the present paper is to propose the use of finite Normal mixture (NM) models (McLachlan and Peel, 2000) for testing symmetry. Indeed, since any continuous distribution - symmetric or skewed - can be approximated arbitrarily well by a finite mixture of Normal distributions with common variance, NM models provide a convenient semiparametric framework by which modeling unknown distributional shapes, while keeping (*i*) a parsimony close to that of full parametric methods, as represented by a single density, and (*ii*) the flexibility of nonparametric methods, as represented by the kernel method.

The paper is organized as follows. Section 2 illustrates the proposed test of symmetry and the NM model on which it is based. Section 3 describes the main results of the Monte Carlo simulation study aimed at comparing the performance of our test with respect to the traditional test of symmetry based on the third sample standardized moment.

2. Mixture-based test of symmetry

Let k be the number of Normal components of the mixture, let α be the centre of the symmetry, and let β be a scale parameter, such that the support points of the mixture are

$$\nu_j = \alpha + \beta\delta_j, \quad j = 1, \ldots, k,$$

where $\delta_1, \ldots, \delta_k$ is a grid of equispaced points between -1 and 1. Then, the density of a mixture of k Normal equispaced components (NM$_k$) is defined as

$$f(x) = \sum_{j=1}^{k} \pi_j \phi(x; \nu_j, \sigma^2),$$

where $\phi(x; \nu_j, \sigma^2)$ denotes the density at x of the distribution $N(\nu_j, \sigma^2)$.

The basic idea behind the proposed approach is that if the sample observations come from a symmetric distribution, then the weights of the mixture components equidistant from the centre of symmetry are equal to each other. Therefore, the hypothesis of symmetry may be formulated as

$$H_0 : \pi_j = \pi_{k-j+1}, \quad j = 1, \ldots, [k/2],$$

where $[z]$ is the largest integer less than or equal to z and k is fixed.

The NM$_k$ model including hypothesis H_0 is nested in the NM$_k$ model with unconstrained weights π_j. Consequently, for testing symmetry we may use a likelihood ratio test, based on the deviance $D = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)]$, where $\hat{\boldsymbol{\theta}}$ is the unconstrained maximum likelihood estimate of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_0$ is that under the constraint imposed by H_0. Both estimates may be computed by the EM algorithm (Dempster et al., 1977). Under H_0, D is asymptotically distributed as a Chi-square with a number of degrees of freedom equal to $[k/2]$.

3. Monte Carlo study

We studied the performance of the proposed test through a Monte Carlo simulation in which this test is compared with the traditional test of Gupta (1967) based on the third sample standardized moment. Our test is evaluated by selecting the number of mixture components (k) by using both Akaike and Bayesian Information Criteria (AIC and BIC, respectively).

Within the Monte Carlo study we simulated 1,000 samples with increasing size ($n = 20, 50, 100$) from the following distributions: standard Normal ($N(0,1)$), Student's t with 5 degrees of freedom (t_5), Laplace or double exponential (Lap), symmetric mixture of three Normal distributions (NM_3), Chi-square with 1 degree of freedom (χ^2_1), Chi-square with 5 degrees of freedom (χ^2_5), Chi-square with 10 degrees of freedom (χ^2_{10}), and Lognormal with mean 0 and variance equal to 1 ($logN$). All tests were performed for nominal levels α equal to $0.01, 0.05, 0.10$.

The comparison between the proposed mixture-based test and the Gupta's test is performed by taking into account the following optimality criteria: (*i*) the empirical type-I error probability must be not higher than the nominal significance level under distributions satisfying the null hypothesis of symmetry and (*ii*) the empirical power under skewed alternatives must be as high as possible. The simulation results, used to evaluate the test procedures in these terms, are reported in Table 1.

As concerns the empirical significance level, the mixture-based test we propose shows performance very similar to that of the Gupta's test when the number k of components is selected by means of BIC. On the contrary, when AIC is used for model selection, the chosen value of k is frequently higher than that chosen by BIC. We also observe that the test based on AIC to select k has an empirical significance level constantly higher than the nominal one: in other words, the type-I error is committed too often. On the other hand, this tendency of AIC of choosing a relatively high number of mixture components results in good performance of the mixture-based test under skewed distributions. In this case, the empirical power is clearly higher with respect to the variant based on BIC and, in particular, to the Gupta's test. However, even the variant of the mixture-based test based on BIC is almost always more powerful than the Gupta's test.

In conclusion, the simulation results show that the behavior of the proposed test depends on the criterion used to select the number of mixture components. In particular, when BIC is adopted, the empirical level of significance of our test is comparable with that of the traditional test, whereas the empirical power results usually higher. This test seems also preferable to that based on AIC to select the number of mixture components.

References

Dempster, J., 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.

Gupta, M. (1967), An asymptotically non parametric test of symmetry, *The Annals of Mathematical Statistics*, 38, 849–866.

Hollander, M. (2006), *Testing for symmetry*, Wiley, New York.

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

Table 1. Empirical significance/power level of the mixture test based on AIC, mixture test based on BIC, and Gupta's test at level of significance of $0.01, 0.05, 0.10$*. The results are based on 1,000 simulated samples of size* $n = 20, 50, 100$ *drawn from certain symmetric and skewed distributions.*

		Symmetric distributions				Skewed distributions			
	n	$N(0,1)$	t_5	Lap	NM_3	χ^2_1	χ^2_5	χ^2_{10}	$logN$
$\alpha = 0.01$									
Mixture test (AIC)	20	0.02	0.02	0.02	0.04	0.35	0.07	0.04	0.20
	50	0.02	0.02	0.02	0.02	0.76	0.50	0.25	0.56
	100	0.03	0.03	0.02	0.02	0.96	0.89	0.65	0.78
Mixture test (BIC)	20	0.01	0.00	0.01	0.03	0.23	0.03	0.02	0.12
	50	0.01	0.01	0.01	0.02	0.65	0.23	0.10	0.42
	100	0.00	0.01	0.01	0.02	0.93	0.61	0.28	0.70
Gupta's Test	20	0.00	0.00	0.00	0.00	0.08	0.02	0.01	0.05
	50	0.00	0.01	0.01	0.00	0.20	0.13	0.08	0.12
	100	0.01	0.00	0.01	0.01	0.33	0.45	0.30	0.15
$\alpha = 0.05$									
Mixture test (AIC)	20	0.06	0.06	0.07	0.09	0.57	0.23	0.14	0.42
	50	0.07	0.08	0.08	0.08	0.87	0.7	0.46	0.71
	100	0.08	0.08	0.10	0.06	0.98	0.95	0.79	0.88
Mixture test (BIC)	20	0.02	0.01	0.03	0.06	0.42	0.12	0.06	0.31
	50	0.01	0.01	0.03	0.06	0.83	0.34	0.15	0.65
	100	0.01	0.03	0.05	0.05	0.97	0.69	0.33	0.83
Gupta's Test	20	0.04	0.03	0.04	0.04	0.36	0.15	0.09	0.27
	50	0.04	0.03	0.04	0.05	0.50	0.54	0.37	0.34
	100	0.04	0.03	0.04	0.05	0.66	0.8	0.71	0.42
$\alpha = 0.10$									
Mixture test (AIC)	20	0.10	0.10	0.13	0.14	0.68	0.33	0.22	0.57
	50	0.09	0.13	0.15	0.13	0.92	0.77	0.53	0.81
	100	0.10	0.13	0.17	0.14	0.99	0.96	0.83	0.91
Mixture test (BIC)	20	0.03	0.03	0.06	0.10	0.56	0.18	0.09	0.45
	50	0.01	0.03	0.06	0.09	0.90	0.39	0.17	0.78
	100	0.01	0.05	0.08	0.11	0.98	0.72	0.34	0.89
Gupta's Test	20	0.10	0.08	0.11	0.09	0.53	0.34	0.21	0.45
	50	0.08	0.08	0.10	0.09	0.69	0.75	0.62	0.54
	100	0.08	0.08	0.09	0.09	0.84	0.92	0.88	0.61

Quaderni di Statistica
Vol. 14, 2012

Evaluation of the degree effect on the work path by a latent variable causal model

Francesco Bartolucci
Department of Economics, Finance and Statistics, University of Perugia
E-mail: bart@stat.unipg.it

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

Giorgio Vittadini
Department of Quantitative Methods for Business and Economic Sciences
E-mail: giorgio.vittadini@unimib.it

Summary: We formulate a causal model for the effect of the degree program on the work path of graduate students, which is based on a Markov process to represent latent characteristics of the subjects. A latent Markov model with covariates and mixed-type responses results, which is estimated by an EM algorithm. We illustrate the proposed approach through an application to a dataset deriving from administrative panel data which concern labor market in Lombardy and allows us to evaluate the effectiveness, in terms of education, of certain Universities in Milan.

Keywords: Causal Inference, Education, Labor Market, Latent Markov Model.

1. Introduction

In this paper, we propose a causal model to evaluate the effect of degree programs on the employment status in terms of income, easiness in switching between types of position, and employment skills of the graduates. The approach is motivated by a dataset deriving from the following databases: (*i*) database of the observatory of the Lombardy labor market, (*ii*) database of graduates from the largest universities in Milan, and (*iii*) database of the office of revenues.

The causal model is formulated following the approach of Heckman (2010) that, in a general context, shows the equivalence between econometric structural models and potential outcome models (Rubin, 1974). The model takes into account the longitudinal structure of the data through a Markov chain that has the role of representing individual characteristics which are not directly observable (latent characteristics). Moreover, it allows for individual covariates, which may directly affect the outcomes or the parameters of the Markov chain, and for response variables of different types, essentially categorical and continuous.

The model based on the above formulation is in practice a latent Markov model (Wiggins, 1973, Bartolucci *et al.*, 2010) with covariates and mixed-type responses. Therefore, from the methodological point of view our main contribution is that of casting this model in the causal literature providing an interpretation of its parameters in terms of causal effects that may be used for evaluation purposes. The approach based on this model may be then compared with recent causal inference approaches for longitudinal data; see, among others, Gill and Robins (2001), Abbring and Van Den Berg (2003), and Aalen *et al.* (2012).

2. The dataset

The dataset concerns 1258 young individuals who graduated in 2004 from the three main universities of Milan: Milano-Bicocca, Milano-Statale, and Cattolica del Sacro Cuore. They have been followed along 20 quarters after the graduation date and four quarters before, covering the years 2003-2008. The choice of the specific 2004 cohort is motivated by the availability of the data of the employment offices from 2004 to 2009. The response variables are: (*i*) annual income in Euro referred to the previous year, (*ii*) employment status, (*iii*) type of position indicating whether a subject is employed with a temporary or permanent contract, and (*iv*) job quality, measured by the skill level of the job (low, medium, or high). The last one is derived by a categorization of the job qualification made by the Italian National Institute of Statistics. The available covariates concern individual characteristics such as: (*i*) gender, (*ii*) age, (*iii*) number of family components, (*iv*) family income, (*v*) place of birth, (*vi*) type of high school, (*vii*) employement during the graduate studies, (*viii*) place of graduation, (*ix*) type of degree program (scientific, humanistic, or social science and business), (*x*) examination grades, (*xi*) final grade at college.

In Table 1 we report descriptive statistics for the distribution of some of the available covariates, whereas in Table 2 we report the descriptive statistics for the response variables for each year of observation.

The main research question concerns the evaluation of the degree program effect on the labor market transitions of the graduates. A question of this type may be addressed in the framework of causal inference by the model for longitudinal data that we illustrate in the following.

Table 1. Descriptive statistics for the distribution of the covariates

Covariate		%	mean	st.dev.
Male		39.11		
Age in 2004			25.59	1.82
Degree grade			101.91	7.28
Family component in 2004			3.19	1.14
Degree type	*Scientific*	15.66		
	Humanistic	25.36		
	Social S. - Bus.	58.98		
Empl. before 2004	*Part-time*	15.02		
	Full-time	5.64		

Table 2. Frequency of every response variable for each year of observation

Year	Average	Temporary	Skill	
	Income	cont.%	high %	medium %
2004	6890	53.88	49.33	46.67
2005	11100	54.01	59.17	38.06
2006	15170	50.55	64.47	33.76
2007	18700	48.40	63.94	34.57
2008	20040	42.86	52.44	28.35

3. The causal model

With reference to a subject in the panel, let $\boldsymbol{Y}^{(t)}$ denote the vector the response variables of interest at time occasion t, $t = 1, \ldots, T$, and $\boldsymbol{X}^{(t)}$ denote the corresponding vector of covariates. The proposed model assumes the existence of a latent process $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(T)})$ which affects the distribution of the response variables. The main assumption of the model is that the vectors $\boldsymbol{Y}^{(t)}$, $t = 1, \ldots, T$, are conditionally independent given the latent process and the covariate vectors $\boldsymbol{X}^{(t)}$, $t = 1, \ldots, T$. In such a context the latent variables summarize the observed outcomes. The latent process is assumed to follow a first-order Markov chain with state space $\{1, \ldots, k\}$, where k is the number of latent states. Note that, contrary to other latent Markov formulations, we do not assume that the response variables in $\boldsymbol{Y}^{(t)}$ are conditionally independent given the corresponding latent variable $U^{(t)}$, but, as in Bartolucci and Farcomeni (2009), these variables may be dependent.

We admit that certain covariates affect the distribution of the response variables given the latent process, whereas other covariates directly affect the response variables. Among these covariates we include dummies for the type of degree and for the University where the degree was earned. We show that the regression coefficients for these dummy variables have a causal interpretation in the sense of Heckman (2010). More-

over, among the covariates we include the lagged response variables, so that we account for state dependence. This means accounting for the effect of having a certain income and employment skill at a given occasion on the probability of having the same income and skill at the following occasion, once observable covariates and subject specific unobservable factors are taken into account.

Maximum likelihood estimation of the model parameters is carried out by the EM algorithm. As usual, this algorithm alternates two steps until convergence in the likelihood. At the first step (E-step), the posterior probability of every latent state and pair of consecutive latent states is computed for every subject in the sample. At the M-step, the expected value of the complete log-likelihood, computed on the basis of these posterior probabilities, is maximized by standard rules.

Acknowledgements: We are grateful to Prof. M. Mezzanzanica and Dr. M. Fontana, CRISP, University of Milano-Bicocca, for providing the dataset.

References

Aalen O. O., Røysland K., Gran J. M., Ledergerber B. (2012), Causality, mediation and time: a dynamic viewpoint, *Journal of the Royal Statistical Society, Series A*, DOI: 10.1111/j.1467-985X.2011.01030.x.

Abbring J., Van Den Berg G. (2003), The nonparametric identification of treatment effects in duration models, *Econometrica*, 71, 1491–1517.

Bartolucci F., Farcomeni, A. (2009), A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure, *Journal of the American Statistical Association*, 104, 816–831.

Bartolucci F., Farcomeni A., Pennoni, F. (2010), An overview of latent Markov models for longitudinal categorical data, http://arxiv.org/abs/1003.2804.

Gill R.D., Robins, J. M. (2001), Causal inference for complex longitudinal data: the continuous case, *Annals of Statistics*, 29, 1785–1811.

Heckman J. J. (2010), Building bridges between structural and program evaluation approaches to evaluating policy, *Journal of Economic Literature*, 46, 356–398.

Rubin D.B. (1974), Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, 66, 688–701.

Wiggins L. M. (1973), *Panel Analysis: Latent probability models for attitude and behaviour processes*, Elsevier, Amsterdam.

Quaderni di Statistica
Vol. 14, 2012

Longitudinal patterns of financial product ownership: a latent growth approach

Francesca Bassi
Department of Statistics, University of Padua
E-mail: francesca.bassi@unipd.it

José G. Dias
ISCTE – Lisbon University Institute
E-mail: jose.dias@iscte.pt

Summary: The main goal of this study is to analyze the dynamic process of financial product ownership of Italian households under the assumption of multiple latent trajectories of growth. Using panel data from the Survey on Income and Wealth conducted by the Bank of Italy, we conclude that the trajectory of Italian households in terms of financial product ownership in the period 2000 to 2008 can be represented by two latent factors describing two specific behaviours: state bonds ownership and all other types of financial assets ownership. We also find that this behaviour is significantly influenced by the area of the country where the family lives and other characteristics of the head of the household, such as labour condition and education.

Keywords: Panel data, Latent growth factor, Intra-individual change.

1. Introduction

Latent growth models (LGM) consider both intra-individual change and inter-individual differences in such change by estimating the amount of variation across individuals in the latent growth factors (random intercept and slope) as well as the average growth (Bollen and Curran, 2006). The assumption of unidimensionality in the growth process – all manifest variables measured by the same latent factor – is not always realistic. If multidimensionality exits and is ignored, statistical results maybe seriously biased. Thus, multidimensional latent growth modelling allows specifying multiple latent trajectories at the same time with varying parameters and inter-relations between the processes.

This paper illustrates the enormous potential of this type of longitudinal latent variable modelling. The application estimates the evolution of financial product ownership at household level in Italy in the period 2000 to 2008. We model the binary indicators of ownership (i.e., whether a given household owns a certain type of financial asset) as multiple indicators of a latent process.

2. The latent growth model

Let us define the structure of the data being modelled. Each household i at a given time t may hold or not the financial product j. Thus, y_{ijt} is a binary variable where 0 is the case when the household does not hold the financial product and 1 otherwise – $i = 1, \ldots, n$; $j = 1, \ldots, J$; $t = 1, \ldots, T$. Let y^*_{ijt} be a continuous score underlying y_{ijt}, i.e., by defining a threshold ν_{jt} it turns out that $y_{ijt} = 0$, for $y^*_{ijt} \leq \nu_{jt}$, and $y_{ijt} = 1$, otherwise. Let h_{ijt} be the latent variable or score of household financial ownership at time t in dimension k. Thus, for each household and time point, it results a k-dimensional factor-item response model for each indicator: $y^*_{ijt} = \lambda_{jtk} h_{itk} + \varepsilon_{ijt}$, where the intercept is omitted, given the threshold, the factor loadings are λ_{jtk}, a latent variable h_{ijt}, and a specific residual ε_{ijt}. The growth model requires measurement invariance of the factors across time, i.e., the thresholds and factor loadings of the indicators are equal over time: λ_{jk} and ν_j, respectively. Moreover, for scaling identification, the first loading is set to 1.

The LGM (Meredith and Tisak, 1990) is defined by the latent process $h_{itk} = h^I_{ik} + (t-1)h^S_{ik} + \varepsilon_{itk}$ that gives the trajectory of the household i financial ownership at time t, where h^I_{ik} and h^S_{ik} are the intercept and slope of the process, respectively, ε_{itk} is the error term.

For the conditional model, the intercept and the slope of each factor are function of the observed variables contained in the vector $\mathbf{x}_i$ (Salgueiro et al., 2011): $h^I_i = \alpha^I_0 + \alpha^I_1 \mathbf{x}_i + \zeta^I_i$; $h^S_i = \alpha^S_0 + \alpha^S_1 \mathbf{x}_i + \zeta^S_i$, where h^I_i and h^S_i are continuous latent variables, α^I_0 and α^S_0 are constant, and α^S_0 and α^S_1 contain the coefficients of the covariates in the conditional model. Model parameters are estimated via maximum likelihood using the EM algorithm.

3. The empirical study

The Bank of Italy (BI) has been running the Survey of Household Income and Wealth (SHIW) since 1965. With a few exceptions, the survey was conducted on a two-year basis since then. SHIW provides information on income, savings, consumption expenditure and the real wealth of Italian households, as well as on household composition and labour force participation. In 1989, BI introduced a longitudinal component in the survey, adopting a peculiar split panel design: at each survey wave, the sample con-

sists of two sections: a panel sub-sample, made up of households who participated in the previous wave; and a fresh cross-sectional sub-sample. In this paper we consider the sub-sample of 1684 households interviewed in the waves from 2000 to 2008 and ownership of 13 different assets[1].

Area of the country and education and professional status of the head of the household are added as time constant covariates (observed in 2000), making the modelling of intercept and slope of the latent processes conditional. In 2000, 45.96% of the households were from the North of Italy (NORTH), and 21.28% and 32.76% were from the Centre and the South (SOUTH), respectively. Moreover 16.6% of heads of household were workmen, 15.9% clerks (CLERK), 3.3% managers (MANAGER), 12.2% were self-employed (SELF) and 52.0% not working (NOTWORK). The distribution by education was: 5.5% did not attend school, 28.7% elementary school (ELEM), 29.1% middle school (MED), 28.2% high school (HIGH) and 8.5% university or higher (UNIV). Estimated loadings (Table 1) for the two latent processes show that all indicators are positively associated with the two factors and lead to interpret factors 1 and 2 as ownership of state bonds and ownership of all other types of financial products, respectively. The

Table 1. Estimated loadings

	Factor 1			Factor 2	
	Estimate	s.e.	Estimate	s.e.	
BOT	1.00	–	SHA	1.00	–
CCT	3.03	0.12	DEP	2.25	0.82
BTP	2.76	0.14	PCD	1.06	0.19
			PCT	1.95	0.30
			BFP	0.56	0.08
			OBB	1.95	0.37
			QFC	2.20	0.44
			PGP	1.24	0.26
			PTE	2.25	0.33
			COO	0.90	0.17

main focus of the analysis is on the structural part of the model, i.e., the latent trajectories measured by the indicators and explained by the covariates (Table 2). For factor 1 – ownership of state bonds, we conclude that: (i) the intercept (expected value in 2000 assuming null covariates) is only significantly influenced by SOUTH, i.e., households from the southern area of Italy are expected to have an initial score of 0.35 lower than those living in the central area of the country; (ii) the slope is significantly different for households living in the North (0.08) compared to those living in the Centre; (iii) middle and high school, and university degree categories have also a significant and positive impact on the slope of the trajectory of owning state bonds compared to the lowest level

[1] Ordinary Treasury bills - BOT, repos - CCT, multi-annual Treasury bills - BTP, shares - SHA, deposits - DEP, certificates of deposits - PCD, Treasury credit certificates - PCT postal bonds - POB, bonds - OBB, mutual funds - QFC, assets under management - PGP, foreign securities - PTE, loans to cooperative - COO.

of education. For factor 2 – owning all other types of financial products, we conclude that: (i) the expected value in 2000 is significantly higher in the North (0.11), and for high school and university degree; (ii) the slope has a significantly negative constant, but being self-employed has a significant positive impact on this ownership behavior; being from the South reduces even more the slope; (iii) all educational dummy variables in the model increase significantly the slope of the trajectory.

Table 2. Estimated parameters – structural conditional model

	Factor 1				Factor 2			
	Intercept		Slope		Intercept		Slope	
	Est.	s.e.	Est.	s.e.	Est.	s.e.	Est.	s.e.
Intercept	0	–	−0.32	0.18	0	–	−0.30	0.09
CLERK	0.12	0.67	0.05	0.20	0.13	0.09	0.03	0.01
MANAGER	0.01	32.91	0.02	0.61	0.08	0.19	0.03	0.07
SELF	0.01	0.65	0.05	0.21	0.19	0.12	0.05	0.02
NOTWORK	0.16	0.61	0.06	0.19	−0.04	0.05	−0.01	0.03
NORTH	0.03	0.05	0.08	0.02	0.11	0.08	0.01	0.02
SOUTH	−0.35	0.08	−0.10	0.07	−0.80	0.36	−0.09	0.01
ELEM	−0.11	0.21	0.05	0.05	0.14	0.12	0.18	0.06
MED	−0.06	0.21	0.12	0.06	0.15	0.09	0.24	0.06
HIGH	0.04	0.29	0.14	0.04	0.39	0.15	0.28	0.07
UNIV	0.31	0.33	0.22	0.06	0.67	0.14	0.28	0.08

4. Conclusions

The paper applies a latent growth model to analyze the recent dynamics of financial product ownership behavior of Italian families. The findings of this study show that one needs two latent factors to summarize this behavior: state bonds ownership and other financial products ownership. These two factors evolve with different trajectories in time and are significantly influenced by family characteristics: area of the country where the family lives and labor condition and education of the head of household.

References

Bollen K.A., Curran, P.J. (2006), *Latent Curve Models: A Structural Equation Approach*, Wiley, New York.

Meredith W., Tisak, J. (1990), Latent curve analysis, *Pshycometrika*, 55, 107-122.

Salguiero M.F., Smith P.W., Vieira M.D.T (2011), A multi-process second-order latent growth curve model for subjective well-being, *Quality and Quantity*, doi: 10.1007/s11135-011-9541.

Quaderni di Statistica
Vol. 14, 2012

Bayesian machine learning approaches for longitudinal latent class modelling to define wheezing phenotypes to elucidate environmental associates

Danielle Belgrave Angela Simpson Iain Buchan Adnan Custovic
The University of Manchester
E-mail: danielle.belgrave@manchester.ac.uk, angela.simpson@manchester.ac.uk, buchan@manchester.ac.uk, adnan.custovic@manchester.ac.uk

Christopher Bishop
Microsoft Research Cambridge
E-mail: chistopher.bishop@microsoft.com

Summary: Accurate phenotypic definition of wheezing in childhood can lead to a greater understanding of the distinct physiological markers associated with different wheeze phenotypes. This paper looks at Bayesian machine learning approaches using Infer.NET to define wheeze phenotypes based on both parental questionnaires and General Practitioner data on patterns of asthma and wheeze consultation within the first eight years of life. We illustrate a taxonomy of longitudinal latent class item response models with varying modelling assumptions to determine wheeze phenotypes (latent classes) for homogenous groups of children.

Keywords: Longitudinal latent class analysis, Bayesian inference, Infer.NET.

1. Introduction

It has been widely recognised that asthma is a heterogeneous disease. Accurate phenotypic definition of wheezing in childhood can elucidate our understanding of this underlying complexity of asthma to identify distinct physiological markers associated with different wheeze phenotypes. We used longitudinal latent class modelling to identify subpopulations (classes) of children who differ in patterns of wheeze trajectories

during childhood, based on both complete medical records and parental assessment of wheeze at different time points. We tested the validity of these classes by examining their relations with measures of lung physiology and atopy.

2. Methods

2.1. Data Description

Data are taken from the Manchester Asthma and Allergy Study (MAAS), an unselected, prospective population-based birth cohort study designed to determine early life factors for the development of asthma and allergic disease. Participants attended follow-up at ages 1, 3, 5 and 8 years. Validated questionnaires were interviewer-administered at each time-point to collect information on parentally-reported symptoms and physician-diagnosed asthma. Current wheeze was defined as parentally-reported wheeze in the past 12 months. Additionally, a trained paediatrician extracted data from electronic and paper-based primary care medical records, including wheeze/asthma diagnosis, and hospital admissions for asthma/wheeze during the first 8 years of life; we calculated the child's age in days for each event. We analysed data from questions assessing the presence of wheeze for 1185 children using these two complementary measures of wheeze.

2.2. Statistical Methods

We analysed all data using Bayesian machine learning approach using Infer.NET (research.microsoft.com/en-us/um/cambridge/projects/infernet/). The Bayesian machine learning method provides a unified framework for modelling and quantifying uncertainty - employing probabilistic modelling strategies based on defining priors in such a way that probabilities can be associated with unknown parameters. This allows us to incorporate and compare different modelling assumptions with a greater degree of flexibility. The three steps for defining a model in Infer.NET are: i) the definition of a probabilistic model; ii) the creation of an inference engine for performing inference; and iii) the execution of an inference query. We used Variational Message Passing (VMP) approximation for inference. VMP gives a factorised approximation for the model distribution.

We modelled a longitudinal latent class item response model to determine latent classes for homogenous groups of children. We analysed data from our two complementary sources: parentally-reported current wheeze at four follow-ups, and physician-confirmed wheeze recorded in medical records within each year from birth to age 8 years. We assumed that each child belongs to one of N latent classes, with the number and size of classes not known a priori. We assumed a discrete trajectory logistic regression model with a class dependent random intercept and slope for the dichotomous variable representing the answer to the question "Has the child wheezed within the given

time period?". This is represented in Equation (1) for child i at time t with wheeze assessed by rater r (physician or parent). Children were assigned to the latent class with the largest posterior probability given a uniform Dirichlet prior.

$$Logit\{Pr(y_{itr}) = 1|x, class_i = k\} = \begin{array}{l}\beta_0 + \beta_1 t + \beta_2[rater] \\ +\beta_3[time \times rater] + \xi_{class} + \xi_{class} t\end{array} \quad (1)$$

We then extended this model by adding a quadratic term to the level-1 linear change trajectory model so that Equation (1) becomes:

$$Logit\{Pr(y_{itr}) = 1|x, class_i = k\} = \begin{array}{l}\beta_0 + \beta_1 t + \beta_2[rater] + \beta_3[time \times rater] \\ +\xi_{class} + \xi_{class} t + \xi_{class} t^2\end{array} \quad (2)$$

This model allows us to assess a higher order of complexity for the latent class trajectory.

We also considered a Hidden Markov Model which takes into account the sequential patterns of wheeze and correlations between observations that are close together. We used wheeze assessment from both physician and parental data to infer a multinomial latent variable for each child in order to assign children to their most probable class. We inferred time-dependent transition probabilities for the eight time points. Children belonging to the same latent class are similar with respect to their transition probabilities which are assumed to come from the same probability distributions, whose parameters are, however, unknown quantities to be estimated.

These modelling approaches with varying assumed number of classes were compared for goodness-of-fit using model evidence. The model evidence evaluates the probability of generating the data from a model whose parameters are sampled from the prior distribution while penalising the model according to its complexity. Classes identified in the optimal model were validated to see whether these patterns of wheeze represented a higher risk of various markers of asthma severity.

3. Results

Based on the model evidence, the optimal model was the discrete trajectory model with class dependent random Intercept and random linear slope with five latent classes of wheeze. Based on our interpretation of the results, we have labelled these classes as "No Wheeze" (53.3%), "Transient Early Wheeze" (13.7%), "Late-onset Wheeze"(16.7%), "Persistent Troublesome Wheeze" (3.2%) and "Persistent Controlled Wheeze" (13.1%) (Figure 1). Mixed effects models revealed significant differences in these 5 wheeze phenotypes."Persistent Troublesome Wheeze", "Late-onset Wheeze" and "Persistent Controlled Wheeze" were found to be associated with atopy ($p<0.001$). There was a significant difference in initial Specific Airway Resistance (sRaw) values for the five classes

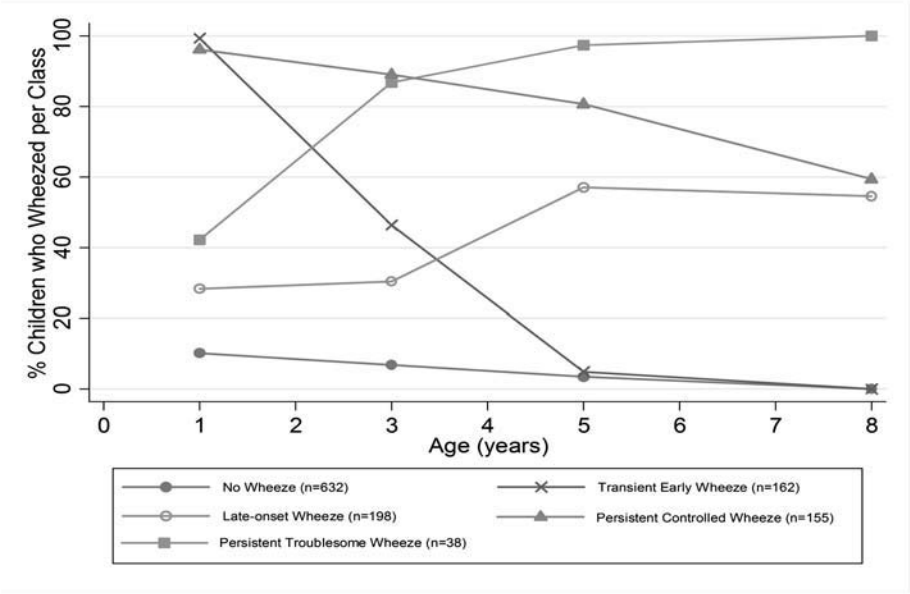


Figure 1. Percentage of children with reported wheezing according to either parental and physician reporting. The number of children in each class is denoted in parentheses.

($p<0.001$). Children with early onset wheeze and persistent controlled wheeze show significant increase in sRaw over time ($p=0.03$ and 0.02 respectively).

4. Conclusion

The joint modelling of observations from the general practitioner and parental reporting of childhood wheeze enables us to identify phenotypes of wheeze with greater accuracy and determine their risk factors and characteristics.

References

Bishop C.M. (2006), *Pattern Recognition and Machine Learning*, Springer, New York.

Minka T., Winn J.M., Guiver J.P., Knowles D.A. (2010), *Infer.NET 2.4* Microsoft Research Cambridge, http://research.microsoft.com/infernet (2nd ed.).

Quaderni di Statistica
Vol. 14, 2012

Mixed response surface models and Bayesian analysis of variance components for Electrically Conductive Adhesives

Rossella Berni
Department of Statistics, University of Florence
E-mail: berni@ds.unifi.it

Valeria L. Scarano, Francesco Bertocci, Marcantonio Catelani
Department of Electronics and Telecommunications, University of Florence
E-mail: valeria.scarano@unifi.it, francesco.bertocci@unifi.it, marcantonio.catelani@unifi.it

Summary: Electrically Conductive Adhesive (ECA) is a material consisting of epoxy resin filled with silver particles. In our specific case we focused on an isotropic material which is conductive in all directions. This presentation deals with an analysis of random effects and variance components for micro-electronics data. By considering the technical challenges related to the resistance of Electrically Conductive Adhesives (ECAs), two adhesives were examined. Random effects are involved in a Response Surface Methodology setting and the results are compared with a Bayesian approach.

Keywords: Variance components, Bayesian analysis, Electrically Conductive Adhesives.

1. Introduction

Since 80's and 90's, in quality and technology fields, Response Surface Methodology (RSM) and experimental designs played a relevant role aimed at improving some specific issues such as: design, modelling (Nelder and Lee, 1991), process optimization (Del Castillo, 2007). By considering the modelling features, many approaches are introduced in order to extend: i) the analysis of variance model; ii) the concept of fixed effect especially when evaluating blocks or noise variables; iii) the relevance of the heteroschedasticity for the error variance. Undoubtedly, when considering source of

variabilities, the connection of RSM with the inclusion of random effects is a notably improvement to study those components, i.e. variance components, which may influence the main and operative variables of the production process, (Khuri, 2006). In this regard, this presentation (Berni et al., 2011; Berni, et al., 2011) is related to the analysis of Electrically Conductive Adhesives (ECAs) through linear mixed models and variance components, (Searle et al., 1992). Thus, we deals with: i) the analysis of experimental data for two ECAs, by exploring the relevance of variance components and random effects due to the presence of the block factor for the two different glues; ii) the comparison of a RSM approach with random effects (Khuri, 2006) with a Bayesian approach (Wolfinger and Kass, 2000).

2. Specimen production and characterization

We focused on an isotropic material which is conductive in all directions (ICA). A one-component adhesive, Heraeus, and a two-component adhesive, Epotek, were examined. In order to study the electrical performance of these materials, specimens were taken according to a specific geometry. The response variable (Y) is the electrical resistance [Ω]. The planned experimental design is a mixed-level fractional factorial design 2_V^{5-1}, with two added center points ($n_0 = 2$) for each block. The type of adhesive is evaluated at two levels as a block factor; independent variables are shown in Table 1. Further details related to specimen production and experimental planning are reported in Catelani et al., (2011).

Table 1. Independent variables and levels

Variable	symbol	original levels
Radial velocity [round-per-minutes] (rpm)	x_1	4000-8000
Curing temp. $[^\circ C]$	x_2	90-120-150
Spin-coating time $[s]$	x_3	10-20-30

3. ECAs through response surface models and mixed effects

In this paper our efforts are aimed at improving the analysis of ECAs through random effects and variance components. This situation has been widely studied in literature, such as in Khuri (2006). We shall consider a response surface model for k variables: $x_1, .., x_j, ..x_k$, defined for the u-th experimental run:

$$y_u = \beta_0 + f'(x_u)\beta + z'_u\gamma + f(x_u)\Lambda z'_u + \epsilon_u; \quad u = 1, .., n \tag{1}$$

Table 2. GLS estimates for fixed effects of model (2)

Coefficient	estimate	st.err.	p-value
β_0	1.86	13.67	0.91
β_1	11.85	11.73	0.49
β_2	-25.61	9.49	0.01
β_3	9.29	9.49	0.33
β_{12}	-22.35	9.49	0.02
β_{22}	24.23	12.73	0.06

where β_0 is the intercept; β is the vector $(\beta_1, ..\beta_p)$ of unknown parameters, F is the extended matrix of dimension $[n \times p]$ formed by the n linear independent functions $f(x_u)$; ϵ_u belongs to the column vector $[n \times 1]$ of random errors. Furthermore, the term $z'_u\gamma$ denotes the presence of an additional sub-experimental factor, such as a block factor at b levels; thus, z_u is a vector of binary variables (values: 0,1), where 1 is equal to the presence of that observation in that block, $\gamma : (\gamma_1, .., \gamma_b)$ is the vector of coefficients related to b blocks. The Λ array is formed by the interaction parameters among blocks and process variables. The estimation (Generalized Least Squares-GLS) of model (1) is carried out through PROC MIXED (SAS Software) on data described in Section 2 (n=36). Thus, the model applied with k=3 process variables (Table 1) and one block factor z, at two levels, is the following:

$$y_u(x,z) = \beta_0 + \beta_1 x_{u1} + \beta_2 x_{u2} + \beta_3 x_{u3} + \beta_{12} x_{u1} x_{u2} + \beta_{22} x_{u2}^2 + \sum_{l=1}^{L=2} \gamma_{1l} z_{ul} + x_{u1} \sum_{l=1}^{L=2} \gamma_{1l} z_{ul} \tag{2}$$

In this case-study only one interaction effect between the process variable "radial velocity" and block is taken into account and it is considered as a random effect. In Table 2 GLS estimates for the fixed effects are shown.

4. A Bayesian approach to estimate variance components for ECAs

The applied algorithm is suggested by Wolfinger and Kass (2000) and it is valid for small and/or unbalanced data-sets. The starting point is the consideration of a general linear mixed-effect model

$$y = X\beta + Z\gamma + \epsilon \tag{3}$$

The two arrays, related to the variance-covariance estimates for random effects (G) and random error (R), are supposedly independent and diagonal. We shall define θ as the vector of variance components $(\theta_1, .., \theta_T)$. Thus, the general joint posterior density is

defined as:

$$p(\beta, \gamma, \theta|y) = p(\beta, \gamma|\theta, y)p(\theta|y) \tag{4}$$

The algorithm accepted 99% of candidates. In Table 3 the results from the posterior sample are shown; mean and interval confidence for the three components: block, block*x_1 and the error.

Table 3. REML estimates and Summary statistics: mean of the Kernel density estimates-Percentiles of the posterior sample of variance components

Parameter	estimate	mean	2.5th Pct.	97.5th Pct.
σ^2-block	5571.10	2304.00	18.16	12483.00
σ^2-block*x_1	4403.30	1364.00	44.29	7217.00
σ^2-error	2087.60	1595.00	1089.00	2321.00

References

Berni R., Scarano V.L., Bertocci F., Catelani M., (2011), Variance components and Bayesian analysis for Electrically Conductive Adhesives, poster presented at the *Workshop on "Statistical Methods Applied to Microelectronics"*, Catholic University of the Sacred Heart, June 2011, Milan.

Berni R., Scarano V.L., Bertocci F., Catelani M., (2011), Mixed response surface models and Bayesian analysis of variance components for Electrically Conductive Adhesives, *submitted*.

Catelani M., Scarano V.L., Bertocci F., Berni R., (2011), Optimization of the soldering process with ECAs in electronic equipment: characterization measurement and experimental design, *IEEE Transactions on components, packaging and manufacturing technology*, 1, 1616–1626.

Del Castillo E., (2007), *Process optimization*, Springer-Verlag, New York.

Khuri A.I., (2006), Mixed response surface models with heterogeneous within-block error variances, *Technometrics*, 48, 206–218.

Nelder J.A., Lee Y. (1991), Generalized linear models for the analysis of Taguchi-type experiments, *Applied Stochastic Model and Data Analysis*, 7, 107–120.

Searle S.R., Casella G., McCulloch C.E. (1992),*Variance components*, John Wiley & Sons, New Jersey.

Wolfinger R.D., Kass R.E., (2000), Nonconjugate Bayesian analysis of variance components models, *Biometrics*, 56, 768–774.

Quaderni di Statistica
Vol. 14, 2012

Generalizing the Rasch model to account for omitting behavior

Lucio Bertoli-Barsotti
Dipartimento di Matematica, Statistica, Informatica e Applicazioni, Università di Bergamo (Italy)
E-mail: lucio.bertoli-barsotti@unibg.it

Antonio Punzo
Dipartimento di Economia e Impresa, Università di Catania (Italy)
E-mail: antonio.punzo@unict.it

Summary: A two-dimensional IRT model is introduced for binary data in the presence of omitted responses. Its parameterization has the advantage to satisfy the following conditions: firstly, the tendency to omit depends on a person latent trait which is distinct from the usual latent ability; secondly, the model belongs to the Rasch family of models, guaranteeing the existence of non-trivial sufficient statistics for the parameters. An application to a real dataset is illustrated.

Keywords: Rasch model, Nonignorable missing data, Multidimensional IRT.

1. General remarks

Within the item response theory (IRT) framework, the "tendency-to-omit" items may be interpreted as a latent variable closely related to cognitive ability. Omitting tendency may also be due to inability to understand, or unwillingness to respond for embarrassment, anger, discomfort, fatigue, or other reason that may, or may not, depend on other latent traits to be measured, as well as to other manifest variables (person factors). As it is well known, in these cases missingness is not generally ignorable for likelihood based inference (Little & Rubin, 2002). In this paper we will consider the specific circumstance in which: 1) a person is administered an item, 2) has time to consider it, but 3) decides, for whatever reason, to skip the item. We will refer to this circumstance as omitting tendency, or omitting behavior.

A well-known IRT "model-based approach" to this specific measurement problem allows missing values to be included into the analysis and then, in particular, it allows information about the tendency to omit to be inferred from nonresponse. This issue has been discussed by Knott, Albanese and Galbraith (1990), Knott and Tzamourani (1997), Moustaki and O'Muircheartaigh (2002). Following these ideas, more recently a unified approach to a wide class of models has been proposed in a seminal paper by Holman and Glas (2005). For a recent comparative study of alternative models within the family of Holman and Glas (2005), see Rose, von Davier and Xu (2010). All the models in this literature are instances of multidimensional IRT (MIRT, Reckase, 2009) models; unfortunately, by construction, they do not belong to the exponential family of distributions. In this paper we present the RRM, a two-dimensional generalization of the simple Rasch Model (RM) for the treatment of binary data in presence of omitting behavior. The model is suited for intentional omissions and belongs to the exponential family of distributions. A case study is considered to illustrate the effectiveness of the proposed model.

2. The basic model

In this article, we will refer to a binary (0/1) outcome variable. Let us consider the events $A =$ omitted response, $B =$ non-omitted response in category "0" and $C =$ non-omitted response in category "1". Further, let π_A, π_B and π_C, the probabilities of the events A, B and C, respectively. The RRM may be defined as

$$\ln \frac{\pi_B}{\pi_A} = \theta_{AB} - \delta_{AB} \quad \text{and} \quad \ln \frac{\pi_C}{\pi_B} = \theta_{BC} - \delta_{BC}. \tag{1}$$

Then, the RRM is an IRT model that is a function of a two-dimensional latent trait, say $(\theta_{AB}, \theta_{BC})'$. It is straightforward to note that – while person and item parameters θ_{BC} and δ_{BC} are interpretable as "ability" and "difficulty" parameters of the outcome variable – the item parameter δ_{AB} indicates the relative "difficulty" of choosing option B rather than A, and the person parameter θ_{AB} the tendency to respond in category 0 rather than A. Note that, by definition, $\theta_{AB} = -\theta_{BA}$, $\theta_{AC} = \theta_{AB} + \theta_{BC}$. Then, the response probability function can be expressed explicitly as $\pi_A = [1 + \exp(\theta_{AB} - \delta_{AB}) + \exp(\theta_{AC} - \delta_{AC})]^{-1}$, $\pi_B = \pi_A \exp(\theta_{AB} - \delta_{AB})$ and $\pi_C = \pi_A \exp(\theta_{AC} - \delta_{AC})$. This model can also be reparameterized to provide possibly more meaningful parameters. For example, one may write $\theta_{AB} = \theta_1 - \theta_2$, $\theta_{AC} = \theta_1 + \theta_2$, $\delta_{AB} = \delta_1 - \delta_2$, $\delta_{AC} = \delta_1 + \delta_2$, where $\theta_2 = 2\theta_{BC}$ represents the (rescaled) ability, while θ_1 represents the tendency to respond, regardless if the chosen option is π_B or π_C. As it can be seen from the expression of π_A, the RRM belongs to the class of models for which the probability to omit item is allowed to depend on both the dimensions of the latent trait.

3. Attitudes to abortion data set

The following analysis focuses on a survey measuring attitude to abortion. The considered 7 items, administered to 410 subjects, was part of an investigation of British social attitudes, held in 1986. For each item, respondents were asked if the law should allow abortion. A number of respondents failed to answer to the items, either completely or in part. The seven items were: 1) the woman decides on her own she does not wish to have the child; 2) the couple agree they do not wish to have the child; 3) the woman is not married and does not wish to marry the man; 4) the couple cannot afford any more children; 5) there is a strong change of a defect in the baby; 6) the woman's health is seriously endangered by the pregnancy; 7) the woman became pregnant as a result of rape. Approval was coded as C, disapproval as B, omitted response as A. This data set was analyzed by Knott et al. (1990) by means of the model

$$\ln \frac{\pi_B + \pi_C}{\pi_A} = a\xi + b\theta - \beta \quad \text{and} \quad \ln \frac{\pi_C}{\pi_B} = c\theta - \delta, \tag{2}$$

where a, b and c are item discrimination parameters. Note that, as for the case of the RRM, model (2) allows the tendency to omit items to depend on both the dimensions, ξ and θ, of the latent trait. As it can be seen, the dimension θ summarizes attitude to abortion, while the trait ξ is related to the tendency to express an opinion. The distinction between omitted and non-omitted responses is governed by both the parameters ξ and θ.

Under the Marginal Maximum Likelihood (MML) approach, we assume that the respondents are sampled at random from a $N_2(\mathbf{0}, \mathbf{\Sigma})$, where the means for both the dimensions are set to zero as an identification constraint, while the covariance matrix $\mathbf{\Sigma}$ has to be estimated. An interesting consequence of treating respondents as random is that the correlation between the latent traits can be incorporated within the model and estimated as an unknown parameter. In order to estimate the unknown parameters, we used ConQuest (Wu et al., 2007). The item parameter estimates, along with their standard errors and fit statistics, are given in Table 1. As it can be seen, ConQuest pro-

Table 1. MML item parameter estimates for the RRM

Item	$\widehat{\theta}_{AB}$	Std Err	MNSQ	T	$\widehat{\theta}_{BC}$	Std Err	MNSQ	T
1	-7.42	0.45	0.96	-0.10	0.85	0.16	0.99	-0.10
2	-6.36	0.44	1.02	0.01	-0.99	0.16	0.97	-0.30
3	-5.16	0.36	0.96	-0.10	-1.45	0.16	0.84	-1.80
4	-5.43	0.37	0.92	-0.30	-1.21	0.16	1.06	0.70
5	-2.63	0.45	1.02	0.01	-6.45	0.26	1.05	0.30
6	-1.69	0.52	0.87	-0.40	-8.30	0.35	1.08	0.40
7	-2.90	0.52	0.88	-0.30	-6.95	0.28	0.99	0.00

duces the mean squared (MNSQ) fit statistic for every estimated parameter, which is based on a standardized comparison between expected and observed scores. When the

model fits the data, the MNSQ statistics have a unitary expected value. These statistics are transformed by ConQuest to approximate normal deviates, and are denoted by T. The software also provides a 95% confidence interval (CI) for the expected value of the MNSQ. If the MNSQ fit statistic lies outside the CI, then the corresponding T statistic will have an absolute value that roughly exceeds 2 (see Wu et al., 2007, p. 23). It is apparent that all the item fit statistics are good. Expected *a posteriori* (EAP) ability estimates $\widehat{\theta}_{BC}$ for the respondents are also given by the software. Their Pearson correlation with the EAP-estimates obtained by Knott et al. (1990) is 0.945.

References

Holman R., Glas C. A.W. (2005), Modeling non-ignorable missing-data mechanism with item response theory models, *British Journal of Mathematical and Statistical Psychology*, 58, 1–17.

Knott M., Tzamourani P. (1997), Fitting a latent trait model for missing observations to racial prejudice data, in Rost J., Langeheine R. (eds.), *Applications of latent trait and latent class models in the social sciences*, Waxmann, Munster, Germany, 244–252.

Knott M., Albanese M., Galbraith J. (1990), Scoring attitudes to abortion, *The Statistician*, 40, 217–223.

Little R.J.A., Rubin D.B. (2002), *Statistical analysis with missing data*, New York, NY: Wiley.

Moustaki I., O'Muircheartaigh C. (2002), Locating "Don't Know", "No Answer" and Middle Alternatives on an Attitude Scale: A latent Variable Approach, in G.A. Marcoulides and I. Moustaki (Eds.), Latent Variable and Latent Structure Models, London: Lawrence Erlbaum Associates, 15–40.

Reckase M.D. (2009), *Multidimensional Item Response Theory*, Springer, New York.

Rose N., von Davier M., Xu X. (2010), Modeling nonignorable missing data with item response theory (IRT), Technical report, ETS, Princeton, New Jersey.

Wu M. L., Adams R. J., Wilson M. R., Haldane S. A. (2007), ACER ConQuest: Generalised Item Response Modeling Software - Version 2.0, ACER Press edition.

Quaderni di Statistica
Vol. 14, 2012

Measuring and analyzing performance in longitudinal data

Annamaria Bianchi Silvia Biffignandi
Department of Mathematics, Statistics, Computing and Applications
University of Bergamo
E-mail: annamaria.bianchi@unibg.it, silvia.biffignandi@unibg.it

Summary: The aim of this research is the modelling of longitudinal data, with specific reference to the analysis of business performance and factors influencing it.
From the methodological point of view, we consider an innovative approach for the measurement of performance in panel data, based on the measure introduced by Kokic et al. (1997). This measure is derived from an M-quantile regression (Breckling and Chambers, 1998) based on assumed production function. We extend this measure to panel data. With reference to the analysis of factors influencing performance, both static and dynamic error component regression models are considered. The empirical analysis is carried out using the Kauffman Firm Survey data.

Keywords: Panel data modelling, Mixed effects models, Performance measure.

1. Performance Evaluation

In this section we show how the performance measure introduced by Kokic et al. (1997) can be extended to panel data. It is a production function approach that relies on the idea that a firm transforms inputs into outputs and the more outputs that a business can produce with fewer inputs the more successful it is. Appealing characteristics of this performance measure are a) outlier robustness, b) the fact that it does not depend on the level of inputs of the business, unlike many other performance measures commonly used (see Fried et al., 2009), and c) its stochastic nature, which allows statistical inference.

Paper supported by the ex 60% 2010 University of Bergamo, Biffignandi grant and PAADEL project (Lombardy Region - University of Bergamo joint project). The authors are thankful to the Kauffman Foundation for providing support. Any opinions, finding, and conclusions or recommendations expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Ewing Marion Kauffman Foundation.

Suppose that the value of the output from a certain business can be measured by a variable y taking real values and the values of the inputs are given by a set of k-dimensional variables x. Further, suppose that a random sample of size N is drawn from the population. For each cross section unit we observe data on the same set of variables for T time periods (x_{it}, y_{it}), $i = 1, \ldots, N$, $t = 1, \ldots, T$. For a given $0 < q < 1$, and under linearity assumptions, the M-quantile of y given x is defined as

$$MQ_q(y_{it}|x_{it}) = x_{it}^T \beta_q, \quad i = 1, \ldots, N, \ t = 1, \ldots, T,$$

where x_{it} contains an intercept term. The coefficient β_q solves

$$E\left[\psi_q\left(\frac{y_{it} - x_{it}^T\beta}{\sigma}\right)x_{it}\right] = 0,$$

σ denoting the standard deviation of y in the population, and $\psi_q(u) = \psi(u)\{qI(u \geq 0) + (1-q)I(u < 0)\}$, with ψ given by the Huber Proposal 2 influence function $\psi(u) = uI(|u| \leq c) + c\text{sgn}(u)I(|u| > c)$. Equivalently, we can write the model in error form

$$y_{it} = x_{it}^T\beta_q + u_{it}(q), \quad i = 1, \ldots, N, \ t = 1, \ldots, T,$$

where, by construction, $MQ_q(u_{it}(q)|x_{it}) = 0$. A natural estimator of β_q is the pooled M-quantile regression estimator $\hat{\beta}_q$, which solves

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\psi_q\left(\frac{y_{it} - x_{it}^T\beta}{s}\right)x_{it} = 0,$$

where s is a suitable robust estimator of scale, such as the median absolute deviation. By applying cross-sectional results (Bianchi and Salvati, 2011), it can be shown that, under suitable assumptions, the M-quantile coefficient estimator is consistent and asymptotically normal with variance covariance matrix $\sigma W_q^{-1} V_q W_q^{-1}$, where $W_q = E(\psi'_{iq} x_i x_i^T)$ and $V_q = E(\psi_{iq}^2 x_i x_i^T)$, with $\psi'_{iq} = \psi'_q(u_{it}(q))$ and $\psi_{iq} = \psi_q(u_{it}(q))$. The asymptotic results are for large N and small T. A fully robust estimator for the asymptotic variance-covariance matrix is $s\hat{W}_q^{-1}\hat{V}_q\hat{W}_q^{-1}/N$, where

$$\hat{W}_q = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\psi}'_{qit}x_{it}x_{it}^T, \quad \text{and} \quad \hat{V}_q = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\psi}_{qit}^2 x_{it}x_{it}^T,$$

with $\hat{\psi}'_{iq} = \psi'_q(\hat{u}_{it}(q))$ and $\hat{\psi}_{iq} = \psi_q(\hat{u}_{it}(q))$, $\hat{u}_{it}(q) = y_{it} - x_{it}^T\hat{\beta}_q$.

Similarly to the case of cross-sectional data, starting from M-quantile regression for panel data, we shall construct a performance measure. To this end, notice that M-quantile regression leads to a family of hyperplanes indexed by the value of the corresponding quantile coefficient q. For each value of q in $(0, 1)$, the corresponding model $x_{it}^T\beta_q$ shows how the M-quantile of order q of the conditional distribution of y given x

varies with x. For large values of q the M-quantile surface describes the 'average' output of relatively efficiently performing businesses, and for small q it describes the 'average' output of relatively inefficiently performing businesses. For a certain business with input x_{it} and output y_{it}, it is therefore natural to define an index of performance measure as the value of the M-quantile surface passing through the observation (x_{it}, y_{it}). Taking the above mentioned characteristics into account, we define the index of performance measure (IPM) for unit i at time t as $IPM_{it} = q$ if $y_{it} = x_{it}^T \beta_q$. The estimator of the performance measure is defined correspondingly.
This index allows meaningful comparisons across the years. In future research we intend to include latent variables in the M-quantile model used for defining the index. This will allow to control for unobserved cross-section heterogeneity, improving the β estimates and, consequently, the measure of the performance.

2. Empirical Analysis

The empirical analysis is carried out using the Kauffman Firm Survey (KFS) data. The KFS is a longitudinal survey of new businesses that were founded in 2004 in the United States. At the moment there are six years of data available, from 2004 until 2009. The Baseline Survey contains information on 4928 firms and it is based on a stratified random sample drawn from the Dun & Bradstreet database. These data include detailed information both on the firm and up to ten business owners per firm.
The proposed index is computed for firms in the KFS, considering the following M-quantile production function model. For firm i in industry j and year t and for $q \in (0, 1)$

$$MQ_q(\log(REV_{ijt})|EXP_{ijt}, LAB_{ijt}, IND_{ij}) = \beta_{0q} + \beta_{1q}\log(EXP_{ijt}) \\ +\beta_{2q}\log(LAB_{ijt}) + \beta_{3jq}IND_{ij} + \beta_{4jq}\log(EXP_{ijt}) \cdot IND_{ij},$$

where REV is the total revenue of the business, EXP is total cost, LAB represents labour (employment and self-employment), and IND is the primary industry of the business. Due to the heterogeneity of businesses in the KFS we decided to take into account the type of industry in the performance measure computation. IND denotes industry dummies which allow for different aggregate industry effects.
For the empirical analysis, we fitted M-quantile regression hyperplanes over a grid of values of q, specifically for $q \in \{0.02, 0.04, \ldots, 0.98\}$. To obtain the regression fits we used a modification of the *rlm* function in R. Some businesses were excluded from the estimation procedure because some of the variables were not available. Adjusted survey weights were used in the estimation to take into account the stratified sample design used in the KFS and the location and response adjustments. The influence function used was Huber's Proposal 2 with $c = 1.345$ and s given by the median absolute deviation (MAD) of the residuals.
Analysis on the index was carried out using the SAS System for Windows (release 9.2; SAS Institute Inc, Cary, NC). The mean values of the index show that there has been

a general increase in performance during the first years. In 2008 and 2009 a decrease in the index means is observed, probably due to the global financial crisis. As far as symmetry and variability of the index are concerned, they remain stable over time.
Finally, as a first step towards understanding how different factors contribute to the variability in the index values (i.e. performance differences), the impact of several factors concerning both the entrepreneur and the firm on the index was modelled. As a first approximation, linear error component regression models (both static and dynamic) were considered. These models include latent variables, capturing unobserved features of the firm - such as managerial quality or structure - that can be viewed as being (roughly) constant over the period in question (Wooldridge, 2010). Special attention was paid to the correct estimation of the variances of the estimators in case of complex survey data. For this purpose we use the research software IVEware, developed at the University of Michigan (Raghunathan et al., 2011). Indeed, SAS, just like many other statistical software used for the analysis of panel data, is based on the simple random sample assumption and do not estimate the variance of estimators correctly. IVEware software uses SAS procedures but at the same time takes into account the complex sample design features. The analysis shows that direct personal involvement in the business affects positively performances. The fact that the entrepreneur is also a paid employee at business results in a 2% increase in IPM. An hour (worked per week) increase corresponds to a 0.06% improvement in the performance. Firms possessing patents and competitive advantages seem to achieve better performances. On the other side, a critical factor for new businesses performances is Hispanic origin.

References

Bianchi A., Salvati N. (2011), Asymptotic properties and variance estimators of the M-quantile regression coefficients estimators, Report 348, DSMAE.

Breckling J., Chambers R. (1988), M-quantiles, *Biometrika*, 75, 761–771.

Fried H.O., Lovell C.A.K., Schmidt S. (Eds) (2009), *The Measurement of Productive Efficiency and Productivity Change*, Oxford University Press, NY.

Kokic P., Chambers R., Breckling J., Beare S. (1997), A measure of production performance, *Journal of Business and Economic Statistics*, 15, 445–451.

Robb A., Reedy E.J., Ballou J., Des Roches D., Potter F., Zhao Z. (2010), An Overview of the Kauffman Firm Survey: Results from the 2004-2008 Data, available at http://www.kauffman.org/uploadedFiles/kfs_2010_report.pdf.

Raghunathan T.E., Solenberger, P.W., Van Hoewyk J. (2011), IVEware: Imputation and Variance Estimation, Version 0.2 Users Guide, available at http://www.isr.umich.edu/src/smp/ive/.

Wooldridge J. (2010), *Econometric analysis of cross section and panel data*, MIT Press, Cambridge.

Quaderni di Statistica
Vol. 14, 2012

Asymptotic properties of adaptive Gauss-Hermite based estimators in latent variable models

Silvia Bianconcini

Department of Statistics, University of Bologna
E-mail: silvia.bianconcini@unibo.it

Summary: Latent variable models have been widely applied in different fields of research in which the constructs of interest are not directly observable, so that one or more latent variables are required to reduce the complexity of the data. In these cases, problems related to the integration of the likelihood function of the model arise since analytical solutions do not exist. A recent applied numerical technique is the Adaptive Gauss-Hermite (AGH) that provides a good approximation of the function to be integrated, and it is also computational feasible in presence of many latent variables and/or random effects. In this paper, we analyze the asymptotic behavior of the AGH-based estimators used to perform inference in generalized linear latent variable models.

Keywords: Numerical integration, M-estimators, Generalized linear models.

1. Introduction

In several scientific fields, researchers often consider models based on one or more latent variables, since the constructs of interest are not directly observable. In these cases, problems related to the integration of the likelihood function arise since analytical solutions do not exist. One of the most recent applied numerical technique is the Adaptive Gauss-Hermite (AGH) quadrature. It consists of adjusting the quadrature locations by taking into account for specific features of the posterior density of the latent variables given the observations. This allows to better approximate the integrand, and to obtain estimates as accurate as those derived by applying the classical Gauss-Hermite (GH). Moreover, AGH requires a reduced number of quadrature points for dimension, being computationally feasible in presence of many latent variables and/or random ef-

fects. In this paper, we discuss the asymptotic behavior of the AGH-based estimators used to perform inference in Generalized Linear Latent Variable Models (GLLVM).

2. Generalized linear latent variable models: specification and estimation

The purpose of GLLVM is to describe the relationship between p manifest variables $\mathbf{y} = (y_1, \cdots, y_p)^T$ and $q < p$ latent variables $z_k, k = 1, 2, ..., q$. Since the latent variables are not observed, their realizations are treated as missing and are integrated out, giving the marginal density of the manifest variables

$$f(\mathbf{y}) = \int_{\mathbb{R}^q} g(\mathbf{y}|\mathbf{z}) h(\mathbf{z}_{(2)}) d\mathbf{z}_{(2)}, \tag{1}$$

where $\mathbf{z}_{(2)}$ is the q-dimensional vector of latent variables, assumed to be normally distributed with zero mean and covariance matrix $\mathbf{\Sigma}$, and $g(\mathbf{y}|\mathbf{z})$ is the conditional probability of the observed variables given $\mathbf{z} = (1, \mathbf{z}_{(2)})^T$. Under the conditional independence assumption of the responses given the latent variables, $g(\mathbf{y}|\mathbf{z})$ is given by

$$g(\mathbf{y} \mid \mathbf{z}) = \prod_{j=1}^{p} g_j(y_j \mid \mathbf{z}) \tag{2}$$

where each $g(y_j \mid \mathbf{z}), j = 1, ..., p$, belongs to the exponential family

$$g_j(y_j \mid \mathbf{z}) = \exp\left[\frac{y_j \boldsymbol{\alpha}_j^T \mathbf{z} - b_j(\boldsymbol{\alpha}_j^T \mathbf{z})}{\phi_j} + c_j(y_j, \phi_j)\right]. \tag{3}$$

Each distribution g_j will then depend on the type of manifest variable y_j, as well as on a set of parameters $\boldsymbol{\alpha}_j = [\alpha_{j0}, ..., \alpha_{jq}]^T$ (also called loadings) and scale ϕ_j.
Our aim is to derive estimators for all the model parameters, and to use them to establish a relationship between the observations $\mathbf{y}$ and the latent variables $\mathbf{z}_{(2)}$. At this regard, we consider a sample $(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n)$, with $\mathbf{y}_i = (y_1, \cdots, y_p)^T, i = 1, \cdots, n$. Let $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_p]$ be a $(q+1) \times p$ matrix of parameters, and $\boldsymbol{\phi} = [\phi_1, ..., \phi_p]^T$ the vector of scale parameters. Then the log-likelihood is

$$\begin{aligned} \ell(\boldsymbol{\alpha}, \boldsymbol{\phi}, \mathbf{\Sigma}) \quad = \quad & \sum_{i=1}^{n} \log \int_{\mathbb{R}^q} \left[\prod_{j=1}^{p} \exp\left\{\frac{y_j \boldsymbol{\alpha}_j^T \mathbf{z} - b_j(\boldsymbol{\alpha}_j^T \mathbf{z})}{\phi_j} + c_j(y_j, \phi_j)\right\}\right] \\ & (2\pi)^{-q/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2} \mathbf{z}_{(2)}^T \Sigma^{-1} \mathbf{z}_{(2)}\right] d\mathbf{z}_{(2)} \end{aligned} \tag{4}$$

where b_j and c_j are known functions that depend on the chosen distribution g_j. Equation (4) contains a multidimensional integral which cannot be computed explicitly, except when $g_j(y_j|\mathbf{z})$ is normal. Consequently, an approximation of this integral is needed, on which the bias and variance of resulting estimators will depend.

3. Estimators based on the adaptive Gauss-Hermite quadrature

The adaptive Gauss-Hermite quadrature technique has been widely discussed in the literature on generalized linear (mixed and latent variable) models (see *e.g.* Schilling and Bock, 2005). It consists of adjusting the quadrature locations with specific features of the posterior density of the latent variables given the observations. This allows to get a better approximation of the function to be integrated. Its application requires to rewrite eq. (1) as follows

$$f(\mathbf{y}) = \int_{\mathbb{R}^q} \frac{g(\mathbf{y}|\mathbf{z})h(\mathbf{z}_{(2)})}{h_1(\mathbf{z}_{(2)};\boldsymbol{\mu},\boldsymbol{\Psi})} h_1(\mathbf{z}_{(2)};\boldsymbol{\mu},\boldsymbol{\Psi})d\mathbf{z}_{(2)} \tag{5}$$

where $h_1(\mathbf{z}_{(2)};\boldsymbol{\mu},\boldsymbol{\Psi})$ is a multivariate normal density with moments

$$\boldsymbol{\mu} = \max_{\mathbf{z}_{(2)}\in\mathbb{R}^q} \left[\log g(\mathbf{y}|\mathbf{z}) + \log h(\mathbf{z}_{(2)})\right], \boldsymbol{\Psi} = \left(-\frac{\partial^2\left[\log g(\mathbf{y}|\mathbf{z}) + \log h(\mathbf{z}_{(2)})\right]}{\partial\mathbf{z}_{(2)}\partial\mathbf{z}_{(2)}^T}\right)|^{-1}_{\mathbf{z}_{(2)}=\boldsymbol{\mu}}$$

Hence, the AGH approximation of the density $f(\mathbf{y})$ is given by

$$f(\mathbf{y}) = 2^{\frac{q}{2}}|\mathbf{T}| \sum_{t_1,\cdots,t_q} g(\mathbf{y}|z^*_{t_1},\cdots,z^*_{t_q})h(z^*_{t_1},\cdots,z^*_{t_q})w^{**}_{t_1}\cdots w^{**}_{t_q} \tag{6}$$

where $\mathbf{T}\mathbf{T}^T = \boldsymbol{\Psi}, \mathbf{z}^* = (z^*_{t_1},\cdots,z^*_{t_q})^T = \sqrt{2}\mathbf{T}(z_{t_1},\cdots,z_{t_q})^T + \boldsymbol{\mu}, w^{**}_{t_k} = w_{t_k}\exp[z^2_{t_k}]$, and z_{t_k} and $w_{t_k}, k = 1,...,q$, are the classical GH nodes and weights, respectively. Hence, the approximate log-likehood function results

$$\begin{aligned}\ell(\boldsymbol{\alpha},\boldsymbol{\phi},\boldsymbol{\Sigma}) &= 2^{\frac{q}{2}}|\mathbf{T}|\sum_{i=1}^{n}\sum_{t_1,\cdots,t_q}\sum_{j=1}^{p}\left\{\frac{y_j\boldsymbol{\alpha}_j^T\mathbf{z}^* - b_j(\boldsymbol{\alpha}_j^T\mathbf{z}^*)}{\phi_j} + c_j(y_j,\phi_j)\right\} \\ &\quad -\frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathbf{z}^{*T}_{(2)}\Sigma^{-1}\mathbf{z}^*_{(2)}w^{**}_{t_1}\cdots w^{**}_{t_q}.\end{aligned} \tag{7}$$

The AGH-based estimators of $\boldsymbol{\alpha}, \boldsymbol{\phi}$, and $\boldsymbol{\Sigma}$ are found by equating the corresponding derivatives of equation (7) to 0, and evaluating the latent variables in their modes.

3.1. Relationship with the Laplace approximation

When only one quadrature point for dimension is used in eq. (7), the approximated log-likelihood is equivalent to the one obtained using the classical Laplace approximation. The statistical properties of the corresponding estimators, called LAMLE (LAplace Maximum Likelihood Estimators), have been studied in GLLVM by Huber et al. (2004). LAMLE belong to the class of M-estimators, hence they are consistent and asymptotically normal, under suitable conditions that must be checked for each particular g_j.

These properties can be extended to AGH estimators based on n_q quadrature points, since, for unidimensional integrals, Liu and Pierce (1994) showed their equivalence with higher order Laplace estimators. This result can be generalized to the q-dimensional integrals considered in this paper. Using the Eistein summation convention, the exact AGH solution of the integral (5) can be derived using the Taylor series expansion of $\nu(\mathbf{z}_{(2)}) = \frac{g(\mathbf{y}|\mathbf{z})h(\mathbf{z}_{(2)})}{h_1(\mathbf{z}_{(2)};\boldsymbol{\mu},\boldsymbol{\Psi})}$ around $\boldsymbol{\mu}$, as follows

$$\nu(\boldsymbol{\mu})\left[1+\sum_{m=2}^{n_q}\sum_{Q}\frac{1}{2m!}\frac{\nu^{i1,\ldots,i_{2m}}(\boldsymbol{\mu})}{\nu(\boldsymbol{\mu})}\nu_{q_1}(\boldsymbol{\mu})\cdots\nu_{q_m}(\boldsymbol{\mu})\right] \tag{8}$$

where $\nu^{i1,\ldots,i_{2m}}(\boldsymbol{\mu})$ denotes the partial derivative of the function ν with respect to $z_{i_1},\cdots,z_{i_{2m}}$ evaluated at the mode $\boldsymbol{\mu}$, and the second sum is over the partition $Q = q_1|\cdots|q_{2m}$ of $2m$ indices into m blocks, each of size 2.
We will show in detail that eq. (8) can be equivalently derived by setting $L(z_{(2)}) = \log g(\mathbf{y}|\mathbf{z}) + \log h(\mathbf{z}_{(2)})$ in eq. (1), and by approximating the integral with an higher order Laplace approximation as follows

$$(2\pi)^{\frac{q}{2}}|\Psi|^{\frac{1}{2}}\exp(L(\boldsymbol{\mu}))\left[1+\sum_{m=2}^{\infty}\sum_{P,Q}\frac{(-1)^t}{2m!}L^{p_1}(\boldsymbol{\mu})\cdots L^{p_t}(\boldsymbol{\mu})L_{q_1}(\boldsymbol{\mu})\cdots L^{q_m}(\boldsymbol{\mu})\right] \tag{9}$$

where the second sum is over all partitions P, Q, such that $P = p_1|\cdots|p_t$ is a partition of $2m$ indices into t blocks, each of size 3 or more, and $Q = q_1|\cdots|q_{2m}$ is a partition of $2m$ indices into m blocks, each of size 2. Shun and McCullagh (1995) showed that, for fixed q, the usual asymptotic order of the term corresponding to the bipartition (P, Q) is $O(p^{t-m})$, p being the number of observed items. Hence, it follows that the asymptotic error of the AGH based on n_q quadrature points is given by $O\left(q^{(n_q+1)}p^{-\left[\frac{n_q}{3}+1\right]}\right)$.

References

Huber P., Ronchetti E., Feser V. (2004), Estimation of Generalized Linear Latent Variable Models, *Journal of the Royal Statistical Society - Series B*, 66, 893–908.

Liu Q., Pierce D.A. (1994), A note on Gauss-Hermite quadrature, *Biometrika*, 81, 624–629.

Schilling S., Bock R.D. (2005), High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature, *Psychometrika*, 70, 533–555.

Shun Z., McCullagh P. (1995), Laplace approximation of high dimensional integrals, *Journal of Royal Statistical Society - Series B*, 57, 749–760.

Quaderni di Statistica
Vol. 14, 2012

Approximate likelihood inference in latent variable models for categorical data

Silvia Bianconcini
Department of Statistics, University of Bologna
E-mail: silvia.bianconcnini@unibo.it

Silvia Cagnone
Department of Statistics, University of Bologna
E-mail: silvia.cagnone@unibo.it

Dimitris Rizopoulos
Department of Biostatistics, Erasmus University Rotterdam
E-mail: d.rizopoulos@erasmusmc.nl

Summary: Latent variable models represent a useful tool for the analysis of complex data characterized by the fact that the constructs of interest are not directly observable. One problem related to these model is that, in presence of categorical data, the integrals involved in the maximization of the likelihood function are not solvable analytically. In this paper we propose a new approach for approximating integrals for latent variable models for binary data. This approach is based on a fundamental theorem by Rahman and Xu (2004) and consists of reducing the dimensionality of the integrals involved in the computations.

Keywords: Dimension reduction method, Binary data, Latent variables.

1. Introduction

Latent variable models represent a useful tool in different fields of research in which the constructs of interest are not directly observable, so that one or more latent variables are required to reduce the complexity of the data. In these cases, problems related to the integration of the likelihood function of the model can arise since analytical solutions do not exist. Usually, the Gauss Hermite (GH) numerical integration quadrature is used to overcome this problem. GH works quite well in several situations, but it requires a

great number of quadrature points per dimension in order to obtain accurate estimates, becoming unfeasible in presence of many latent variables. The most recent Adaptive Gauss-Hermite (AGH) numerical quadrature, that consists of adjusting the quadrature locations (GH nodes) taking into account for specific features of the posterior density of the latent variables given the observations, allows a better approximation of the function to be integrated. As a consequence, the AGH requires a smaller number of quadrature points per dimension than GH to obtain the same accuracy of the estimates, being more feasible in presence of many latent variables and/or random effects. However, the AGH requires a quite high computational effort.
The aim of this study is to propose an alternative approach based on a reduction in the dimensionality of the integrals involved in the computations with significant computational savings. This approach is based on a fundamental theorem by Rahman and Xu (2004) that provides a convenient way to represent the Taylor series expansion of the integrand up to a specific dimension without involving any partial derivative. The dimension reduction is applied to latent variable models for binary data.

2. Model specification

Let $\mathbf{y}' = (y_1, \ldots, y_p)$ be a vector of p observed binary variables and $\mathbf{z}' = (z_1, \ldots, z_q)$ be a vector of q continuous latent variables. The probability associated to the individual response pattern can be expressed as

$$f(\mathbf{y}_l) = \int_{R^q} g(\mathbf{y}_l|\mathbf{z}_l)h(\mathbf{z}_l)d\mathbf{z}_l, \qquad l = 1, ..., n. \tag{1}$$

where

$$g(\mathbf{y}_l|\mathbf{z}_l) = \prod_{j=1}^{p} g(y_{jl}|\mathbf{z}_l) = \prod_{j=1}^{p} \pi_j(\mathbf{z}_l)^{y_{jl}}(1 - \pi_j(\mathbf{z}_l))^{(1-y_{jl})} \tag{2}$$

under the assumption of conditional independence of the observed variables given the latent variables. $\pi_j(\mathbf{z})$ is the probability of a positive response to the item j, and it is dependent on the latent variables through the logit link function as follows

$$\text{logit}[\pi_j(\mathbf{z})] = \alpha_{0j} + \sum_{k=1}^{q} \alpha_{jk} z_k.$$

α_{0j} and α_{jk} $(j = 1, ..., p; k = 1, ..., q)$ can be interpreted as the intercept and the loadings of a factor model. For sake of simplicity, we assume that $\mathbf{z}$ follows a standard normal distribution.
For a random sample of size n, the observed data log-likelihood is defined as

$$\ell(\boldsymbol{\theta}) = \sum_{l=1}^{n} \log f(\mathbf{y}_l) = \sum_{l=1}^{n} \log \int_{R^q} g(\mathbf{y}_l \mid \mathbf{z}_l)h(\mathbf{z}_l)d\mathbf{z}_l \tag{3}$$

where $\boldsymbol{\theta}' = (\alpha_{01}, ..., \alpha_{0p}, \alpha_{11}, ..., \alpha_{1q}, ..., \alpha_{p1}, ..., \alpha_{pq})$ refers to the model parameters. The integrals involved in eq. (3) do not admit an analytical solution, since they are defined on the latent variable space. In the following section we show the solution based on the dimension reduction method.

3. Dimension Reduction method

The dimension reduction can be applied to integrals of the form

$$E[f(\mathbf{z})] = \int_{R^q} f(\mathbf{z})h(\mathbf{z})d\mathbf{z}. \tag{4}$$

Hence, in order to apply the transformation to the integrals involved in formula (3) we apply the following transformation

$$\int_{R^q} g(\mathbf{y}|\mathbf{z})h(\mathbf{z})d\mathbf{z} = \int_{R^q} \exp[\log g(\mathbf{y}|\mathbf{z}) + \log h(\mathbf{z})]d\mathbf{z} = \int_{R^q} \exp[L(\mathbf{z})]d\mathbf{z}, \tag{5}$$

where $L(\mathbf{z}) = \log g(\mathbf{y}|\mathbf{z}) + \log h(\mathbf{z})$. Considering the Taylor series expansion of $L(\mathbf{z})$ around its mode $\hat{\mathbf{z}} = \operatorname{argmax}_{\mathbf{z}\in R^p}[L(\mathbf{z})]$, the higher order terms of the decomposition can be expressed as follows $\nu(\mathbf{z}) = L(\mathbf{z}) - L(\hat{\mathbf{z}}) - \frac{1}{2}(\mathbf{z} - \hat{\mathbf{z}})'L''(\hat{\mathbf{z}})(\mathbf{z} - \hat{\mathbf{z}})$ so that

$$\int_{R^q} \exp[L(\mathbf{z})]d\mathbf{z} = \exp[L(\hat{\mathbf{z}})](2\pi)^{q/2}|\hat{\boldsymbol{\Sigma}}|^{1/2} \int_{R^q} \exp[\nu(\mathbf{z})]\phi(\mathbf{z}; \hat{\mathbf{z}}, \hat{\boldsymbol{\Sigma}})d\mathbf{z} \tag{6}$$

where $\phi(\mathbf{z}; \hat{\mathbf{z}}, \hat{\boldsymbol{\Sigma}})$ is the normal density function with mean vector $\hat{\mathbf{z}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$. The latter is the Hessian matrix of $L(\mathbf{z})$ evaluated at the mode $\hat{\mathbf{z}}$, *i.e.* $\hat{\boldsymbol{\Sigma}}^{-1} = -L''(\hat{\mathbf{z}})$. The integral involved in expression (6) can be viewed as the expected value defined in eq. (4), where $f(\mathbf{z}) = \exp[\nu(\mathbf{z})]$ and $\phi(\mathbf{z}; \hat{\mathbf{z}}, \hat{\boldsymbol{\Sigma}}) = h(\mathbf{z})$.
Since the dimension reduction requires standardized integrating variables, we apply the Cholesky decomposition $\hat{\boldsymbol{\Sigma}} = \mathbf{C}\mathbf{C}'$ getting

$$\int_{R^q} \exp[\nu(\mathbf{C}\mathbf{z}^* + \hat{\mathbf{z}})]\phi(\mathbf{z}^*; \mathbf{0}, \boldsymbol{I})d\mathbf{z}^* = \int_{R^q} f(\mathbf{z}^*)\phi(\mathbf{z}^*; \mathbf{0}, \boldsymbol{I})d\mathbf{z}^*. \tag{7}$$

The solution of the integral (7) is obtained by approximating $f(\mathbf{z}^*)$ as follows $\hat{f}(\mathbf{z}^*) = \sum_{w=0}^{s} t_w$ where

$$t_w = \sum_{j_1, j_2, ..., j_w} \sum_{k_1 < k_2 < ... < k_w} \frac{\partial^{j_1 + j_2, ..., + j_w} f}{\partial z_{k_1}^{*j_1} \partial z_{k_2}^{*} j_2 \ldots \partial z_{k_w}^{*j_w}}(\mathbf{0}) z_{k_1}^{*j_1} z_{k_2}^{*j_2} \ldots z_{k_w}^{*j_w}.$$

If in eq. (7) we replace $f(\mathbf{z}^*)$ with the approximation $\hat{f}(\mathbf{z}^*)$, we get a sum of uni-dimensional, bi-dimensional ,..., s-dimensional integrals that can be solved analytically

since they are related to the moments of standardized univariate, bivariate,..., s-variate normal random variables. That is,

$$\int_{R^q} f(\mathbf{z}^*)\phi(\mathbf{z}^*)d\mathbf{z}^* \approx f(\mathbf{0}) + \sum_{j_1=1}^{\infty} \frac{1}{j_1!} \sum_{k=1}^{q} \frac{\partial^{j_1} f}{\partial z_k^{*j_1}}(\mathbf{0}) \int_{-\infty}^{\infty} z_k^{*j_1} \phi(z_k^*) dz_k + \cdots$$

$$+ \sum_{j_1,\ldots,j_s} \sum_{k_1<\ldots<k_s} \frac{\partial^{j_1+j_2,\ldots,+j_s} f}{\partial z_{k_1}^{*j_1} \partial z_{k_2}^{*j_2} \ldots \partial z_{k_s}^{*j_s}}(\mathbf{0}) \int_{R^s} z_{k_1}^{*j_1} z_{k_2}^{*j_2} \ldots z_{k_s}^{*j_s} \prod_{w=1}^{s} \phi(z_w^*) dz_{k_1^*} \ldots dz_{k_s^*}$$

The approximation depends on the choice of (i) the number s of terms to be considered, but also on the choice of (ii) the order $j_1, j_2, \ldots j_w, w = 1, \ldots, s$, of the derivatives at which we truncate each term. The main advantage of this approach relies on the fact that the integrals can be solved analytically but it requires the computation of partial derivatives of f with respect to $\mathbf{z}$. Hence, the computational complexity of the approach increases as the order of the derivatives increases. To overcome this latter drawback, Xu and Rahman (2004) show that the s-variate approximated function $\hat{f}(\mathbf{z}^*)$ can be equivalently expressed as

$$\hat{f}(\mathbf{z}^*) = \sum_{i=0}^{s} (-1)^i \begin{pmatrix} q-s+i-1 \\ i \end{pmatrix} f_{s-i}(\mathbf{z}^*) \tag{8}$$

where $f_{s-i}(\mathbf{z}^*)$ is the $(s-i)$-variate function, $i = 0, \ldots, s$,

$$f(0, \cdots, z_{k_1}^*, 0, \cdots, 0, z_{k_2}^*, 0, \cdots, 0, z_{k_{s-i}}^*, 0, \cdots, 0).$$

Replacing $f(\mathbf{z}^*)$ with the approximated function as expressed in eq. (8), the integral (7) results

$$\int_{R^q} f(\mathbf{z}^*)\phi(\mathbf{z}^*)d\mathbf{z}^* \approx$$

$$\sum_{i=0}^{s} (-1)^i \begin{pmatrix} q-s+i-1 \\ i \end{pmatrix} \int_{R^{s-i}} \sum_{k_1<\ldots<k_{s-i}} f_{s-i}(\mathbf{z}^*)\phi(z_{k_1}^*) \cdots \phi(z_{k_{s-i}}^*) dz_{k_1}^* \ldots dz_{k_{s-i}}^*$$

where each integral depends on $s, s-1, \cdots, 0$ variables, respectively, and they can be easily approximated by using the classical Gauss Hermite quadrature technique.

References

Rahman S., Xu H. (2004), A Univariate Dimension-Reduction Method for Multi-Dimensional Integration in Stochastic Mechanics, *Probabilistic Engineering Mechanics*, 19, 393-408.

Xu H., Rahman S. (2004), A Generalized Dimension-Reduction Method for Multi-Dimensional Integration in Stochastic Mechanics, *International Journal for Numerical Methods in Engineering*, 61, 1992-2019.

Quaderni di Statistica
Vol. 14, 2012

Bayesian nonparametric predictions for count time series

Luisa Bisaglia
Department of Statistics, University of Padua
E-mail: bisaglia@stat.unipd.it

Antonio Canale
Department of Economics and Statistics, University of Turin
E-mail: antonio.canale@unito.it

Summary: Nonetheless the central role of the Box-Jenkis Gaussian autoregressive moving average models for continuous time series, there is no such a leading technique for count time series. In this paper we introduce a Bayesian nonparametric methodology for producing coherent predictions of a count time series $\{X_t\}$ using the nonnegative INteger-valued AutoRegressive process of the order 1 (INAR(1)) introduced by Al-Osh and Alzaid (1987) and McKenzie (1988). INAR models evolve as a birth-and-death process where the value at time t can be modeled as the sum of the survivors from time $t-1$ and the outcome of an innovation process with a certain discrete distribution. Obviously such components are not observable. Our predictions are based on estimates of the p-step ahead predictive mass functions assuming a nonparametric prior distribution for the innovation process. Precisely we model this distribution with a Dirichlet process mixture of rounded Gaussians (Canale and Dunson, 2011). This class of prior has large support on the space of probability mass functions and is able to generate almost any count distribution including over/under-dispersion or multimodality. An efficient Gibbs sampler is developed for posterior computation and the methodology is used to analyze real data sets.

Keywords: INAR(1), Dirichlet process mixtures, Gibbs sampling algorithm.

1. Introduction

Recently, there has been a growing interest in studying nonnegative integer-valued time series and, in particular, time series of counts. Examples are categorical time series,

binary processes, birth-death models and counting series.

The most common approach to build an integer-valued autoregressive processes is using a probabilistic operation called thinning. Using binomial thinning, Al-Osh and Alzaid (1987) and McKenzie (1988) first introduced integer-valued autoregressive processes (INAR). A recent review on integer-valued AR processes can be found in Silva et al. (2005) and Jung and Tremayne (2011). While theoretical properties of INAR models have been extensively studied in the literature, relatively few contributions discuss the development of forecasting methods that are coherent, in the sense of producing only integer forecasts of the count variable. Freeland and McCabe (2004), in the context of INAR(1) process with Poisson innovations suggest some solutions that are somewhat problem-specific. Thus, McCabe and Martin (2005) consider the Bayesian point of view and present a methodology for producing coherent forecasts of low count time series that is completely general. The predictive probability mass function, defined only over the support of the discrete count variable, is a natural outcome of Bayes theorem.The results are valid for any sample size and not only asymptotically, moreover the innovations can be any arbitrary discrete distribution, within a specified finite set of distributions. In particular, the authors focus on Poisson, binomial and negative binomial distributions.

In this paper, we consider INAR(1) models with flexible specifications of the error term under a Bayesian nonparametric approach. The assumption of a nonparametric prior with large support for the innovation distribution, bypasses the need to specify a finite set of discrete distribution as in McCabe and Martin (2005). Our approach leads to two main improvements: first we overcome the specification of the predictive probability as a mixture of K predictive distributions, and second we do not rely on the usual strict parametric models. Among the different proposal made in the Bayesian nonparametric literature to model count distributions, we use that of Canale and Dunson (2011).

2. Model specification

To introduce the class of INAR model we first recall the thinning operator, '$\circ$', defined as follows.

Definition *Let Y be a non negative integer-valued random variable, then for any $\alpha \in [0, 1]$*

$$\alpha \circ Y = \sum_{i=1}^{Y} X_i$$

where X_i is a sequence of iid count random variables, independent of Y, with common mean α.

The INAR(1) process $\{Y_t; t \in \mathbf{Z}\}$ is defined by the recursion

$$Y_t = \alpha \circ Y_{t-1} + \epsilon_t \tag{1}$$

where $\alpha \in [0,1]$, and ϵ_t is sequence of iid discrete random variables with finite first and second moment. The components of the process $\{Y_t\}$ are the surviving elements of the process Y_{t-1} during the period $(t-1,t]$, and the number of elements which entered the system in the same interval, ϵ_t. Each element of Y_{t-1} survives with probability α and its survival has no effect on the survival of the other elements, nor on ϵ_t which is not observed and cannot be derived from the Y process in the INAR(1) model. In the next section we discuss a nonparametric prior for the distribution of the error term.

To define a nonparametric model for counts, Canale and Dunson (2011) proposed to round an underlying variable having an unknown density given a Dirichlet process mixture of Gaussians prior. Such rounded mixture of Gaussians (RMG) have been showed to be highly flexible and having excellent performance in small samples while having appealing asymptotic properties in terms of large support and strong posterior consistency. Let the probability that the discrete error equals j, for $j \in \mathbf{N}$ to be

$$p(j) = g(f)[j] = \int_{a_j}^{a_{j+1}} f(y^*)dy^* \tag{2}$$

with the thresholds chosen as $a_0 = -\infty$ and $a_j = j-1$ for $j \in \{1,2,\dots\}$ and modelling the underlying f as the mixture model

$$f(y^*;P) = \int \phi(y^*;\mu,\tau^{-1})dP(\mu,\tau), \quad P \sim DP(\eta P_0). \tag{3}$$

Here, $\phi(y;\mu,\tau^{-1})$ is a Gaussian density having mean μ and precision τ and $DP(\eta P_0)$ corresponding to the Dirichlet process with P_0 chosen to be Normal-Gamma and $\eta > 0$. Equations (2)–(3) induce a prior $p \sim \Pi$ over $\mathcal{C}$, the space of the probability mass functions on the non negative integers.

3. p-step ahead predictive probability mass function

Exploiting the birth-and-death process interpretation of the INAR(1) model, the distribution of Y_t given y_{t-1}, α and p is

$$Pr(Y_t = y_t \mid y_{t-1},\alpha,p) = \sum_{s=0}^{\min\{y_t,y_{t-1}\}} Pr(B_{y-1}^{\alpha} = s) \times p(y_t - s) \tag{4}$$

where p is a random probability measure obtained through (2)–(3) and $B_k^{\pi} \sim \text{Be}(k,\pi)$.

The likelihood function given $\mathbf{y} = (y_1,\dots,y_T)$ of α and the random discrete measure p turns out to be

$$\ell(\theta \mid \mathbf{y}) \quad \propto \quad \prod_{t=2}^{T} \sum_{s=0}^{\min\{y_t,y_{t-1}\}} \alpha^s(1-\alpha)^{y_{t-1}-s}p(y_t - s) \tag{5}$$

where $\theta \in \Theta$ and $\Theta = \mathbf{R} \times \mathcal{C}$. The posterior distribution can be obtained as

$$\pi(\theta \mid \mathbf{y}) \propto \ell(\theta \mid \mathbf{y})\pi(\theta) \tag{6}$$

where $\pi(\theta)$ is the prior probability. Given the nonparametric prior $p \sim \Pi$ it is sufficient to elicit a prior for $\alpha \sim \pi_\alpha$. In presence of prior information we can use a beta distribution with given mean corresponding to one's prior belief about α. Being noninformative one can assume a uniform prior distribution between zero and one. Assuming that α and p are independent a priori, the prior $\pi(\theta)$ is $\pi(\theta) = \Pi \times \pi_\alpha$.

The p-step ahead probability mass function is here defined as

$$Pr(Y_{T+p} = j \mid \mathbf{y}) = \int_\Theta Pr(Y_{T+p} = j \mid \mathbf{y}, \theta) d\pi(\theta \mid \mathbf{y}) \tag{7}$$

where $\pi(\theta \mid \mathbf{y})$ is the posterior distribution (6).

The following Gibbs sampler computes the quantity in (7) iterating the steps:

1. Data augmentation step given p and α.
   - For $t = 2, \ldots, T$, simulate $B_t \sim \text{Be}(y_{t-1}, \alpha)$
   - For $t = 2, \ldots, T$, simulate $\epsilon_t^* \sim f$ where f is as in (2)–(3) under the constraints $a_{y_t - B_t} \leq \epsilon_t^* \leq a_{y_t - B_t + 1}$
2. Update the parameters of the RMG as in Canale and Dunson (2011)
3. Update α from its conditional posterior distribution via Metropolis-Hastings step
4. After burn in, simulate Y_{t+p} as in equation (4)

References

Al-Osh, M. A. and A. A. Alzaid (1987), First order integer-valued autoregressive INAR(1) process. *Journal of Time Series Analysis*, 8(3), 261–275.

Canale, A. and D. B. Dunson (2011), Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106, 1528–1539.

Freeland, R. K. and B. P. M. McCabe (2004), Forecasting discrete valued low count time series. *International Journal of Forecasting*, 20, 427–434.

Jung, R. C. and A. R. Tremayne (2011), Useful models of time series of counts or simply wrong ones? *Advances in Statistical Analysis*, 95, 59–91.

McCabe, B. P. M. and G. M. Martin (2005), Bayesian predictions of low count time series. *International Journal of Forecasting*, 21, 315–330.

McKenzie, E. (1988), Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability*, 20, 822–835.

Silva, I., M. E. Silva, I. Pereira, and N. Silva (2005), Replicated inar(1) process. *Methodology and Computing in Applied Probability*, 7, 893–899.

Quaderni di Statistica
Vol. 14, 2012

Reliability measurement for polytomous ordinal items: the empirical polychoric ordinal Alpha

Andrea Bonanomi Marta Nai Ruscone Silvia Angela Osmetti
Department of Statistical Sciences, Università Cattolica del Sacro Cuore di Milano
E-mail: andrea.bonanomi@unicatt.it, marta.nairuscone@unicatt.it, silvia.osmetti@unicatt.it

Summary: We aim at proposing a new reliability measurement for polytomous ordinal items. Conventionally, reliability coefficients, such as Cronbach Alpha, are calculated using the Pearson correlation matrix. We suggest a modification of the classical Cronbach Alpha for ordinal variables, by using the empirical polychoric correlation coefficient. It builds upon the theoretical framework of the classic polychoric correlation coefficient, but relaxes its fundamental assumption that the ordinal variables have a specific underlying continuous joint distributions. The proposed index, defined by the means of the empirical copula, is a non parametric reliability measure particularly suitable for ordinal data. A simulation study is conducted in order to compare the proposed index to classical reliability measures.

Keywords: Reliability, Cronbach Alpha, Polychoric correlation.

1. Introduction

Reliability is consistency of a set of measurements in research involving test construction and use. In literature several reliability measures have been proposed. The most widely used reliability coefficient in the social science is Cronbach Alpha (Cronbach, 1951).

Consider a set of items Y_j for $j = 1, \cdots, k$, the Cronbach Alpha is given by

$$\alpha = \frac{k}{k-1} \frac{\sum \sum_{i \neq j} \sigma_{ij}}{\sum \sum_{i,j} \sigma_{ij}}$$

where σ_{ij} is the covariance of the pair (Y_i, Y_j) (see Cortina, 1993). It is easy to estimate from the data simply by using sample variances and sample covariances.

If we consider standardized variables, the index becomes

$$\alpha = \frac{k * \overline{\rho}}{1 + (k-1)\overline{\rho}},$$

where $\overline{\rho} = \sum \sum_{i \neq j} \sigma_{ij} \setminus (k(k-1))$.

This index is frequently applied when analyzing items on self-report instruments such as personality tests and surveys that often use rating scales with a small number of response options. If the items have a Likert type rating response scale a modification of the classical Cronbach Alpha is necessary. Zumbo *et al*. (2007) propose an ordinal Alpha; they use the polychoric correlation, instead of the Pearson correlation, thereby taking into account the ordinal nature of the data. The polychoric correlation (see Pearson, 1900) is a measure of bivariate association arising when both observed variables are ordered categorical variables deriving from polychotomizing underling latent continuous variables. If the underling continuous distribution is not specified, the empirical polychoric correlation can be used (Ekström, 2009). By using the empirical polychoric correlation we propose a new reliability index for categorical ordinal variables, the empirical polychoric ordinal Alpha.

2. Polychoric correlation coefficient

Let Y_i and Y_j be two ordinal variables with categories r and s, respectively. The fundamental idea of the polychoric correlation is to assume that the two ordinal variables are discretized random variables with a underlying continuous joint distribution belonging to bivariate distributions. The discretization consists to impose thresholds on the domain of the bivariate density function into rectangles corresponding to the cells of the contingency table of the two ordinal variables. The volumes of the rectangles should equal to the joint probabilities of the two categorical variables. This measure was originally proposed by Karl Pearson (1900) under the assumption of a underling standard normal distribution. In this case the polychoric correlation corresponds to the linear correlation coefficient.

Formally, let H_θ be the supposed bivariate cumulative distribution function of (Y_i, Y_j) and F_1 and F_2 the marginal cumulative distribution functions. Let u_i for $i = 1, 2, .., r$ and υ_j for $j = 1, 2, .., s$ be the probabilities of the observed values with order less than or equal to i and j respectively, and let u_0 and υ_0 equal to zero. Let $A_1, ..., A_{r \cdot s}$ be the rectangles obtained by the discretization of the domain of the bivariate distribution function, such that $A_{ij} = [F_1^{-1}(u_{i-1}), F_1^{-1}(u_i)] \times [F_2^{-1}(\upsilon_{i-1}), F_2^{-1}(\upsilon_i)]$. Let $p_1, ..., p_{r \cdot s}$ be the joint probabilities of the ordinal variables corresponding to the rectangles $A_1, ..., A_{r \cdot s}$. The volumes of the rectangles should be equal to the joint probabilities $p_1, ..., p_{r \cdot s}$:

$$(H_\theta(A_1), ..., H_\theta(A_{r \cdot s})) = (p_1, ..., p_{r \cdot s}). \tag{1}$$

The polychoric correlation is

$$r_{pc} = 2\sin(\rho_S(H_\theta))\pi/6, \tag{2}$$

for the values of θ satisfying the above equation, where ρ_s denotes the Spearman correlation coefficient.

3. Cronbach Alpha via Empirical correlation coefficient

For the estimation of the polychoric correlation coefficient in (2) it is necessary to specify the family of distributions H_θ. Instead of considering the assumed bivariate distribution it is often more convenient to consider its corresponding family of copulae C_λ (see Nelsen, 2006). The copula is a bivariate cumulative distribution function with uniform marginal random variables in $[0, 1]$. Consequently, the two ordinal variables are without loss of generality assumed to be discretized numerical random variables with support on the unit interval. Therefore, the rectangles $A_1, A_2, ..., A_{r\cdot s}$ are defined as $A_{i\cdot j} = [u_{i-1}, u_i] \times [v_{j-1}, v_j]$ and the condition in (1) becomes

$$(C_\lambda(A_1), \cdots, C_\lambda(A_{rs})) = p_1, \cdots, p_{rs}.$$

The polychoric correlation coefficient is

$$r_{pc} = 2\sin(\rho_s(C_\lambda)\pi/6). \tag{3}$$

In order to estimate the polychoric correlation coefficient is necessary to specify the copula C_λ. Therefore, the polychoric correlation coefficient depends on the choice of the family of copulae. We consider the empirical version of the polychoric correlation coefficient proposed by Ekström (2009). The distribution family does not need to be specified. It only needs to be assumed that an underlying continuous distribution exists.

The empirical polychoric correlation coefficient is based on the empirical copula (Deheuvels, 1979)

$$\widehat{C}_n(u_i, v_j) = \frac{1}{n}\sum_{k=1}^{n} I_{[0,u_i]\times[0,v_j]}(x_k, y_k),$$

where I_A is the indicator function. Let

$$\widehat{E}_n(u, v) = \sum_{k=1}^{r\cdot s} a_k I_{A_k}(u, v),$$

where

$$a_k = \frac{1}{4}\left(\widehat{C}_n(u_i, v_j) + \widehat{C}_n(u_i, v_{j-1}) + \widehat{C}_n(u_{i-1}, v_j) + \widehat{C}_n(u_{i-1}, v_{j-1})\right).$$

The empirical polychoric correlation coefficient is then, in analogy with the polychoric correlation coefficient, defined as

$$r_{epc} = 2\sin(\rho_s(\widehat{E}_n)\pi/6 \tag{4}$$

We use the index in (4) to define the empirical polychoric ordinal α. Supposing that the k items have an underlying standardized continuous distribution, the the empirical polychoric ordinal α is

$$\alpha_{epc} = \frac{k * \overline{r}_{epc}}{1 + (k-1)\overline{r}_{epc}}$$

where $\overline{r}_{epc}$ is the mean of r_{epc} for the k items.

A simulation study is conducted in order to compare the proposed index to classical reliability measures. We compare the bias and the mean square error (MSE) of the indexes assuming different underlying copulae (Gaussian, Frank, Clayton), sample size and numbers of alternative response scales. We obtaine good results for the MSE for the proposed index respect to the ones obtained with the classical Cronbach Alpha and the ordinal Alfa.

References

Cortina J.M. (1998), What is coefficient Alpha? An examination of theory and applications, *Journal of Applied Psychology*, 78, 98–104.

Cronbach L.J. (1951), Coefficient alpha and the internal structure of tests, *Psychometrika*, 16, 297–334.

Deheuvels P. (1979), La fonction de dépendence empirique et ses propriétés. Un test non paramétrique d'indépendence, *Acad. Roy. Belg. Bul. Cl. Sci.*, 65, 274–292.

Ekström J. (2009), Contributions to the theory of measures of association for ordinal variables, *Digital comprehensive summaries of Uppsala dissertation from the Faculty of Social Sciences* n. 50, ACTA Universitatis Upsaliensis, Uppsala.

Nelsen R.B. (2006), *An Introduction to Copulas*, Springer, New York.

Pearson K. (1900), Mathematical contribution to the theory of evolution, *Philosophical Transactions of the Royal Society of London. Series A*, 195, 1–47.

Zumbo B.D., Gadermann A.M., Zeisser C. (2007), Ordinal versions if coefficient Alpha and Theta for Likert rating scales, *Journal of Modern Applied Statistical Methods*, 6, 21–29.

Quaderni di Statistica
Vol. 14, 2012

Regression trees for change point analysis of ordinal time series

Carmela Cappelli Francesca Di Iorio
Dipartimento di Teorie e Metodi delle Scienze Umane e Sociali, Università di Napoli
E-mail: carcappe@unina.it, E-mail: fdiiorio@unina.it

Pierpaolo D'Urso
Dipartimento di Scienze Sociali, Sapienza, Università di Roma
E-mail: pierpaolo.durso@uniroma1.it

Summary: In this paper we describe how to conduct a change-point analysis when dealing with time ordered data that are measured on an ordinal scale. In order to treat such time series we propose to employ a fuzzy coding i.e. the ordinal scale is converted into a fuzzy variable. Then, to detect the number and location of change points we employ in the framework of Atheoretical Regression Trees (ART) a deviation measure for fuzzy variables. The proposal is illustrated by an application to a real ordinal time series.

Keywords: Regression trees, Ordinal time series, Fuzzy coding.

1. Motivation

Change-point analysis comprises various statistical tools which are employed for determining if and when a change in a data set has occurred. In case of multiple changes in mean Cappelli *et al.* (2008) have proposed a method called ART (Atheoretical Regression Trees) that employs Least Square Regression Trees (so forth denoted LSRT) to estimate the number and location of multiple change points; extensive simulation studies, comparison with current methods and applications to various real time series have provided evidence of the usefulness of the approach (see Rea *et al.*, 2010).

In this paper we describe how to employ ART to conduct a change-point analysis of time ordered data that are measured on an ordinal scale. Indeed, in many real life situations we meet data that derive from human perception or expert judgment and sometimes the data are ordered sequences of observations and by definition time series. In case such as these change-point analysis is a useful tool for monitoring and control. In general, treating ordinal data as either numerical or categorical might entail loss of information

or inaccuracy because judgments and evaluations are discrete measures of continuous latent variables and they are intrinsically accompanied by a vagueness (uncertainty) that needs to be properly taken into account. At this aim we consider a fuzzy coding i.e. the ordinal scale is converted into a fuzzy variable and, in order to estimate the number and location of change points of the fuzzified time series we employ, in the framework of ART, a deviance measure decomposition for fuzzy variables (D'Urso and Santoro, 2006) based on the Yang-Ko's metric (Yang and Ko, 1996).

In the following we briefly introduce the issue of detecting multiple changes in mean and the ART procedure and then we describe how to deal with time series measured on an ordinal scale. Then, we present the results of an application of the proposed approach to a time series of Italian wine judgments.

2. Basics and method

The issue of estimating multiple changes in mean can be briefly illustrated as follows. Let y_t be a time series characterized by $m+1$ regimes and m change points so that $t = T_{(j-1)} + 1, \ldots, T_j$ and $j = 1, \ldots, m+1$ (we adopt the convention that $T_0 = 0$ and $T_{m+1} = T$ where T is the length of the series). A common estimation method of the set of unknown break dates is that based on the least square principle i.e. the estimated break points $(\hat{T}_1, \ldots, \hat{T}_m)$ are such that:

$$(\hat{T}_1, \ldots, \hat{T}_m) = argmin_{(T_1,\ldots,T_m)} SSR(T_1, \ldots, T_m) \tag{1}$$

where $SSR(T_1, \ldots, T_m) = \sum_{j=1}^{m+1} \sum_{t=T_{(j-1)}+1}^{T_j} (y_t - \mu_j)^2$ is the sum of squared residuals. To detect the presence of such structural changes Cappelli *et al.* (2008) have proposed a procedure based on LSRT. In LSRT a node h is split into its left and right descendants h_l and h_r, respectively, to reduce the deviation of the dependent variable y_t, thus, the algorithm selects the split that minimizes:

$$SSR(h_l) + SSR(h_r) = \sum_{g \in \{l,r\}} \sum_{y_t \in h_g} (y_t - \hat{\mu}(h_g))^2 \tag{2}$$

where $\hat{\mu}(h_g)$ is the mean of the y_t values in node h_g ($g \in \{l, r\}$) thus, the splitting criterion (2) corresponds to the objective function (1) computed for a binary partition and it is based on the decomposition property. Once the binary partition of a node is performed, the splitting process is recursively applied to each subnode until either the subnodes reach a minimum size or no improvement of the criterion can be achieved. Indeed, tree regressing a time series y_t on a sequence of completely ordered numbers $k = 1, \ldots, T$ provides a partition of the series into contiguous segments such that $\hat{\mu}_j \neq \hat{\mu}_{j+1}$; the partition is represented as a binary tree whose split points identify candidate change points, tree pruning together with model selection criteria gives their actual number.

When a time series arises from human judgments, perceptions or evaluations, whatever the quality scale employed by the expert, the corresponding items are intrinsically

accompanied by a (non probabilistic) vagueness and imprecision. We argue that, in a case such as this, instead of treating the data as either numerical or categorical, the adoption of a fuzzy scale is more expressive and accurate and it represents a proper way to take into account the above mentioned vagueness. Thus, the observed ordinal time series y_t is converted into a LR *fuzzy time series* $\widetilde{y}_t \equiv (c_t, l_t, u_t)_{LR}$, where c_t denotes the center at time t and l_t and u_t the left and right *spreads*, respectively, with the following *membership function*:

$$\mu(y_t) = \begin{cases} L(\frac{c_t - y_t}{l}) & y_t \leq c_t \quad (l > 0) \\ R(\frac{y_t - c_t}{u_t}) & y_t \geq c_t \quad (u > 0); \end{cases}$$

Based on the squared Euclidean distance of Yang and Ko (1996) we can define the deviation of the *fuzzy time series* over the entire sample period ($t = 1, \ldots, T$) as:

$$\begin{aligned} SS(\widetilde{y}_t) &= 3\sum_{t=1}^{T}(c_t - \bar{c}_t)^2 - 2\lambda\sum_{t=1}^{T}(c_t - \bar{c}_t)(l_t - \bar{l}_t) + \lambda^2\sum_{t=1}^{T}(l_t - \bar{l}_t)^2 + \\ &+ 2\rho\sum_{t=1}^{T}(c_t - \bar{c}_t)(u_t - \bar{u}_t) + \rho^2\sum_{t=1}^{T}(u_t - \bar{u}_t)^2 \end{aligned} \tag{3}$$

where $\bar{c}_t$, $\bar{l}_t$, $\bar{u}_t$ are the mean values of c_t, l_t and u_t, respectively and λ and ρ are parameters that summarize the shape of the membership function. D'Urso and Santoro (2006) proved that for deviation (3) it holds the decomposition property. Thus, we can apply ART to detect changes in mean of the fuzzified time series $\widetilde{y}_t$ using the above introduced deviation measure in the computation of splitting criterion (2).

3. Application

In this section we present the results of an application of the proposed approach to detect change points in the annual time series of Barolo wine judgments, the sample period is 1949-2010 ($T = 63$) and it is freely available at `www.podericolla.it`.

The judgments, issued by the oldest and more experienced member of the Colla family, are on a five point scale, a sort of Likert-like evaluation scale, expressed in terms of number of glasses ranging from 1 glass= *poor* to 5 glasses= *excellent*. The ordinal time series has been fuzzified using the following fuzzy coding: $\{poor = (3, 3, 1.5), fair = (4, 1.5, 1.5), good = (6, 1, 0.5), very\ good = (8, 1.75, 0.25), excellent = (10, 2, 0)\}$; the centers are depicted in Figure 1, left panel, we see that there's no graphical evidence of any change point. We have applied ART to the fuzzified series employing deviance measure (3) and setting a minimum number of observations of 5 years.

Using the BIC, whatever the value of parameters λ and ρ, we found evidence of a single break at time 46, year 1995, when, indeed, the evaluation was limited to the grapes and wines produced in the Colla farms. In Figure 1, right side panel, it is reported the

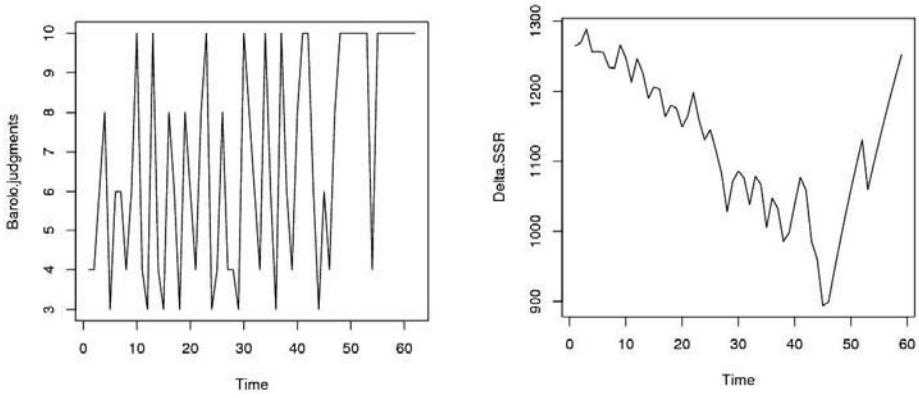


Figure 1. Time series of the centers and values of the splitting criterion

value of the splitting criterion corresponding to all possible splits of the series into two segments; as we can see a very evident minimum is reached at time 46. The results of this application suggest that the proposed approach represents a useful tool to investigate the presence of change points in ordinal time series.

References

Cappelli C., Penny R., Rea W., Reale M. (2008), Detecting multiple mean breaks at unknown points in official statistic, *Mathematics and Computers In Simulation*, 78, 351–356.

Rea W., Reale M., Cappelli C., Brown J.A. (2010), Identification of changes in mean with regression trees: an application to market research, *Econometric Reviews*, 29, 754–777.

D'Urso P., Santoro A. (2006), Goodness of Fit and Variable Selection in the fuzzy multiple linear regression, *Fuzzy Sets and Systems*, 157, 2627–2647.

Yang M.S., Ko C.H.(1996), On a class of fuzzy c-numbers clustering procedures for fuzzy data, *Fuzzy Sets and Systems*, 84, 49–60.

Hung, W.L., Yang M.S. (2005), Fuzzy clustering on LR-type fuzzy numbers with an application in Taiwanese tea evaluation, *Fuzzy Sets and Systems*, 150, 561–577.

Quaderni di Statistica
Vol. 14, 2012

Using the GME estimator with the Rasch analysis in the multigroup SEM

Maurizio Carpita
Department of Quantitative Methods, University of Brescia, Italy
E-mail: carpita@eco.unibs.it

Enrico Ciavolino
Department of Social Sciences, University of Salento, Italy
E-mail: enrico.ciavolino@unisalento.it

Summary: We develope a multigroup structural equation model (SEM) based on the generalized maximum entropy (GME) estimator, allowing integration with Rasch Analysis. The proposed method can simplify the development of the final model, with more control on the statistical properties of the manifest variables (with the Rasch Analysis) and integrating the obtained measures in the multigroup SEM by using GME estimator.

Keywords: Multigroup SEM, Generalized maximum entropy, Rasch Analysis.

1. Introduction

The GME represents a semi-parametric estimator for the SEM which works well in case of ill-behaved data or minimal distributive assumptions (Golan *et al.*, 1996, Ciavolino and Al-Nasser, 2009). We use the following two-step approach:

- In the first step, using the Rasch analysis with the Rating Scale Model, we construct and evaluate the reliability of some unidimensional measures obtained from subjective data based on likert scales;
- In the second step, with the GME estimator we consider the multigroup SEM, integrating the previous measures and taking into account their reliability indices (i.e. measurement errors).

We apply our approach to the data from a recent survey on a sample of workers employed in the Italian Social Cooperatives (Carpita and Golia, 2011).

2. GME estimator for the multigroup SEM

GME approach for the SEM considers the re-parameterization of the unknown parameters and error terms as a convex combination of expected value of discrete random variables. The coefficient matrices of classical three equations system of SEM (Bollen, 1989), $\boldsymbol{B}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}^y, \boldsymbol{\Lambda}^x$ are re-parameterized as expected values of discrete random variable with *M* fixed points for the coefficients and *N* for the errors:

$$\begin{array}{ll} \boldsymbol{B}_{m,m} = \boldsymbol{Z}^{\beta}_{m,mM} \cdot \boldsymbol{P}^{\beta}_{mM,m} & \boldsymbol{\Gamma}_{m,n} = \boldsymbol{Z}^{\Gamma}_{m,mM} \cdot \boldsymbol{P}^{\Gamma}_{mM,n} \\ \boldsymbol{\Lambda}^{y}_{p,m} = \boldsymbol{Z}^{\Lambda^y}_{p,pM} \cdot \boldsymbol{P}^{\Lambda^y}_{pM,m} & \boldsymbol{\Lambda}^{x}_{q,n} = \boldsymbol{Z}^{\Lambda^x}_{q,qM} \cdot \boldsymbol{P}^{\Lambda^x}_{qM,n} \\ \boldsymbol{\tau}_{m,1} = \boldsymbol{Z}^{\tau}_{m,mN} \cdot \boldsymbol{P}^{\tau}_{mN,1} & \boldsymbol{\epsilon}_{p,1} = \boldsymbol{Z}^{\epsilon}_{p,pN} \cdot \boldsymbol{P}^{\epsilon}_{pN,1} \qquad \boldsymbol{\delta}_{q,1} = \boldsymbol{Z}^{\delta}_{q,qN} \cdot \boldsymbol{P}^{\delta}_{qN,1} \end{array}$$

The fixed points define the support space for parameters and the error terms, in order to write both in term of expected value. To simplify the notation, in the remaining formulas, we omit the matrices and vectors subscripts. The three SEM equations can be re-formulated as a unique function model:

$$\boldsymbol{y} = \boldsymbol{\Lambda}^{y} \cdot (\boldsymbol{I} - \boldsymbol{B})^{-1} \cdot [\boldsymbol{\Gamma} \cdot \boldsymbol{\Lambda}^{x^{-s}} \cdot (\boldsymbol{x} - \boldsymbol{\delta}) - \boldsymbol{\tau}] + \boldsymbol{\epsilon} \tag{1}$$

With *m* endogenous latent variables, *n* exogenous latent variables and *p* and *q* manifest endogenous ($\boldsymbol{y}$) and exogenous ($\boldsymbol{x}$) variables. Given the re-parameterization and the re-formulation, the GME system can be expressed as a constrained non-linear programming problem. The coefficients and the error terms are estimated by recovering the probability distribution of the discrete random variables set. Applying the *vec* operator to the seven $\boldsymbol{P}^*$ matrices, we obtain the seven vectors $\boldsymbol{p}_* = vec(\boldsymbol{P}^*)$ and the vectors of probabilities for parameters and errors are defined as follow: $\boldsymbol{p}' = [\boldsymbol{p}'_B, \boldsymbol{p}'_\Gamma, \boldsymbol{p}'_{\Lambda^y}, \boldsymbol{p}'_{\Lambda^x}]$ and $\boldsymbol{p}'_E = [\boldsymbol{\tau}', \boldsymbol{\epsilon}', \boldsymbol{\delta}']$. These probabilities are calculated by the maximization of the following entropy function:

$$H(\boldsymbol{p}, \boldsymbol{p}_E) = H(\boldsymbol{p}) + H(\boldsymbol{p}_E) = -\boldsymbol{p}' \cdot ln(\boldsymbol{p}) - \boldsymbol{p}'_E \cdot ln(\boldsymbol{p}_E) \tag{2}$$

subjected to the consistency constraints, which are represented by: the re-formulated model in the equation (1) expressed as re-parameterized coefficients and errors; the normalization constraints, which guarantee the sum of each coefficient and error probability vector equal to 1.

The multigroup SEM-GME is an extention of the above model where the parameters may differ in the groups object of study. The model is defined by establishing a fit function, called H_G, as a weighted combination of the fit function defined in the equation (2) for all groups:

$$H_G(\boldsymbol{p}, \boldsymbol{p}_E) = \sum_{g=1}^{G} \frac{N_g}{N} \cdot H(\boldsymbol{p}_g, \boldsymbol{p}_{Eg}) \tag{3}$$

where, N_g is the sample size of the g^{th} group, $N = \sum_{g=1}^{G} N_g$ and $H(\boldsymbol{p}_g, \boldsymbol{p}_{Eg})$ is the entropy function for the g^{th} group. For assessing the model goodness of fit, we use the normalized entropy index:

$$S(\tilde{\boldsymbol{p}}) = -[\boldsymbol{p}' \cdot ln(\boldsymbol{p})]/[K \cdot ln(M)] \tag{4}$$

that measures the reduction in uncertainty information. The numerator $-\boldsymbol{p}' \cdot ln(\boldsymbol{p})$ is the Shannon's entropy function, as reported in the equation (2), relative only to the parameters. The quantity $K \cdot ln(M)$ represents the maximum of uncertainty, where K is the number of parameters and M is the number of fixed points. If $S(\tilde{\boldsymbol{p}}) = 0$ there is no uncertainty; if $S(\tilde{\boldsymbol{p}}) = 1$ it means total uncertainty.

3. A quality of work analysis in Italian social cooperatives

In this section we apply our approach to the data from a recent survey on a sample of workers employed in the Italian Social Cooperatives ($ICSI^{2007}$); in particular, we use a random sample of 360 workers stratified with respect to three level of education (Low, Medium and High). The SEM used for the analysis of the quality of work is defined by two endogenous (EXTRINSIC and INTRINSIC JOB SATISFACTION) and three exogenous (MOTIVATIONS, DISTRIBUTIVE and PROCEDURAL FAIRNESS) Latent Variables (LVs). The endogenous LVs are measured by two manifest variables (MVs), called $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$; the exogenous LVs are measured respectively by three, two and two MVs, called $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_7$. All the MVs are obtained by taking into account the several Likert-type scales included in the $ICSI^{2007}$ questionnaire, firstly we have construct some Rasch measures of work quality related to motivations, fairness and job satisfaction (Carpita and Golia, 2011). The corresponding Alpha reliability coefficients are used to define the bound values in the GME estimator for the SEM in way to calculate the variance of the measurement errors.

The index of uncertainty (4) for the full SEM is equal to 0.425 and the GME estimated gamma coefficients are reported in table 1.

Table 1. GME estimated gamma coefficients for the full sample SEM

LVs	MOTIVATIONS	DISTR. FAIRNESS	PROC. FAIRNESS
$INTRINSIC J.S.$	.092	.393	.695
$EXTRINSIC J.S.$	.130	.286	.483

For the full sample of workers, job satisfaction for the extrinsic aspects of job depends on workers choices (increasing their motivations increase this job satisfaction dimension) and on workers organizational identification and relationship (increasing their fairness perceptions increase this job satisfaction dimension); instead, only the two fairness perceptions affects intrinsic job satisfaction, where the gamma coefficient of the motivations is the only one statistically not different from zero.

Table 2 reports the GME estimated lambda coefficients for the full sample SEM.

Table 2. GME estimated lambda coefficients for the full SEM

MVs	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	MVs	$\boldsymbol{y}_1$	$\boldsymbol{y}_2$
$\hat{\lambda}^x$	.097	.220	.322	.441	.374	.474	.391	$\hat{\lambda}^y$	1	1
$SE(\hat{\lambda}^x)$	.207	.236	.184	.137	.151	.1417	.137	$SE(\hat{\lambda}^y)$	-	-
$var(\boldsymbol{\delta}_i)$	.325	.284	.449	.128	.251	.219	.196	$var(\boldsymbol{\epsilon}_i)$	.212	.116

The lambda coefficients measure the relationship between LVs and MVs (Rasch measures): the only not significant coefficient is that related to the $\boldsymbol{x}_1$ for motivations. Moreover, the estimated error variances are very similar to the Alpha reliability coefficients, showing the opportunity to combine the Rasch and GME approaches.

Finally, we assess the multigroup educational effect in the measurement model of SEM. The index of uncertainty (4) for the multigroup SEM is equal to 0.381, showing taking into account the group effect increase the level of fit of the model.

Table 3. GME estimated lambda coefficients for the multigroup SEM (education level)

	Low Level (.292)			Medium Level (.358)			High Level (.350)		
MV	$\hat{\lambda}^x$	$SE(\hat{\lambda}^x)$	$var(\boldsymbol{\delta}_i)$	$\hat{\lambda}^x$	$SE(\hat{\lambda}^x)$	$var(\boldsymbol{\delta}_i)$	$\hat{\lambda}^x$	$SE(\hat{\lambda}^x)$	$var(\boldsymbol{\delta}_i)$
$\boldsymbol{x}_1$	.061	.161	.138	.064	.161	.217	.103	.158	.151
$\boldsymbol{x}_2$	.305	.143	.117	.521	.192	.123	.488	.200	.080
$\boldsymbol{x}_3$	.292	.144	.182	.327	.148	.302	.291	.134	.286
$\boldsymbol{x}_4$	.432	.100	.557	.493	.114	.582	.446	.100	.571
$\boldsymbol{x}_5$	.486	.114	.281	.227	.106	.543	.420	.109	.425
$\boldsymbol{x}_6$	.387	.105	.255	.419	.109	.529	.497	.114	.363
$\boldsymbol{x}_7$	.479	.104	.309	.461	.104	.486	.408	.105	.072
MV	$\hat{\lambda}^y$	$SE(\hat{\lambda}^y)$	$var(\boldsymbol{\epsilon}_i)$	$\hat{\lambda}^y$	$SE(\hat{\lambda}^y)$	$var(\boldsymbol{\epsilon}_i)$	$\hat{\lambda}^y$	$SE(\hat{\lambda}^y)$	$var(\boldsymbol{\epsilon}_i)$
$\boldsymbol{y}_1$	1	-	.336	1	-	.457	1	-	.266
$\boldsymbol{y}_2$	1	-	.284	1	-	.395	1	-	.247

In table 3, GME estimated lambda coefficients and error variances are different between groups and therefore with respect to those obtained with the full sample SEM.

References

Bollen K. A. (1989), *Structural Equations with Latent Variables*, Wiley.

Carpita M., Golia S. (2011), Measuring the quality of work: the case of the Italian social cooperatives, *Quality & Quantity, on line first, June*, 1–27.

Ciavolino E., Al-Nasser A.D. (2009), Comparing GME and PLS methods for SEM. *Journal of Nonparametric Statistics*, 21 (8), 1017–1036.

Golan A., Judge G.G., Miller D. (1996), *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, Wiley.

Quaderni di Statistica
Vol. 14, 2012

Permutation-based control charts for ordered categorical response variables with application to monitoring of customer satisfaction

Eleonora Carrozzo Iulia Cichi Livio Corain Luigi Salmaso
Department of Management and Engineering, University of Padova, Italy
E-mail: eleonora.carrozzo@gest.unipd.it, iulia.cichi@gest.unipd.it, livio.corain@unipd.it, luigi.salmaso@unipd.it

Summary: The purpose of this paper is to extend the multivariate nonparametric control chart, called NPC chart, proposed by Corain and Salmaso (2012) to the case of ordered categorical response variables which is a situation particularly difficult to treat by traditional statistical process control methods. As confirmed by the simulation study and by the application to a real case study in the field of monitoring of customer satisfaction, we can state that the proposed NPC chart for ordered categorical response variables may represent a good alternative with respect to existing techniques.

Keywords: Permutation tests, Nonparametric combination, NPC chart.

1. Introduction and motivation

Recently Corain and Salmaso (2012) proposed the application of the NonParametric Combination (NPC) methodology (Pesarin and Salmaso, 2010) to develop a novel type of multivariate nonparametric control chart called NPC chart which has proved to be particularly effective as statistical process control (SPC) tool when the underlying data generation mechanism is non-normal in nature. The purpose of this paper is to extend the proposal of Corain and Salmaso (2012) to the case of ordered categorical response variables which is a situation difficult to treat by traditional SPC methods such as the Shewart-type charts. The NPC chart provides a flexible and effective analysis in terms both of the specification of the inferential hypotheses and of the nature of the variables involved without the need of modelling, in case of multivariate response, the dependence among variables. Design and implementation of traditional Shewart-based control charts requires the assumption that the process response distribution follows a parametric form

(e.g., normal).

2. NPC Charts for Ordered Categorical Response Variables

The theory of permutation tests and of nonparametric combination represent the methodological background from which a nonparametric multivariate control chart for ordered categorical response variables can be developed. Let be the two p-variate samples Y_0 and Y_1 related to the so-called control chart phase I and II, more specifically, denoting with n ($n > 1$) the sample size of the rational subgroup, Y_0 has to be considered as the $n_0 \times p$ pooled sample, $n_0 = n \times m$, of the m in-control samples, used to retrospectively testing whether the process was under control and where the first m subgroups have been drawn. Note that without loss of generality we are assuming that all m initial subgroups are actually in-control samples. The sample Y_1 has to be considered as one of the actual subgroups of size $n_1 \times p$, possibly out-of-control samples ($n = n_0$), used for testing whether the process remains under control when further subgroups will be drawn. Since our reference response is a multivariate ordered categorical variables, for each univariate component we consider suitable permutation test statistics:

- Multi-focus statistic (Pesarin and Salmaso, 2010): this approach suggests to decompose the categorical response variable of interest into k binary variables each one related to one category of the response; in this way it is possible to refer to a further decomposition of the null univariate sub-hypothesis $H_{0(jh)}$ into k additional sub-hypothesis each one suitable for testing the equality in distribution of each one of the k category of the ordered categorical response variable; this is done by taking into account a set of k Chi-squared based tests calculated from k 2×2 contingency sub-tables to be then combined into a final statistic;
- Anderson-Darling statistic (Pesarin and Salmaso, 2010):

$$T^*_{AD} = \sum_{r=1}^{k-1} \left(N^*_{hr} - N^*_{jr}\right) \left[2\frac{N_{\cdot r}}{n}\left(\frac{2n - N_{\cdot r}}{2n}\right)\frac{n^2}{2n-1}\right]^{-\frac{1}{2}},$$

  where $N_{\cdot r} = N_{jr} + N_{hr} = N^*_{jr} + N^*_{hr}$ are the observed and the permutation cumulative frequencies in which $N^*_{sr} = \sum_{q \leq r} f^*_{sq}, r = 1, \ldots, k-1, s = j, h$, and f_{sq} is the frequencies of the s-th treatment for the q-th category of the response.

Within our approach we do not need to estimate any parameters of the in-control process. This is consistent with the usual rationale behind the nonparametric permutation approach, i.e. the in-control pooled sample Y_0 plays the role of reference dataset to be kept out and recursively compared with Y_1, which is one of the subgroups to be tested in the future. This conceptual framework leads to a procedure where we consider a sequence of independent multivariate two-samples hypotheses testing problems in case of

unbalanced designs ($n_0 > n_1$). Some important remarks have to be pointed out: the control limit for the multivariate control chart at the desired α-level has to be simply calculated as the $(1-\alpha)$ quantile of the null permutation distribution of the multivariate combined test statistic T'' (for details on how to perform a multivariate permutation test via nonparametric combination methodology see Pesarin and Salmaso, 2010); noting that necessarily the limits will differ for any given subgroup under testing. Finally, we are implicitly assuming that the process parameters are unknown but the extension to the case of known parameters is straightforward; in fact, this case may reduced to the so-called multivariate one-sample problem where a combination-based solution already does exist (Pesarin and Salmaso, 2010). In this work we focus on the case of unknown limits, the most interesting for real applications.

3. Simulation study

In order to validate the proposed NPC Chart and to evaluate its relative performance when compared with a traditional multivariate and control chart such as Hotelling T^2 and X-bar, we carried out a suitable Monte Carlo simulation study. The real context we are referring to is a typical customer satisfaction study where a group of 20 people provides their evaluations by a Likert 1-5 rating ordinal scale, where we suppose that the 0.5 scores are admitted as well. Note that we are actually considering a 9 point ordered categorical response variable. Let us consider the following simulation setting:
- number of response variables: $p = 5$; where the number of active variables (under the alternative hypothesis) was 1, 2 and 4 (three setting), more precisely while in all cases $\mu_0 = [0, 0, 0, 0, 0]$, under H_1 the mean values were set for the three settings respectively as $\mu_1 = [0, 0, 0, 0, 1]$, $\mu_1 = [.5, 0, 0, 0, 1]$ and $\mu_1 = [.5, .5, 1, 1, 0]$;
- two type of multivariate distributions for random errors: Normal and a moderate heavy tailed and skewed distribution (with kurtosis and skewness both equal to 1). In order to guarantee an ordered categorical response variable we rounded the continuous error to the nearest 0.5 value;
- two type of variance/covariance matrices: i. I_p (identity matrix, i.e. the case of independence where $\sigma_{jh} = 0, \forall j, h = 1, ..., p$) and Σ_p is such that each univariate random component has $\sigma_j^2 = 1, \forall j = 1, ..., p$ and $\sigma_{jh} = 0.4$.

The performance of univariate and multivariate NPC Charts has been evaluated in term of Average Run Length - ARL, i.e. the average number of samples needed before to get the first out-of-control (reject the null hypothesis that a truly out-of-control process is under control). Simulations are designed so that for each of the 1,000 data generations 5 independent samples of size $n = 20$ are created and the control charts are applied until the first out of control is reported. The in-control pooled dataset Y_0 was generated by $n_0 = n \times 2$ random values ($m = 2$). Main results of the simulation study (not reported here for the sake of space) are available at http://www.gest.unipd.it/ salmaso/.

4. Application to Monitoring of Customer Satisfaction

The proposed permutation-based control chart for ordered categorical variables has been applied to a case study in the field of monitoring of customer satisfaction called 'SESTO' (Statistical Evaluation of a Skischool from Tourists Opinions) which is the first Italian survey on the evaluation of Ski Instructors and it is a pilot study performed in the Sesto Ski School (Italian Alps, north Italy). Nine satisfaction variables towards different aspects of ski teaching have been evaluated in a rating scale 1-10 and the application of NPC Chart allowed us to monitor the customer satisfaction during the ski season, week by week. Results shows that both at a univariate level and globally the NPC Chart is effective in detecting weak and strength aspects of ski teaching.

5. Conclusion and future research

As confirmed by the simulation study and by the real case study, we can state that the proposed NPC chart for ordered categorical response variables may represent a good alternative with respect to existing techniques. Furthermore, control charts based on NPC can manage with any dependence relation among variables and any kind of variable (even mixed variables with possible presence of missing observations). An important property which can be very useful in real applications is represented by the finite sample consistency of NPC-based tests (see Pesarin and Salmaso, 2010) which can help in gain power keeping fixed the sample size and increasing the number of informative variables.

Acknowledgements: Authors wish to thank the University of Padova (CPDA092350/09) and the Italian Ministry for University and Research MIUR project PRIN2008 -CUP number C91J10000000001 (2008WKHJPK/002) for providing the financial support for this research.

References

Corain L., Salmaso L. (2012), Nonparametric Permutation and Combination-based Multivariate Control Charts with Applications in Microelectronics, Working Paper no. 12-2011, Dept. of Management and Engineering, Univ. of Padova, submitted for the publication on *Applied Stoch. Model in Business and Industry*.

Pesarin F., Salmaso L. (2010), *Permutation tests for complex data: theory, applications and software*, Wiley, Chichester.

Quaderni di Statistica
Vol. 14, 2012

Non parametric models for credit rating assessment

Paola Cerchiello Paolo Giudici
Department of Economics and Management, University of Pavia
E-mail: paola.cerchiello@unipv.it, giudici@unipv.it

Summary: In this contribution we propose to estimate the probability of financial default of companies and the correlated rating classes, using efficiently the information contained in different databases. We want to classify companies according to the target variable in a supervised way. Our approach is ordinal: covariates induce partitioning of companies, on the basis of their measurement levels, when categorical, or to their quantiles, when they are continuous. We were provided by an Italian bank with a database containing 13 covariates regarding both micro-economic and macro-economic information to test the usefulness and efficiency of our approach.

Keywords: Credit Risk, Rating classes, Bayesian variable averaging.

1. Background

The financial meltdown of 2008-2009 questioned the validity of risk models and their practical implications. Within the framework of existing regulatory models, Basel II and III, banks have a tendency to uniform their models of risk evaluation and enterprise funding generating a pro-cyclical approach that in the current economic and financial environment highlights and exacerbates the difficult conditions in which the firms operate.

The high dimensional data available from public financial statements make credit analysis difficult and the problems are exacerbated by the necessity to account jointly for qualitative and quantitative data. In order to improve empirical results and to obtain credit scores that are more predictive and less procyclical, research is needed in the area of "scoritisation" of ordinal variables and in variable selection, preliminary to the inclusion in a full Bayesian model averaging perspective. One possibility is to follow what recently proposed by Cerchiello et al (2011) that suggest to employ stochastic dominance and quantile-based indicators to summarise ordinal variables in a scoring in-

dicator. A further need in modelling of microeconomic credit risk data is to take into account interdependencies between risk variables, and their causal factors; one possibility is to employ Bayesian network modelling as suggested in Bonafede et al. (2007).

2. Proposal

In this contribution we propose to estimate the probability of financial default of companies and the correlated rating classes, using efficiently the provided information typically contained in several and not homogeneous databases.

We want to classify companies into groups (that are rating classes) in a supervised way. Such groups to comply with BASEL II requirements, have to be: homogeneous with regard to target variable (that is default- not default), order preserving (that is ordering ability) and stable with regard to horizon time. In this context we are typically provided with databases of various origin, often not transparent and made of qualitative and quantitative variables. Our proposal is to build, effective but easy to explain, ordinal rating models integrated by means of Bayes theorem.

The model we propose, can be essentially described as follows:

$$E(\theta_i \mid \underline{X}) = \sum_{k=1}^{K} E(\theta_j|\underline{X}, g_k) \cdot p(g_k, \underline{X}) \tag{1}$$

where g_k is a partition induced by each covariate k (to be combined) that classify each unit i into one and only one level j. The model is elaborated by using a non parametric framework based on mixture of product of Dirichlet processes.

We apply our method to a real database provided by an Italian bank and we are able to select the most important covariates in terms of predicting power with regards to the target event: default or not default.

More precisely, we move from Giudici et al. (2003) that proposed a mixtures of products of Dirichlet process in the survival analysis context in order to compare the explanatory power of each available covariate. In this paper, we cope with a completely different framework but we borrow the general idea adapting it to the credit rating assessment issue.

If we consider a collection of n companies, let be Y a random variable distributed as a Bernoulli variable representing the default (θ) or not default event of the $i-th$ company. Let θ parameter be distributed as a Beta r.v. with parameters α and β. Given un unknown partition of the $i-th$ company, we get that the marginal likelihood is:

$$p(\underline{x}|g_k) = \prod_{j=1}^{J} \frac{M^{d_j}}{M^{[d]}} \prod_{r=1}^{R} (n_{j(r)} - 1)! \times \left[\frac{\beta}{\alpha+\beta} I_{[0,1)}(x_j) + 1 I_{[1]}(x_j) \right] \tag{2}$$

where M is a known precision measure, d_j is the number of default companies in each

level j of the $k-th$ covariate, $n_{j(r)}$ is the number of distinct observations with regard to the target variable (in non descending order) in level j, α and β are the parameters of the Beta r.v. and $I(\cdot)$ is the indicator function.

The conditional mean is:

$$E(\theta_j|\underline{X}, g_k) = \frac{M}{M+n_j}\left[\theta I_{[0,1)}(x_j) + 1I_{[1]}(x_j)\right] + \frac{n_j}{M+n_j}\hat{F}_{Beta} \qquad (3)$$

where n_j is the number of companies in the $j-th$ level of the covariate j and $\hat{F}_{Beta}$ is the empirical cumulative distribution function of the Beta r.v.

3. Empirical Evidence and Remarks

In order to evaluate the performance of the proposed non parametric model, we were provided from an Italian bank with a database containing a list of variables on a set of small and medium enterprisers. The set is made of 1000 companies and 14 variables: a target one describing the default or not default event; the others on an ordinal measurement. Those 13 variables can be divided into two subset: the first containing 4 quantitative variables, giving information on the economic context, are measured on a ordinal scale made of 9 levels. These variables are: 'Ai' describing the banking transactions,'Cr' from historical defaults database, 'Dir' on macro-economic scenarios and 'Cebi' for balance sheets information. On the qualitative side we have 9 ordinal variables characterized by a 4 levels measurement, coming from the internal questionnaire compiled when a loan greater than 30.000 euro is requested. Those variables give information either on the historical relation between the company and the bank (if there exists) or on the management and structure of the company itself. The final aim is to predict the probability of default of each company given the available covariates described above. In Table 1 we report the marginal likelihoods of the 13 covariates in order to evaluate their importance with regards to the target variables. It clearly appears that the highest marginal likelihood is obtained by item 3, item 1 and item 4 from the questionnaire, instead the 4 economic covariates seem not to be relevant. Moreover, we remark that the analysis has been carried out by varying the sensitivity parameter M, but with no significant impact on the final results. The three items from the questionnaire regard: competitive positions of the company in the market (item 1); payment of the furnishers (item3); the number of clients which the 50% of the sales figures refer to (item 4). Thus it is reasonable to assume that company specific characteristics are more relevant for the default event rather than macro-economic figures.

Finally we use formula 2 and 3 according to formula 1 to obtain the probability of default for each companies. From the distribution of the probabilities of default on the 1000 available companies we derive 9 rating classes by considering the 9-th quantile of that distribution. Such method allows us to guarantee the order preserving property requested by Basel II accord.

Table 1. Marginal Likelihood of analyzed covariates

Covariates	Non parametric
Ai	0.11e40
Cr	0.38e40
Dir	0.20e40
Cebi	0.55e40
Item 1	1.10e66
Item 2	8.40e56
Item 3	5.49e67
Item 4	1.11e66
Item 5	3.66e60
Item 6	3.08e53
Item 7	2.08e63
Item 8	9.46e54
Item 9	1.68e56

Moreover, an aspect that we want to further investigate, is the impact of the measurement levels of the two sets of covariates on the final results. We shall deepen the analysis by homogenizing the categories of the covariates. A possible approach will be based on the application of an ordinal index based on the quantile of the distribution or on the cumulative distribution function itself as in Cerchiello et al. (2010).

References

Bonafede E., Cerchiello P., Giudici P. (2007), Statistical models for Business Continuity Management, *Journal of Operational Risk*, 2(4), 79–96.

Cerchiello P., Dequarti E., Giudici P., Magni C. (2010), Scorecard models to evaluate perceived quality of academic teaching, *Statistica & Applicazioni*, 2, 145–156.

Giudici P., Mezzetti M., Muliere P. (2003), Mixtures of products of Dirichlet processes for variable selection in survival analysis, *Journal of Statistical Planning and Inference*, 111, 101–115.

Quaderni di Statistica
Vol. 14, 2012

Multivariate permutation and combination-based composite indicators with application to the evaluation of indoor environment

Livio Corain Luigi Salmaso
Department of Management and Engineering, University of Padova, Italy
E-mail: livio.corain@unipd.it, luigi.salmaso@unipd.it

Valeria De Giuli Roberto Zecchin
Department of Industrial Engineering, University of Padova, Italy
E-mail: valeria.degiuli@unipd.it, roberto.zecchin@unipd.it

Summary: The aim of this work is extend the methodology of multivariate permutation and combination-based composite indicators to the case of an observational study where measures of interest are typically multivariate ordered categorical variables. The proposed solution is applied to a real case study related to the evaluation of indoor environment.

Keywords: Environmental quality, Permutation tests, Nonparametric combination.

1. Introduction

When the response variable is multivariate in nature, the need to define an appropriate composite indicator related to items of interest such as products, services, teaching courses, degree programs, and so on is very common in both experimental and observational studies. From the methodological point of view, the problem of defining a suitable composite indicator related to several multivariate populations is dealt with by referring to a so called ranking parameter, that is a suitable aspect of the populations, such that the rank transformation of that aspect, which may be able to provide a meaningful ranking of the populations from a multivariate point of view (Arboretti et al., 2010). Depending on the assumptions made about the random errors, the distribution of ranking parameter estimators can be derived in a parametric or nonparametric way. However, the parametric approach presents a number of drawbacks and limitations; conversely, thanks to

its robustness and flexibility, the permutation and combination-based approach appears to be more reliable and powerful. The aim of this work is extend the methodology of multivariate permutation and combination-based composite indicators to the case of an observational study where measures of interest are typically a multivariate ordered categorical variable. Such real case study is related to the evaluation of indoor environment. Indoor environmental comfort is, together with the project of an efficient building, the main purpose to reach, especially for commercial and educational buildings where it has been deeply demonstrated that a comfortable building makes people feel satisfied and therefore their productivity and learning increase.
This research presents part of the environmental analysis carried out in three Italian primary schools (X, Y, Z) that have been monitored from February to June 2011. The study involved 8 classrooms with around 160 pupils. The aim of the study is to apply a suitable statistical method able to effectively summarize all the information collected from the survey to classify school buildings in terms of indoor environmental quality. The survey has been distributed three times (in February, April and in the end of May), to see whether seasonal variations affect children impact on environmental perception, but mostly to make pupils friendly to this kind of approach. Before starting with statistical evaluation, a specific methodology has been used to impute missing data: the multiple imputation and propensity score method (Shafer, 1999) implemented in the software Solas (Horton and Lipsitz, 2001) so that three dataset have been imputed. These three dataset have been then analyzed with the nonparametric combination (NPC) methodology (Pesarin and Salmaso, 2010): a C-related test has been carried out in order to check if significant differences could be found along the three time evaluations. The result confirms the hypothesis that pupils became friendly with the survey and that the third time evaluation might be the more reliable one. Finally, a multivariate permutation and combination-based composite indicators analysis has been applied to quantify the differences among schools and to rank the schools from different multidimensional aspects, that is environmental quality, interaction user-building and indoor environment.

2. Main results on global rankings to classy three schools

A new multivariate ranking procedure extending the results from Corain and Salmaso (2007) and Arboretti et al. (2010) has been applied to four different domains of items (building-related factors, psychological factors, interaction building-occupant, frequency of discomfort and lighting and acoustic quality); moreover the aspects connected to the interaction between building and occupant have been considered also separately, therefore eight rankings have been computed. Building related factors deal with room interior, amount of space, classroom appearance, etc.), while psychological factors investigate on feeling sensation of pupils towards the school and the classmates. Table 1 shows how the three schools are ranked according to the aspects listed above: ”1” refers to the school placed first and so on. School Z shows to be the worst school in terms of

building-related and psychological factors and of frequency of discomfort: it is the one in which shadings are more operated and windows more opened, just because its indoor conditions need to be changed as a consequence of unacceptable conditions. School Y resulted to be the most comfortable building, especially for thermal and visual comfort: this school registered the lowest temperature profiles and it is efficiently shaded by solar radiation, therefore no problems connected to glare are present. No significant differences were found for thermal comfort in wintertime and for air drafts, which also do not represent a discomfort: in fact, around the 70% of pupils feel cold rarely during winter and around the 90% do not perceive air drafts. On the contrary, in the summer period, the 70% of pupils in school Z complain about thermal sensation, while only 24% in school X and 30% in school Y. Visual discomfort connected to solar radiation represents a huge problem for school Z, where 52% of pupils complain about that, while only the 29% in school X and only the 9% in school Y. Artificial lighting does not represent a relevant discomfort in any school, except for school Z where someone complains about that.

Table 1. Global ranking of schools with respect to each domain.

domain	school X	school Y	school Z
building-related factors	1	1	3
psychological factors	1	2	2
interaction building-occupant	1	3	2
switching of light	1	3	2
air changing	1	1	1
shadings operation	1	3	1
frequency of discomfort	2	1	3
lighting and acoustic quality	1	1	1

School Y turns to be the best school from environmental comfort point of view, while it is the school in which pupils rarely interact with the building (shadings, lighting and windows opening): this confirms the fact that it is the building that requires the less environmental changing, just due to the more comfortable indoor conditions.

3. Conclusions

Our case study emphasized the role of multivariate nonparametric ranking methods in assessing a classification of objects. In the context of ranking multivariate populations the parametric approach presents a number of drawbacks that have to be taken into account. First of all, when keeping the sample size fixed, increasing dimensionality (number of variables) results in a loss of degrees of freedom, hence the estimation procedures may become inaccurate. Moreover, under non normal errors, inferential achievements are valid solutions only asymptotically, hence for finite samples (often very small

in many real applications) the approximation accuracy has to be carefully considered along with validity of results. Conversely, a permutation-based nonparametric inference may offer a number of advantages:

- it is a robust solution, with respect to the true underlying distribution of response variables;
- the dependence structure of the response variables is implicitly captured, so there is no need to estimate any dependence coefficient or hypothesize a dependence model;
- it can be used with whatever ranking parameter, arbitrarily complex although its expression could be very complex and its distribution could be virtually impossible to derive parametrically.

A specific web-based software performing such procedures have been implemented and it is available at http://stat.dft.unipd.it/globalranking/login.aspx.

Acknowledgements: Authors wish to thank the University of Padova (CPDA092350/09) and the Italian Ministry for University and Research MIUR project PRIN2008 -CUP number C91J10000000001 (2008WKHJPK/002) for providing the financial support for this research.

References

Arboretti Giancristofaro R., Corain L., Gomiero D., Mattiello F. (2010), Nonparametric Multivariate Ranking Methods for Global Performance Indexes, *Quaderni di Statistica*, 12, 79–106.

Corain L., Salmaso L. (2007), A nonparametric method for defining a global preference ranking of industrial products, *Journal of Applied Statistics*, 34, 2, 203–216.

Horton N. J., Lipsitz S.R. (2001), Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables, *The American Statistician*, 55, 3, 244–254.

Pesarin F., Salmaso L. (2010), Permutation tests for complex data: theory, applications and software, Wiley, Chichester.

Schafer J.L. (1999), Multiple imputation: a primer, *Statistical Methods in Medical Research*, 8, 3–15.

Quaderni di Statistica
Vol. 14, 2012

A statistical analysis of consumer perception of wine attributes

Marcella Corduas
Department TEOMESUS, University of Naples Federico II
E-mail: corduas@unina.it

Luciano Cinquanta Corrado Ievoli
Department STAAM, University of Molise
E-mail: cinquant@unimol.it, ievoli@unimol.it

Summary: The importance of extrinsic and intrinsic attributes of wine for purchase decision is object of a lively debate. This work provides some insights into the problem by analyzing data from a survey on consumers' perceptions using CUB models.

Keywords: Ordinal data, CUB model, Consumer perceptions, Food quality.

1. Introduction

As for other food products wine is characterized by so called intrinsic attributes, relating to taste, color, aroma, and extrinsic attributes such as labeling information, price, region of origin, brand, packaging. The purchase decision is, therefore, originated by an inferential process which is aimed to assess wine quality from the extrinsic attributes. All the other features can, in fact, be judged only during consumption. Of course, some of those attributes are under producers' control who plan strategic and product policy. In this regards the literature is wide ranging from studies on the consumers' perception of wine attributes in different countries to investigations of relative importance of intrinsic and extrinsic features (for instance, see Martinez-Carrasco *et al.* 2006, Goodman 2009, and, with specific reference to the case of Italian wines, Hertzberg *et al.* 2008, Lai *et al.* 2008, Casini *et al.* 2009). This contribution discusses how CUB models (Corduas *et al.* 2009) can be applied in order to identify: i) fundamental elements which affect purchase decisions, ii) significant similarities and differences in the overall judgements expressed by raters on various attributes, iii) the dependence of ratings from consumers' profile.

2. The methodology

We briefly recall that CUB model describes preferences or ratings by a random variable R such that:

$$P(R=r) = \pi \left(\begin{array}{c} m-1 \\ r-1 \end{array} \right) (1-\xi)^{r-1} \xi^{m-r} + (1-\pi)\frac{1}{m}, \qquad r = 1, 2, ..., m \quad (1)$$

where $\xi \in [0,1]$, $\pi \in (0,1]$ and $m > 3$ is the number of modalities for evaluating an item. The parameter π determines the role of *uncertainty* in the final judgment: the lower the weight $(1-\pi)$ the smaller the contribution of the Uniform distribution in the mixture. The parameter ξ characterizes the shifted Binomial distribution and $(1-\xi)$ denotes the positive/negative feeling that the rater has towards the object. Corduas et al. (2009) review various statistical properties and several generalization of the model. For our aims, in the next section, the clustering of estimated CUB distributions is performed by testing the homogeneity hypothesis of rating distributions by Kullback-Liebler (KL) divergence and producing the subsequent classification using the BEA algorithm (Corduas, 2011). In addition, the effect of covariates on raters' judgements is estimated relating the CUB parameters to significant covariates by means of a logistic link function.

3. The empirical study

The study refers to a sample of 192 consumers. In order to increase the range of competence and the level of knowledge about wine, half of the respondents were randomly selected among visitors at Vitigno Italia and Vinitaly during the general admission days. Each interviewee was asked to rate the importance of 13 wine attributes in determining his/her purchase decision. The rates were expressed on a 7 point Likert scale.

The representation of the estimated CUB models in the parameter space shows that consumers assess the items with a different degree of uncertainty (horizontal axis). They have a clear and precise opinion about the importance that they attach to the "wine complexity and taste", the "aroma/bouquet", the "quality-price ratio", the "region of origin" and the "food-pairing" whereas they are more uncertain when they rate the wine producer, the packaging and the protected geographical status (Figure 1). Moreover, the items appear to be well separated with respect to the degree of importance in the purchase decision (vertical axis): the "brand name" and the "bottle shape" are considered rather unimportant with respect to the remaining items.

The rating distributions of the considered wine attributes are clustered in 4 groups (Figure 2): G1 ("producer", "wine label information", "alcoholic degree", "color", "drink pleasantness"); G2 ("food-pairing","aroma/bouquet","wine complexity","quality/price ratio"); G3 ("grape variety" and "region of origin"); G4 ("bottle shape","brand name and label appearance"). The graphs allow to capture the different emphasis that consumers put on the various items moving from the cluster of most influential attributes

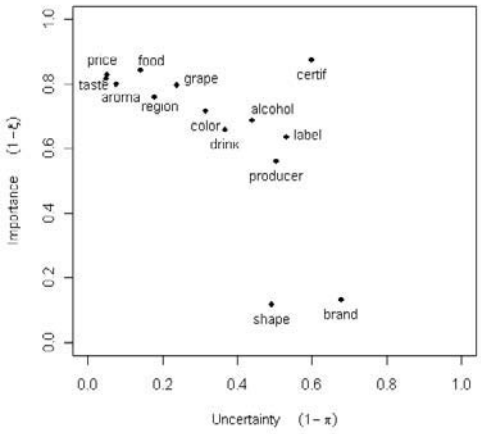


Figure 1. Estimated CUB models

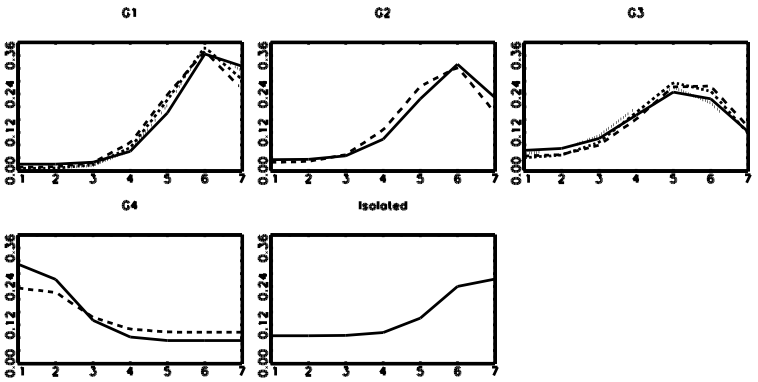


Figure 2. Clustered CUB models (by rows: G1, G2, G3, G4, isolated item)

(G1) to that of least important ones (G4). As far as the believed importance increases the uncertainty in the responses decreases. The "protected geographical status", instead, appears isolated with respect to all the other items because of the great uncertainty that the raters expresses.

Finally, differences in ratings due to consumers' profile has been investigated introducing some socio-demographic variables, the consumption occasions and frequency, a self-assessed measure of competence as covariates in the CUB models. For sake of space, the following table summarizes part of the significant relationships which have been revealed by the estimated CUB models.

Table 1. The effect of some covariates on the importance of wine attributes

Covariate	Item
expertise level	producer, alcoholic degree, aroma, wine complexity, quality/price ratio, grape variety, region of origin, geog. protected status
consumpt. frequency	drink pleasantness
consumpt. occasion	color, grape variety, region of origin, geog. protected status, bottle shape, brand, quality/price ratio
place of purchase	aroma, wine complexity, qual/price ratio, grape variety, region of origin

Acknowledgements: This work has been partly supported by MIUR-PRIN2008 grant (CUP n.E61J10000020001) of the Research Unit at University of Naples Federico II.

References

Casini L., Corsi A. M., Goodman S. (2009), Consumer preferences of wine in Italy applying Best-Worst scaling, *Int. Journal of Wine Business Research*, 21,1, 64–78.

Corduas M., Iannario M., Piccolo D. (2009), A class of statistical models for evaluating services and performances, in M. Bini, P. Monari, L. Salmaso (Eds), *Statistical methods for the evaluation of educational services and quality of products*, Physica-Verlag, Heidelberg, 99–117.

Corduas M. (2011), Assessing similarity of rating distributions by Kullback-Liebler divergence, in B. Fichet, D. Piccolo, R. Verde, M. Vichi (Eds), *Classification and Multivariate Analysis for Complex Data Structures*, Springer-Verlag, Heidelberg, 221–228.

Goodman S. (2009), An international comparison of retail consumer wine choice, *Journal of International Wine Business Research*, 21, 1, 41–49.

Hertzberg A., Malorgio G. (2008), Wine demand in Italy, an analysis of consumer preferences, *New Medit*, 4, 40–44.

Lai M. B., Del Giudice T., Pomarici E. (2008), Unobserved heterogeneity in the wine market, an analysis of Sardinian wine using Mixed Logit, *A.A.W.E. WP*, 28.

Martínez-Carrasco L., Brugarolas M., Del Campo F.J., Martínez A. (2006), Influence of purchase place and consumption frequency over quality wine preferences, *Food Quality and Preference*, 17, 315–327.

Quaderni di Statistica
Vol. 14, 2012

Purchasing in European Union: a multilevel latent class application

Chiara Dal Bianco
Department of Economics, University of Padua
E-mail: chiara.dalbianco@unipd.it

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

Roberta Varriale
ISTAT - Italian National Statistical Institute
E-mail: varriale@istat.it

Summary: This work aims at investigating similarities and differences in the ways of purchasing goods and services by the European citizens, in particular the consumer behaviour on the preferred purchasing channels among web, phone, mail and sales representatives, exploiting data collected by the Eurobarometer 69.1 survey in 2008. To this aim, we adopt a multilevel latent class solution, which allows to simultaneously cluster individuals and countries. The overall result is that most countries can be grouped in classes that follow a geographical division, while the European citizens can be divided in classes with some specific profiles.

Keywords: Consumer behaviour, Multilevel latent class, Segmentation.

1. Introduction

The ongoing political process of European unification, the introduction of Euro currency and new developments in information and communication technology have contributed to create a potential, large single market, the European Union. In spite of this, European Union (EU) countries are clearly separated each other according to national, cultural and economic dimensions. Treating the different member states as a unique market can lead to plan some attractive (because of economies of scale) unified market-

ing strategies. Nevertheless, the identification of substantial differences among countries (because of differences in economic structures, cultural identities or regulations) might support the use of multi-domestic or multi-regional policies. In such a complex context, international segmentation should combine the benefits of standardization (lower costs and better quality) with the benefit of adaptation (closeness to the needs of consumers), by structuring the heterogeneity that exists among countries and units (individuals, firms, etc.) in order to identify relatively homogeneous groups of countries and/or units (Steenkamp and Ter Hofstede, 2002).
Using data collected in 2008 by the Eurobarometer 69.1 survey, the aim of this work is to investigate similarities and differences in the ways of purchasing goods and services in the last 12 months by the European citizens, adopting a multilevel Latent Class (LC) analysis as the segmentation tool. In particular, the focus is on the consumer behaviour concerning the national and international preferred purchasing channels among web, phone, mail and sales representatives within the European member countries.
Eurobarometer consists in a series of surveys regularly performed on behalf of the European Commission. In the wave 69.1 of these surveys, 26746 citizens of the 27 EU member countries were interviewed: purchasing in the EU was one of the major areas of interest. There is evidence of different utilization of the purchase channels across countries and across age groups. Some channels are evaluated as risky (for instance, because of frauds) by most consumers.
The statistical solution adopted in this work is briefly introduced in the next section. The main results and conclusions of our analysis are reported in Section 3 and 4 respectively.

2. The model

The approach applied in this paper is the multilevel Latent Class analysis (Vermunt, 2003). On the one hand, LC analysis has been recently introduced in the literature as a model-based tool for regular market and international segmentations. On the other hand, the multilevel solution may take into account the dependencies between the lower-level units resulting from the hierarchical data structure (individuals nested into countries). Hence, the application of multilevel LC models typically aims at obtaining simultaneously country segmentation and cross-national consumer segmentation.
Let $i = 1, \ldots, I$ be the international sample of consumers, living in one of the 27 European member countries, denoted by $j = 1, \ldots, 27$. Let $\ell = 1, \ldots, L$ be the investigated purchasing channel. Thereby, $Y_{ij\ell} = 1$ if individual i of country j purchased a good or service through the channel $\ell = 1$ (for example, via web within the national boundaries) in the last 12 month, 0 otherwise. Let Y_{ij} and Y_j be the full vector of responses (channels) of individual i in country j and the full set of responses of country j respectively. Let M be the number of higher-level LCs, C be the number of lower-level LCs. Consumer and country segment membership is represented by the discrete latent variables X_{ij} and X_j^g respectively; Z_{ij} represent the concomitant variables at the consumer level.

At the individual level, the model specifies the conditional probabilities of purchasing through channel ℓ for individual i in country j, given that country j belongs to LC x^g:

$$P(\mathbf{Y_{ij}} = \mathbf{y_{ij}}|X_j^g = x^g) = \sum_{x=1}^{C} P(X_{ij} = x|X_j^g = x^g) \prod_{\ell=1}^{L} P(Y_{ij\ell} = y_{ij\ell}|X_{ij} = x, Z_{ij} =$$

The purchases through different channels are assumed to be independent, conditional to the consumer latent class membership and concomitant variables Z.
At the country level, the model specifies the marginal probabilities of purchasing channel for a country j (the observations of the n_j respondents in each country j are assumed to be independent of one another given the country LC membership):

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{x^g=1}^{M} P(X_j^g = x^g) \prod_{i=1}^{n_j} P(\mathbf{Y_{ij}} = \mathbf{y_{ij}}|X_j^g = x^g)$$

Combining the two equations we obtain the following formulation:

$$\begin{aligned} P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{x^g=1}^{M} [P(X_j^g = x^g) \prod_{i=1}^{n_j} [\sum_{x=1}^{C} P(X_{ij} = x|X_j^g = x^g) \cdot \\ \cdot \prod_{\ell=1}^{L} P(Y_{ij\ell} = y_{ij\ell}|X_{ij} = x, Z_{ij} = z)]] \end{aligned} \tag{1}$$

which states the three components (each of them modeled as logit equations) of the model: the probability that country j belongs to a particular country segment; the probability that individual i belongs to a particular respondent segment; the probability that individual i purchases via the channel ℓ, given his/her segment membership.

3. Results and Main Conclusions

Model expressed in equation (1), including the effect of some covariates by means of concomitant variables (age, gender, education, occupation, household size, area of residence) is estimated by using the Latent GOLD software (Vermunt and Magidson, 2008). After the comparison of different model specifications (different number of consumer and country segments) using the information criteria CAIC, the solution with 7 country segments and 9 consumer segments is chosen.
The first level-1 class has low probability for all purchase channels (the *no-purchasers*). In the second class there is a high probability of using the web for national purchases, in particular among students and workers with high education (*national web purchasers*). An attractive segment is the third one: it has a high probability of having middle-age women who buy by mail (*national mail purchasers*). Consumers of the fourth class are

called *potential purchasers*. On the one hand, this segment has low probability for all purchase channels, like the *no-purchasers* class. On the other hand, the components are more likely to be similar to the *national web purchasers* class. The *national purchasers* describes the fifth segment of consumers, who have high probability for national buying with all channels. The sixth class (*in person purchasers*) is more likely to be composed by individuals who buy from sales representatives. Individuals who purchase via internet, both within and outside the national boundaries (*web purchasers*) have high probability of being included in the seventh class. The last two classes have high penetration rates for purchasing only outside the national boundaries (*international purchasers*) and purchase via all the aforementioned channels (*general purchasers*). Age and education are the main determinants across all classes.

Using the empirical Bayes modal prediction, the countries have been assigned to one of the seven segments. Mediterranean countries, Romania and Bulgaria belong to the first segment, that is mainly composed by *no-purchasers*. The second class is characterized by a wide use of purchase via mail (Belgium, Slovenia, Czech Republic, Baltic countries). The third class is made up of central Europe countries, in which national and web purchasers are the most common level-1 classes. The *international purchasers* are mostly present in the fourth segment, like in Luxembourg and Malta. Denmark, the Netherlands and Sweden belong to the fifth segment, which is characterized by the high use of internet purchases. The sixth segment shows an interesting geographic proximity of the countries (Poland, Slovakia and Hungary), where the most common channels are mail and sales representatives. Austria belongs to the last (country-specific) segment, which shows an interesting mix of *no-purchasers* and *national mail purchasers*.

Summarizing, multilevel LC analysis provides a new and powerful tool to identify market segments in target marketing. The overall finding is that it is difficult to treat the different European member states as a unique market and most countries can be grouped in classes that follow a geographical division. Then, European consumers can be divided in classes with some interesting profiles. On the one hand, these findings could be used by policymakers to promote and guarantee similar service levels across all European citizens. On the other hand, our results could be helpful for companies to develop and implement suitable cross-country strategies.

References

Steenkamp J-B.E.M., Ter Hofstede F. (2002), International market segmentation: issues and perspectives, *Internation Journal of Research in Marketing*, 19, 185–213.

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

Vermunt J.K., Magidson J. (2008), *LG-Syntax Users Guide: Manual for Latent GOLD 4.5 Syntax Module*, Statistical Innovations Inc., Belmont (MA).

Quaderni di Statistica
Vol. 14, 2012

Bayesian covariate selection in CUB model: some considerations

Laura Deldossi Roberta Paroli
Department of Statistical Science, Università Cattolica del Sacro Cuore, Milano
E-mail: laura.deldossi@unicatt.it, roberta.paroli@unicatt.it

Summary: The aim of the paper is to analyze and to compare the performance of some of the most popular Bayesian variable selection methods adapted to the CUB(p, q) model. Here the variables selection mechanism is based on the marginal posterior inclusion probability, instead of the highest posterior probability model. Following this kind of procedure for the same simulation cases analyzed in Deldossi and Paroli (2011), we show that the ability of this criteria in selecting the true covariates is in general very appealing since the marginal posterior inclusion probabilities of the true covariates are often greater than 0.8. Furthermore our aim is to compare the various approaches exploring higher complex CUB models, with an increasing number of predictors and for different orders p and q.

Keywords: Ordinal data, MCMC sampling methods, Variables selection methods.

1. Introduction

The CUB model for ordinal data has been recently introduced to analyse the consumer preferences with regards to items or services in the customer satisfaction surveys (D'Elia and Piccolo (2005)). It is assumed that data are modeled as a two components mixture of discrete distributions (Shifted Binomial and Uniform) with covariates related to some individual informations on each respondent. In Deldossi and Paroli (2011) inference on the model's parameters has been developed in a Bayesian perspective.

The problem of variable selection can rise when many covariates are available and we have to choose the best subset of them. In Deldossi and Paroli (2011) three methods of classical Bayesian variable selection have been developed and adapted to the CUB model. The computational algorithms are based on the Kuo and Mallick (KM) method (Kuo and Mallick, 1998), Metropolized Kuo and Mallick (MKMK) method (Paroli

and Spezia, 2008) and Stochastic Search Variable Selection (SSVS) method (George and McCulloch, 1993). They compute the highest posterior probability of the model through suitable sampling MCMC schemes. Here we compute an alternative posterior model probability: the marginal posterior probability of inclusion of the covariates in the model. It is a method frequently used for finding active factors in Bayesian DOE experiments (see e.g. Meyer et al. (1996)) and introduced by Barbieri and Berger (2004) to identify the median probability model. We want to check the performance of this alternative criteria and compare it with the classical ones. For this aim, according to a special experimental design and assuming independence among the covariates, a simulation study has been introduced.

2. The CUB(p, q) model and the variables selection problem

The general formulation of a CUB(p, q) model is expressed by a mixture of a shifted Binomial($m-1, 1-\xi$) random variable and a discrete Uniform(m) random variable, whose probability distribution depends on the unknown parameters $\pi \in (0, 1]$, the mixture proportion, and $\xi \in [0, 1]$, related to the shifted Binomial parameter. Let R_i, $i = 1, \ldots, n$, be the ordinal random variable that describes the rate assigned by the i-th respondent to a given item of a preferences' test, with $r \in \{1, \ldots, m\}$; its probability function is expressed as:

$$P(R_i = r|\pi_i, \xi_i) = \pi_i \left(\begin{array}{c} m-1 \\ r-1 \end{array} \right) (1-\xi_i)^{r-1} \xi_i^{m-r} + (1-\pi_i)\frac{1}{m}. \qquad (1)$$

The maximum rate m, that has to be greater than 3 due to the identifiability conditions (Iannario (2010)), is known. A systematic component that links π_i and ξ_i to some subject's covariates vectors, Y_i and W_i, of dimension $p+1$ and $q+1$, respectively, can added as a logistic function:

$$\pi_i = \frac{\exp(Y_i'\beta)}{1+\exp(Y_i'\beta)}; \qquad \xi_i = \frac{\exp(W_i'\gamma)}{1+\exp(W_i'\gamma)}. \qquad (2)$$

The parameters of the model are then the vectors $\beta = (\beta_0, \beta_1, \ldots, \beta_p)\prime$ and $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_q)\prime$. Due to the choice of the logistic function, the parametric space of π_i and ξ_i are restricted to $\xi_i \in (0, 1)$ and $\pi_i \in (0, 1)$.

When many covariates are available the problem of variables selection can rise and the best subset of them have to be chosen. Classical Bayesian approaches to variables selection are based on the computation, through suitable MCMC algorithms, of the highest posterior probability of the model, i.e. the "best" subset of predictors is that with the most frequent appearance in the sequence of the MCMC iterations. Here we consider, instead of the highest posterior probability model criteria, the one based on the marginal posterior inclusion probability of the covariates in the model. In all the three considered methods, to compute the highest posterior probability model, two vectors of binary

indicators are associated respectively to the p-dimension and q-dimension covariates' vectors. They regulate the inclusion of the exogenous variables and they are either part of the model or included in the prior of the parameters. The posterior probability model is so defined as the p-dimensional or the q-dimensional marginal posterior probabilities of the indicators vectors. In the alternative criteria we compute the unidimensional marginal posterior probability for each component of the indicators vectors. Then the covariates with marginal posterior probability greater than a fixed threshold α will be included to defined optimal the model. In many cases this optimal probability model can differ from the highest probability model.

3. Simulation results

Following the simulation study in Deldossi and Paroli (2011), we start generating a series of length $n = 500$ from a CUB(p, q) model with $p = 2$, $q = 3$, $m = 7$ and 5 potential covariates. The exogenous variables are generated both from discrete ($X_1 \sim Be(0.5)$; $X_2 \sim U(5)$) and continuous random variables ($X_3, X_4, X_5 \sim N(0, 1)$). We assume that all the exogenous variables are mutually independent and the true model contains covariates X_1 and X_4 for the model on π_i and covariates X_2, X_3, X_5 for the model on ξ_i. So, the true binary vectors are $\delta_B = (110010)$ and $\delta_G = (101101)$. By varying the parameters of the CUB(2,3) model, as it is shown in Table 1, we simulate 19 different cases such that the whole parametric space of π_i and ξ_i has been explored. We implement the three MCMC sampling algorithms for each of the dataset simulated.

Table 1. Values of the parameters of the 19 simulated models

				β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
γ_0	γ_2	γ_3	γ_5	0	0	0	0.1	1	1	0.1	2	0.1	0.1	5	10
0	0	0	0	case 1						case 2			case 3		
1	0.5	0.1	0.5	case 4			case 5			case 6			case 7		
-4	0.5	0.1	0.5	case 8			case 9			case 10			case 11		
-2	1	5	5	case 12			case 13			case 14			case 15		
-0.5	0.1	0.1	0.5	case 16			case 17			case 18			case 19		

For each method and for all the 19 cases of Table 1 we obtained the values of the marginal posterior inclusion probabilities for every covariates in (2). The performance of the three MCMC algorithms are summarized in Table 2, where the relative frequency of including in the model the true covariates (right choice) and the wrong ones (false choice) is reported, by varying the threshold probability level α that regulate the decision to insert or not a variable in the final model. For the 19 cases of Table 1 and $\alpha \geq 0.6$, KM seems to have the best performance in correctly identifying the true model even if MKMK is the method that guarantees zero false choice. In order to prove that these results may be considered valid for a general CUB(p, q) model, we would increase the number of covariates and the orders p and q.

Table 2. Relative frequency of right and false choice for models in Table 1 varying α

	right choice				false choice		
α	KM	MKMK	SSVS		KM	MKMK	SSVS
0.5	0.882	0.633	1.000		0.732	0.160	1.000
0.6	0.789	0.544	0.607		0.105	0.000	0.368
0.7	0.779	0.533	0.353		0.084	0.000	0.204
0.8	0.726	0.512	0.342		0.053	0.000	0.105
0.9	0.689	0.484	0.342		0.021	0.000	0.056

Acknowledgements: The paper has been prepared within a MIUR grant (code 2008WKH JPKPRIN2008 - PUC number E61J10000020001) for the project: "Modelli per variabili latenti basati su dati ordinali: metodi statistici ed evidenze empiriche" (Research Unit University of Naples Federico II)

References

Barbieri M.M., Berger J. (2004), Optimal Predictive Model Selection, *The Annals of Statistics*, 32, 870–897.

Deldossi L., Paroli R. (2011), Some notes on Bayesian inference for CUB model, in: *8th International Meeting of the CLADAG*, Pavia.

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

George E.I., McCulloch R.E. (1993), Variables selection via Gibbs-Sampling, *Journal of the American Statistical Association*, 88, 881-889.

Kuo L., Mallick B. (1998), Variable Selection for Regression Models, *Sankhya, Series B*, 60, 65-81.

Meyer R.D., Steinberg D.M., Box G. (1996), Follow-up designs to resolve counfounding in multifactor experiments, *Technometrics*, 38, 4, 303–313.

Paroli R., Spezia L. (2008), Bayesian Variable Selection in Markov Mixture Models, *Communication in Statistics - Simulation and Computation*, 37, 25-47.

Quaderni di Statistica
Vol. 14, 2012

Evaluating R&R of ordinal classifications with CUB model

Laura Deldossi Diego Zappa
Department of Statistical Science, Università Cattolica del Sacro Cuore, Milano
E-mail: laura.deldossi@unicatt.it, diego.zappa@unicatt.it

Summary: The CUB model is a class of model for ordinal data obtained as a mixture distribution. It is adopted to assess the reliability of categorical measurement in several and different contexts, such as marketing and sensometrics. One of the main opportunities of this model is to include subjects' covariates in its probability structure. Motivated by the work of De Mast and Van Wieringen (2010), in this paper we propose to modify the CUB model in order to define Repeatability and Reproducibility indexes (R&R) for ordinal data in business and industry framework.

Keywords: Ordinal data, Measurement System Analysis, Gauge Capability

1. Introduction

Ordinal data are frequently used in business and industry, e.g. to classify manufacturing faults into *minor*, *major*, *critical* or for quality evaluations using a scale such as *good*, *acceptable*, *questionable*, *rejected*. In such a cases there are no instruments and only subjective ability and/or the expertise of the appraisers may contribute to the capability of the measurement system. Since measures are collected on ordinal scale, standard methods for *numerical* Measurement System Analysis (MSA) such as Gauge R&R indexes, typically based on variance components, can not be applied. Other methods offered in the literature treats ordinal data as either nominal data (Kappa index), numerical data (weighted Kappa index), or rankings that in turn are treated as numerical data (Kendall's W), so not providing satisfactory results. Searching for an approach that specifically takes into account the features of ordinal data, De Mast and Van Wieringen (2010) propose to modify a class of model from Item Response Theory, usually adopted to assess the reliability of categorical measurements in psychometrics, to deal with ordinal data in business and industry. They model the ordinal measurements using the

Partial Credit Model. According to their proposal the repeatability for each appraiser (intra-appraiser) is expressed as the probability of correct ordering and consistent classification (see formula (5) and (6) in De Mast and Van Wieringen (2010) for their definition). Then reproducibility is obtained extending the above concepts in a inter-appraisers variant. Their approach is very appealing and it seems to offer satisfactory interpretations, but, in our opinion, some drawbacks are present. In fact their definition of R&R is slightly different with respect to the established meaning of numerical MSA where repeatability is the variation in measurements taken by a single person on the same item and under the same conditions, while reproducibility is the degree of agreement between measurements conducted varying the experimental conditions, e.g. by different people.

Motivated by this work, our aim is to study whether we can modify the CUB model, commonly used to explain the behavior of respondents to items in education or customer satisfaction context, in order to define R&R indicators for ordinal data. To illustrate the approach and to validate the results an example is presented.

2. The CUB model

CUB models have been introduced in the literature by Piccolo (2003), D'Elia and Piccolo (2005) and generalized by Piccolo and D'Elia (2008) and Iannario (2012). Its acronym stands for Combination of Uniform and Shifted Binomial. It has been introduced to explain the judgment/rate y, with $y \in \{1, \ldots, m\}$ by

$$P(Y = y|\pi, \xi) = \pi \binom{m-1}{y-1} (1-\xi)^{y-1} \xi^{m-y} + (1-\pi)\frac{1}{m} \qquad (1)$$

where the unknown parameters are $\pi \in (0, 1]$, i.e. the mixture proportion, and $\xi \in [0, 1]$ relative to the shifted Binomial parameter. π is inversely related to the *uncertainty* in judging an item on the categorical scale, while ξ is positively related to the degree of liking/disliking *feeling* expressed by the raters towards the item. The maximum rate m, that must be greater then 3 for identifiability conditions (Iannario (2010)), is assumed known.

Inference on model (1) has been developed both in a classical framework via maximum likelihood methods (see Piccolo (2006)) and from a Bayesian point of view via a suitable MCMC algorithm (see Deldossi and Paroli (2011)).

3. R&R for ordinal data using the CUB model

In MSA context, to assess the R&R of a classification procedure one takes I objects, which are evaluated K times by each of the J appraisers into one of the categories $y = 1, ..., m$. Then data y_{ijk} denote the evaluation of the object $i \in \{1, \ldots, I\}$ from the appraiser $j \in \{1, \ldots, J\}$ in the k-th replication, with $k \in \{1, \ldots, K\}$.

For $K = 1$ the measurement scheme described above is analogous to that of the CUB model in the preference analysis issue where: items become objects, raters become appraisers and the couple (π, ξ) is estimated for each object. However some problems exist in the application of CUB to MSA, as it has pointed out in Deldossi and Zappa (2011). The main question is that the number of appraisers (i.e. the sample size used for the estimation procedure) is really small: in general the number J of the operators is not greater than 10. This constraint leads to a twofold consequence: first, we cannot apply any asymptotic results; second, the π parameter turns out to be bounded from below. Both to overcome these problems and to define proper R&R indexes for ordinal data, we exploit CUB model to estimate (π, ξ) for each appraiser, using the sample of I objects for inference. As a consequence we assume a measurement system with fixed appraisers. Depending on how the experimental design to collect data is executed, independence among replications may be guaranteed.

As it is reasonable for an *in control* process, we will assume that the objects in the sample are homogeneous, that is the parts are all e.g. *acceptable* or *good*. Observe that assessing the quality of the process is not the aim of MSA.

In this framework π can be interpreted as the degree of the appraiser's expertise (π closer to 1 implies greater experience of the operator) while ξ his/her ability to recognize the quality of the evaluated parts. After having estimated π_{jk} and ξ_{jk}, the corresponding cumulative probability function (cdf) of $(Y|\pi_{jk}, \xi_{jk})$, for every $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$ are computed. From suitable comparisons of the above cdfs we are able to define R&R indicators.

In particular, since in numerical MSA repeatability concerns the variability among replications, we can obtain information about repeatability comparing for each appraisers $j \in \{1, \ldots, J\}$ the cdfs of $(Y|\pi_{jk}, \xi_{jk})$ by varying $k \in \{1, \ldots, K\}$.

We conclude that the measurements are repeatable if, for all the J appraisers, the cdfs among replications result overlapped. Otherwise we deduce some problems have occurred and we can identify the appraisers that caused it.

Analogously, reproducibility concerns variability among operators, then we can deduce proper conclusions comparing for each replication $k \in \{1, \ldots, K\}$ the cdfs of $(Y|\pi_{jk}, \xi_{jk})$ by varying $j \in \{1, \ldots, J\}$. We assess that the measurements are reproducible if, for all the K replications, the cdfs among appraisers result overlapped. Otherwise some problems have occurred and we can understand the reason for it.

To verify the similarity among CUB cdfs we will compute the envelope of indifference that corresponds to an acceptance region of the hypothesis of a capable measurement system.

To illustrate the approach and to validate the results we consider the example introduced in De Mast and Van Wieringen (2010) related to the redesignation of a process for soldering printed circuit boards (PCB) in an electronics manufacturer where the solder joints quality was judged by means of a visual inspection from $J = 3$ appraisers, twice ($K = 2$) on $I = 45$ PCB (initial experiment) and, three weeks later, on $I = 30$ PCB (follow-up experiment).

Preliminary results of the application of our proposal are encouraging since our conclusions are in line with those of De Mast and Wieringen (2010) with the additional property of greater simplicity in the interpretation of results.

Moreover our approach allows the introduction in the model of parts' covariates, if available, to better explain the performance of the measurement system.

Acknowledgements: The paper has been prepared within a MIUR grant (code 2008WKH JPKPRIN2008 - PUC number E61J10000020001) for the project: "Modelli per variabili latenti basati su dati ordinali: metodi statistici ed evidenze empiriche" (Research Unit of University of Naples Federico II).

References

Deldossi L., Paroli R. (2011), Inference on the CUB model: an MCMC approach, in press on "Studies in Classification, Data Analysis, and Knowledge Organization" Series, Springer, Berlin.

Deldossi L., Zappa D. (2011), Measurement errors and uncertainty: a statistical perspective, in: Ingrassia S., Rocci S., Vichi M. (eds.), *New Perspective in Statistical Modeling and Data Analysis*. Springer, Berlin, 145–153.

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

De Mast J.,Van Wieringen W.N. (2010), Modeling and Evaluating Repeatability and Reproducibility of ordinal classifications, *Technometrics*, 52, 1, 94–106.

Iannario M. (2010), On the identifiability of a mixture model for ordinal data, *Metron*, LXVIII, 87–94.

Iannario M. (2012), Modeling shelter choices in a class of mixture models for ordinal responses, *Statistical Methods and Applications*, 21, 1–22.

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

Quaderni di Statistica
Vol. 14, 2012

Simulation based estimation for Generalized Latent Linear Variables Models

Elise Dupuis-Lozeron Maria-Pia Victoria-Feser
Research Center for Statistics, Faculty of Economics and Social Sciences, University of Geneva
E-mail: Elise.Dupuis@unige.ch, Maria-Pia.VictoriaFeser@unige.ch

Summary: Generalized Linear Latent Variables Models (GLLVM) constitute a broad class of models that offer a general framework for modeling relationships between manifest and latent variables, as the manifest variables can follow any distribution of the exponential family (e.g, binomial, multinomial or normal). However, the estimation of such models is quite difficult due to the complexity of the associated log-likelihood function which contains integrals without closed form expression, except in the normal case. We propose a method based on indirect inference (Gouriéroux, Monfort, and Renault 1993) which starts from an easy to compute estimator that is then corrected for bias.

Keywords: Indirect inference, GLM, Ordinal variables.

1. Introduction

Latent variables models have now become very popular in various area of research such as psychology, social sciences or economics. When manifest (or observed) variables are supposed to be multivariate normal, like in Factor Analysis (FA), maximum likelihood estimation is performed from the sample covariance matrix of the manifest variables. However, in many applied fields, the measures are very seldom taken on a normal scale. Generalized Linear Latent Variable Models (GLLVM), that were developped for example by Bartholomew (1984) and Moustaki and Knott (2000), propose a more general framework for latent variable models in the case of normal and non-normal observations. Nevertheless, there is an important difficulty with GLLVM, namely the estimation of their parameters. Indeed, as the latent variables are not observed they must be integrated out from the likelihood function which implies the non-analytic evaluation of multiple integrals. Several methods of numerical integration have been proposed to

approximate the likelihood before maximizing it (e.g. Huber, Ronchetti, and Victoria-Feser 2004, Rabe-Hesketh, Skrondal, and Pickles 2002). However these numerical integration methods coupled with the maximum likelihood estimating equations require an expert tuning of the optimization parameters to make the optimization of the approximated likelihood successful.
In order to avoid part of this numerical problems, we propose an alternative estimator, based on indirect inference (Gouriéroux, Monfort, and Renault 1993, Gallant and Tauchen 1996).

2. Generalized Linear Latent Variables Models

Latent variables models combine a set of factors denoted by $z^{(k)}$, $k = 1, \ldots, q$, that convey most of the information contained in a set of response variables denoted by $x^{(j)}$, $j = 1, \ldots, p$, with p (much) larger than q. A crucial assumption is that all the dependence structure of the manifest variables is explained by the factors. This is known as the assumption of conditional or local independence.
In GLLVM the conditional expectation of the manifest variables is linked to the latent variables through a link function in the following manner

$$\nu_j(E(x^{(j)}|\mathbf{z})) = \boldsymbol{\alpha}^{(j)\mathrm{T}}\mathbf{z} \tag{1}$$

where $\boldsymbol{\alpha}^{(j)} = (\alpha_0^{(j)}, \ldots, \alpha_q^{(j)})^{\mathrm{T}} = (\alpha_0^{(j)}, \boldsymbol{\alpha}_{(2)}^{(j)})^{\mathrm{T}}$ are the loadings,
$\mathbf{z} = (1, z_1, \ldots, z_q)^{\mathrm{T}} = (1, \mathbf{z}_{(2)})^{\mathrm{T}}$ and ν_j is the link function which can be any monotonic differentiable function and may be different for different manifest variables $x^{(j)}, j = 1, \ldots, p$. The latent variables are used as covariates in the linear predictor $\boldsymbol{\alpha}^{(j)\mathrm{T}}\mathbf{z}$. The conditional distribution of the manifest variables given the latent ones is supposed to belong to the exponential family, i.e.

$$g_j(x^{(j)}|\mathbf{z}) = \exp\left\{\frac{x^{(j)}u_j(\boldsymbol{\alpha}^{(j)\mathrm{T}}\mathbf{z}) - b_j(u_j(\boldsymbol{\alpha}^{(j)\mathrm{T}}\mathbf{z}))}{\phi_j} + c_j(x^{(j)}, \phi_j)\right\} \tag{2}$$

where ϕ_j is the scale parameter and $u_j(\boldsymbol{\alpha}^{(j)\mathrm{T}}\mathbf{z})$ is the so-called canonical parameter. The form of the fonctions $u_j(.)$, $b_j(.)$ and $c_j(.)$ depends on the specific distribution $g_j(x^{(j)}|\mathbf{z})$, $u_j(.)$ being the identity function when ν_j is the canonical link.
Due to the assumption of conditional independence, the joint conditional distribution of the manifest variables is $\prod_{j=1}^{p} g_j(x^{(j)}|\mathbf{z})$. The density of the latent variables, denoted by $h(\mathbf{z}_{(2)})$ is assumed to be multivariate standard normal. We will also assume that they are independant. The joint distribution of the manifest and latent variables is then $\prod_{j=1}^{p} g_j(x^{(j)}|\mathbf{z})h(\mathbf{z}_{(2)})$. As the latent variables are not observed, we consider their realizations as missing and we integrated them out. This gives the following marginal

density for the manifest variables

$$f_{\boldsymbol{\alpha},\boldsymbol{\phi}}(\mathbf{x}) = \int \dots \int \Big\{ \prod_{j=1}^{p} g_j(x^{(j)}|\mathbf{z}) \Big\} h(\mathbf{z}_{(2)}) d\mathbf{z}_{(2)} \tag{3}$$

Given a sample of n observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{\mathrm{n}}$ where $\boldsymbol{x}_1 = (x_i^{(1)}, \dots, x_i^{(p)}), i = 1, \dots, n$, the log-likelihood of the loadings $\boldsymbol{\alpha}$ and the scale parameters $\boldsymbol{\phi}$ is

$$\begin{aligned} l(\boldsymbol{\alpha}, \boldsymbol{\phi}|\mathbf{x}) &= \sum_{i=1}^{n} \log f_{\boldsymbol{\alpha},\boldsymbol{\phi}}(\mathbf{x}_i) \\ &= \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \prod_{j=1}^{p} \exp\Big\{ \frac{x_i^{(j)} u_j(\boldsymbol{\alpha}^{(j)\mathrm{T}}\mathbf{z}) - b_j(u_j(\boldsymbol{\alpha}^{(j)\mathrm{T}}\mathbf{z}))}{\phi_j} \\ &\quad + c_j(x^{(j)}, \phi_j) \Big\} h(\mathbf{z}_{(2)}) d\mathbf{z}_{(2)} \end{aligned} \tag{4}$$

MLE for $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ is obtained by derivating (4) with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$.
The multidimensional integral does not have a closed form expression except in the case where the manifest variables are multivariate normal. This is also the case for Generalized Linear Mixed Models (GLMM) (see Pinheiro and Chao 2006 for example). Consequently, numerical integration can be used in order to optimize the log-likelihood for GLLVM. Another approach is to apply simulation-based techniques to get rid of the multidimensional integral. We propose to use indirect inference, which is such a technique.

3. Indirect inference for GLLVM

Indirect inference (see Gouriéroux, Monfort, and Renault 1993, Gallant and Tauchen 1996) has been proposed as an estimation procedure for a complex model $F_{\boldsymbol{\theta}}$ with intractable likelihood function. An auxiliary estimator $\hat{\boldsymbol{\pi}}$, which is known to be a biased estimator of $\boldsymbol{\theta}$, is first computed from a sample $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ supposedly generated from $F_{\boldsymbol{\theta}}$. For instance, the auxiliary estimator could be an approximation of the exact likelihood or the exact likelihood of an approximated model. Let $\hat{\boldsymbol{\pi}}$ be an M-estimator defined as the solution of

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{x}_i; \hat{\boldsymbol{\pi}}(F_n)) = \mathbf{0} \tag{5}$$

where F_n is the empirical distribution. A consistant estimator for the complex model is obtained implicitly by the solution in $\boldsymbol{\theta}$ of

$$\int \boldsymbol{\psi}(\mathbf{x}; \hat{\boldsymbol{\pi}}(F_n)) dF_{\boldsymbol{\theta}} = \mathbf{0} \tag{6}$$

In general (6) is not analytically tractable because of the multidimentional integral. The integral in (6) can however be estimated by the empirical mean over $n^\star$ observations $\mathbf{x}_i(\boldsymbol{\theta})$ simulated from $F_{\boldsymbol{\theta}}$ for a given $\boldsymbol{\theta}$ whith $n^\star = n \times l$ and l being as large as possible. For our purpose, we define an auxiliary estimator based on the Generalized Linear Model (GLM) in which the covariates are fixed values for the latent variables and are estimated first as the Barelett's score of a FA performed on the original data. The estimates obtained by the GLM are the auxiliary estimates of the parameters of the GLLVM. Then, we debias the auxiliary estimator by means of a numerical method based on resampling that solve (6). Simulations from models with ordinal manifest variables and several latent variables show that our estimator has good finite sample properties
Our estimator is actually very general and can be applied to many different latent variable models provided that a reasonnable auxiliary estimator can be found. The same method has been used successfully by Mealli and Rampichini (1999) in GLMM. We think that it is a really interesting alternative to other estimation method for GLLVM based on complex numerical integration methods.

References

Bartholomew D.J. (1984), The foundations of factor analysis, *Biometrika*, 71, 221–232.

Gallant A.R., Tauchen G. (1996), Which moments to match, *Econometric Theory*, 12, 657–681.

Gouriéroux C., Monfort A., Renault E. (1993), Indirect inference, *Journal of Applied Econometrics*, 8 (supplement), S85–S118.

Huber P., Ronchetti E., Victoria-Feser M.P. (2004), Estimation of generalized linear latent variable models, *Journal of the Royal Statistical Society, Series B*, 66, 893–908.

Mealli F., Rampichini C. (1999), Estimating binary multilevel models through indirect inference, *Computational Statistics & Data Analysis*, 29, 313–324.

Moustaki I., Knott M. (2000), Generalized latent trait models, *Psychometrika*, 65, 391–411.

Pinheiro J., Chao E. (2006), Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models, *Journal of Computational and Graphical Statistics*, 15, 58–81.

Rabe-Hesketh S., Skrondal A., Pickles A. (2002), Reliable estimation of generalized linear mixed models using adaptive quadrature, *The Stata Journal*, 2, 1–21.

Quaderni di Statistica
Vol. 14, 2012

Qualitative latent variables: a comparison between SEM and LCA

Daniele Durante
Department of Statistical Sciences, University of Padua
E-mail: daniele.durante.2@gmail.com

Summary: In this paper differences and analogies between estimates of structural equation models (SEM) and those of latent class models (LCA) are analyzed, considering in particular quantitative elements. We propose a method to compare the estimated coefficients from the SEM to the conditional probabilities resulting from LCA, and we discuss whether and to what extent one of these techniques is the approximation of the other (at least from the point of view of the final results). We analyze this problem by considering the contingency tables obtained by using the values predicted by the SEM and comparing the predictions of SEM with the conditional probabilities obtained through LCA. An application of the comparison methodology to customer satisfaction analysis of a business company is also presented.

Keywords: Structural equation models, Latent class analysis, Ordinal variables.

1. Introduction

The identification of latent constructs and the analysis of the relationships among them, are very important in many fields of application. This is especially true in marketing where, for example, customer satisfaction is related to multivariate latent structures whose analysis allows to formulate operational strategies.

The variables measured in marketing research are often expressed in a qualitative ordinal scale (e.g., Likert scales), and often analysts want to identify latent variables when ordinal categorical data are observed. The latent class analysis (LCA) developed by Goodman (1974) represent an ad-hoc solution to this problem. However, in practice, structural equation models (SEM, Bollen, 1989) are more often used. For example the American Customer Satisfaction Index suggests the use of this class of models to analyze customer satisfaction of business companies. One possible approach to include

qualitative variables in SEM has been proposed by Muthén (1983) through the analysis of threshold values; however, the applicability of this statistical technique is subject to a number of problems: the estimation is computationally laborious, and the model requires distributional assumptions for the calculation of polychoric correlations. For these reasons, in applied research, usually qualitative variables are considered and treated as if they were quantitative.

In this paper we propose a method to compare the estimated coefficients from the SEM to the conditional probabilities resulting from LCA, and we discuss whether and to what extent one of these techniques is the approximation of the other. We analyze this problem by considering the contingency tables obtained by using the values predicted by the SEM (in which the qualitative variables are assumed to be quantitative), and comparing them with the conditional probabilities obtained through latent class analysis.

An application of the comparison methodology is applied to the identification of latent variables related to customer satisfaction of a business company.

2. Comparison between models

In our analysis we focus our attention on the comparison between the estimates of SEM and LCA considering factor analysis model. Let $X = \Lambda_x \xi + \delta$ be the model of factor analysis in SEM form, where X is a vector $q \times 1$ of observed variables typically considered quantitative, Λ_x a matrix $q \times k$ of factor loadings to be estimated, ξ a vector $k \times 1$ of latent factors and δ a vector of errors with distribution $N_q(0, \Theta_\delta)$.

The latent factors have zero mean, variance and covariance matrix Φ and are assumed uncorrelated with the errors of the model. Under these assumptions the model can be consistently estimated by MLE, using the information contained in the sample correlation matrix of observed variables X.

The corresponding specification of the model of factor analysis in LCA form is:

$$p(X = x_i) = \sum_{j_1=1}^{z_1} .. \sum_{j_k=1}^{z_k} \left[p(\xi_1 = j_1, .., \xi_k = j_k) \prod_{h=1}^{q} p(X_h = x_{h,i} | \xi_1 = j_1, .., \xi_k = j_k) \right] \tag{1}$$

where X is a vector $q \times 1$ of variables measured on qualitative scale (or appropriately transformed into qualitative from quantitative variables), $\xi_1, ..., \xi_k$ a set of latent factors assumed qualitative with joint distribution $p(\xi_1 = j_1, ..., \xi_k = j_k)$. Here, $p(X_h = x_{h,i} | \xi_1 = j_1, ..., \xi_k = j_k)$ $h = 1, ..., q$ are the unknowns conditional probabilities that describe the relationship between the observed variables and latent classes. The MLE estimate has not explicit form and is done via EM algorithm using information contained in the table of absolute frequencies of the vector X observed on the sample analyzed.

The aim of our work is to develop a technique to compare the $p(X_h = x_{h,i} | \xi_1 = j_1, ..., \xi_k = j_k)$ $h = 1, ..., q$ fitted by LCA with the same quantities obtained by transforming the predictions given by SEM. The idea is to hold firm estimates of the LCA

and work on the structural equation model trying to express the estimates in the form of conditional probabilities.

Let $\hat{\Lambda}_x$ be the estimated matrix of the SEM model; for each statistical unit we estimate the vector $\hat{\xi}$ by using the observed variables X. Therefore, each unit $i, i = 1, ..., n$ is associated with a vector containing the q observed variables and the k estimated factors. In a second step we calculate the same conditional probabilities estimated by the LCA by using the predictions of the SEM model, recoding both variables X and latent factors $\hat{\xi}$ in the same scale in which they were transformed in order to fit LCA. We will have, therefore, a measure of $p(X_h = x_{h,i}|\xi_1 = j_1, ..., \xi_k = j_k)\ h = 1, ..., q$ where $\hat{\xi}_1 = j_1, ..., \hat{\xi}_k = j_k$ are the latent factors predicted by SEM model, that can be compared with LCA estimates. To evaluate significance of observed differences between the measures of conditional probabilities implied by the predictions of the SEM model and those estimated by LCA, we implemented a bootstrap procedure.

3. Results

The data analyzed refer to a customer satisfaction survey of a company of computer services. The survey, based on 324 individuals, is composed of 29 questions measured on a scale from 1 to 10 and divided into 7 main areas.

For simplicity, we focused on 6 items, 3 of which relate to *customer care* and the other 3 to *product quality*. Using these variables a factor analysis with two latent factors (customer care, product quality) and two groups of indicators (each item measures a single factor) has been estimated according to the SEM model specified above.

The estimation of LCA model has been conducted on variables transformed into dichotomous by associating value 1 to values greater or equal to the sample mean of the variable considered, 0 otherwise. The two factors have been assumed qualitative and dichotomous. The model specification is consistent with the structure of the SEM: the conditional probability of the items concerning the quality of the product depends only on the factor product quality, the same goes for customer care. So that, the two models have the same dipendence structure.

Results in Table 1 show that the estimates are very similar. With the exception of the probability $p(X_5 = 0|\xi_2 = 0)$ where there is a marked difference. Table 2 shows the 95% bootstrap confidence intervals for the differences between estimates based on 1000 resamplings. All intervals (also that for $p(X_5 = 0|\xi_2 = 0)$) contain the 0.

4. Conclusion

The method proposed allows to compare SEM and LCA models. We considered a specific case related to analysis of customer satisfaction, but the proposed method is more general as shown in a simulation study.

Table 1. Estimates of the conditional probabilities from LCA model and from SEM model.

	$p(X_i = 0\|\xi_j = 0)$	$p(X_i = 0\|\xi_j = 0)$	$p(X_i = 0\|\xi_j = 1)$	$p(X_i = 0\|\xi_j = 1)$
	SEM	LCA	SEM	LCA
$j = 1$ - *customer care*				
X_1	0.934	0.943	0.004	0.002
X_2	0.824	0.842	0.043	0.037
X_3	0.967	0.959	0.094	0.099
$j = 2$ - *product quality*				
X_4	0.762	0.760	0.063	0.049
X_5	0.950	0.900	0.063	0.068
X_6	0.881	0.898	0.202	0.179

Table 1. 95% bootstrap confidence intervals for the differences in the LCA and SEM.

	mean	s.d.	CI 95%
$p(X_1 = 0\|\xi_1 = 0)$	-0.014	0.041	$[-0.095; 0.065]$
$p(X_1 = 0\|\xi_1 = 1)$	0.007	0.005	$[-0.009; 0.011]$
$p(X_2 = 0\|\xi_1 = 0)$	-0.026	0.030	$[-0.084; 0.033]$
$p(X_2 = 0\|\xi_1 = 1)$	0.006	0.007	$[-0.008; 0.020]$
$p(X_3 = 0\|\xi_1 = 0)$	-0.001	0.031	$[-0.062; 0.060]$
$p(X_3 = 0\|\xi_1 = 1)$	-0.004	0.011	$[-0.026; 0.018]$
$p(X_4 = 0\|\xi_2 = 0)$	-0.012	0.050	$[-0.111; 0.086]$
$p(X_4 = 0\|\xi_2 = 1)$	0.013	0.014	$[-0.014; 0.041]$
$p(X_5 = 0\|\xi_2 = 0)$	0.044	0.035	$[-0.025; 0.114]$
$p(X_5 = 0\|\xi_2 = 1)$	-0.013	0.030	$[-0.072, 0.046]$
$p(X_6 = 0\|\xi_2 = 0)$	-0.029	0.037	$[-0.102; 0.044]$
$p(X_6 = 0\|\xi_2 = 1)$	0.021	0.017	$[-0.013; 0.055]$

References

Bollen K.A. (1989), *Structural equations with latent variables*, Wiley, New York.

Goodman L.A. (1974), The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach, *Journal of Time Series Analysis*, 79, 1179–1259.

Muthén B. (1983), Latent variable structural equation modeling with categorical data, *Journal of Econometrics*, 22, 48–65.

Quaderni di Statistica
Vol. 14, 2012

Globally-optimized latent variable extraction in formative-reflective models

Marco Fattore Giorgio Vittadini
Department of Quantitative Methods, University of Milano-Bicocca
E-mail: marco.fattore@unimib.it, giorgio.vittadini@unimib.it

Matteo Pelagatti
Department of Statistics, University of Milano-Bicocca
E-mail: matteo.pelagatti@unimib.it

Summary: In this paper, we propose a novel globally optimal procedure to extract exogenous and endogenous latent variables (LVs) in formative-reflective structural equation models. The procedure is a valuable alternative to PLS-PM and Lisrel, since it is fully consistent with the causal structure of formative-reflective schemes and extracts both the structural parameters and the factor scores without identification or indeterminacy problems. The algorithm estimates the structural model taking into acccount both the capability of the exogenous LVs to represent their manifest formative blocks and the capability of the endogenous LVs to explain the manifest reflective blocks. It can be applied to virtually any kind of formative-reflective scheme and can be easily implemented in any programming language with numerical optimisation capabilities.

Keywords: Structural Equation Models, Formative-reflective scheme, PLS-PM.

1. Introduction

Formative-reflective models are a standard tool in socio-economic research, particularly in the fields of causal modeling and multidimensional evaluation. Despite the relevance of the topic there are still unsolved methodological problems when dealing with formative constructs (Howell et al., 2007; Wilcox, 2008). In fact, these models are usually estimared using the Lisrel algorithm, which is affected by indeterminacy problems (Vittadini, 1989), or using the PLS-PM algorithm, which cannot handle reflective relationships properly (Vittadini et al., 2007). In a formative-reflective scheme,

the exogenous latent variables play a double role. On the one hand, they should summarize their formative blocks; on the other hand, they should mediate, *via* the system of endogenous latent variables, the causal relationships linking the formative side to the reflective side. Realizing this, the proposed procedure extracts the exogenous latent variables balancing between these two aspects.

2. Formative-reflective models

Let $\boldsymbol{x}_i$, $i = 1, \ldots, p$, and $\boldsymbol{y}_j$, $j = 1, \ldots, q$, be vectors of zero-mean manifest variables and let $\boldsymbol{\omega}_i, i = 1, \ldots, p$, be vectors of real coefficients. According to the formative-reflective scheme, each exogenous LV ξ_i is expressed as a linear combination of the MVs of the corresponding formative group:

$$\xi_i = \boldsymbol{\omega}_i' \boldsymbol{x}_i, \qquad i = 1, \ldots, p. \tag{1}$$

By stacking $\xi_1, \ldots, \xi_p$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ into the vectors $\boldsymbol{\xi}$ and $\boldsymbol{x}$ respectively, definitions (1) can be cast in the following compact form

$$\boldsymbol{\xi} = \boldsymbol{\Omega} \boldsymbol{x}, \tag{2}$$

with

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\omega}_1' & \mathbf{0}' & \ldots & \mathbf{0}' \\ \mathbf{0}' & \boldsymbol{\omega}_2' & \ldots & \mathbf{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \ldots & \boldsymbol{\omega}_p' \end{bmatrix}, \quad \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_p \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_p \end{bmatrix}.$$

In an analogous way, the vector $\boldsymbol{y}_j$ of MVs of the j-th reflective block is assumed to be built as sums of a rescaled common scalar endogenous LV η_j plus a residual $\boldsymbol{\varepsilon}_j$,

$$\boldsymbol{y}_j = \boldsymbol{\lambda}_j \eta_j + \boldsymbol{\varepsilon}_j, \qquad j = 1, \ldots, q,$$

where $\boldsymbol{\lambda}_j, j = 1, \ldots, q$, are real vectors. Again, stacking $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q$ and $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_q$ into the vectors $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$, respectively, we get

$$\boldsymbol{y} = \boldsymbol{\Lambda} \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{3}$$

with

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\lambda}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\lambda}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\lambda}_q \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_p \end{bmatrix}.$$

Finally, the endogenous LVs stacked in vector $\boldsymbol{\eta}$ are built as linear combinations of the exogenous LVs stacked in vector $\boldsymbol{\xi}$, namely

$$\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi}, \tag{4}$$

where $\boldsymbol{\Gamma}$ is a conformable matrix of real coefficients. By putting (2), (3) and (4) together we obtain the final model, linking the formative MVs and the reflective MVs, *via* the constrained latent structure expressed by the matrices $\boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}$:

$$\boldsymbol{y} = \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{x} + \boldsymbol{\varepsilon}.$$

2.1. Latent variables extraction

The extraction of variables $\xi_1, \ldots, \xi_p$ requires a compromise between two goals: on one hand, each of them should summarize effectively its own formative block; on the other hand, as a whole, they should indirectly predict (*via* the set of endogenous LVs) the variables in the reflective groups. The first goal would be achieved extracting exogenous LVs through the minimization of the following loss function:

$$L_x(\boldsymbol{\Pi}, \boldsymbol{\Omega}) = \frac{1}{p}\sum_{i=1}^{p} \frac{\mathbb{T}\mathrm{r}\{\mathbb{E}[(\boldsymbol{x}_i - \boldsymbol{\pi}_i\boldsymbol{\omega}_i'\boldsymbol{x}_i)(\boldsymbol{x}_i - \boldsymbol{\pi}_i\boldsymbol{\omega}_i'\boldsymbol{x}_i)']\}}{\mathbb{T}\mathrm{r}\{\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i']\}} \tag{5}$$

where $\boldsymbol{\pi}_i, i = 1, \ldots, p$, are vectors of regression coefficients of $\boldsymbol{x}_i$ on $\boldsymbol{\omega}'\boldsymbol{x}_i$, and

$$\boldsymbol{\Pi} = \begin{bmatrix} \boldsymbol{\pi}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\pi}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\pi}_p \end{bmatrix}.$$

The second goal would be achieved by extracting the exogenous (and thus the endogenous) LVs through the minimization of the following loss function

$$L_y(\boldsymbol{\Omega}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}) = \frac{1}{q}\sum_{j=1}^{q} \frac{\mathbb{T}\mathrm{r}\{\mathbb{E}[(\boldsymbol{y}_j - \boldsymbol{\lambda}_j\boldsymbol{\gamma}_j'\boldsymbol{\Omega}\boldsymbol{x})(\boldsymbol{y}_j - \boldsymbol{\lambda}_j\boldsymbol{\gamma}_j'\boldsymbol{\Omega}\boldsymbol{x})']\}}{\mathbb{T}\mathrm{r}\{\mathbb{E}[\boldsymbol{y}_j\boldsymbol{y}_j']\}}, \tag{6}$$

where $\boldsymbol{\gamma}_j'$ is the j-th row of matrix $\boldsymbol{\Gamma}$ and the vectors $\boldsymbol{\lambda}_j$ are regression coefficients of $\boldsymbol{y}_j$ on $\boldsymbol{\gamma}_j'\boldsymbol{\Omega}\boldsymbol{x}$. If the matrix $\boldsymbol{\Omega}$ were given, and so the latent variables $\boldsymbol{\xi} = \boldsymbol{\Omega}\boldsymbol{x}$, then L_y would be minimised by a rank-one reduced rank regression of each $\boldsymbol{y}_j$ on $\boldsymbol{\xi}$. The problem of extracting the exogenous and endogenous LVs taking into account both competitive goals can be solved by minimizing the following global loss function

$$L^{(\alpha)}(\boldsymbol{\Pi}, \boldsymbol{\Omega}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}) = (1-\alpha)L_x(\boldsymbol{\Pi}, \boldsymbol{\Omega}) + \alpha L_y(\boldsymbol{\Omega}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}). \tag{7}$$

where $\alpha \in [0,1]$ determines the relative weight given to each goal. When $\alpha = 0$, $\boldsymbol{\omega}_i$ is just the first eigenvector of $E[\boldsymbol{x}_i \boldsymbol{x}_i']$, ξ_i is the first principal component of the variables in the i-th formative block and $\boldsymbol{y}_j = \boldsymbol{\lambda}_j \boldsymbol{\gamma}_j' \boldsymbol{\xi} + \boldsymbol{\varepsilon}_j$ is an ordinary rank-one reduced rank regression of $\boldsymbol{y}_j$ on $\boldsymbol{\xi}$. On the contrary, when $\alpha = 1$ the latent variables ξ_i's are built as the linear combinations of the respective $\boldsymbol{x}_i$ that best fit, *via* the endogenous latent variables, the vector $\boldsymbol{y}$, and the whole problem reduces to a multivariate regression with many constraints, implied by the form of matrix $\boldsymbol{B} = \boldsymbol{\Lambda\Gamma\Omega}$.

3. First results and conclusion

The extraction methodology has been applied to both simulated and real datasets and its performance has been compared to that of PLS-PM. In general terms, the methodology proves effective in extracting latent variables and capturing causal links among them. Letting the parameter α moving in $[0,1]$, and comparing the extracted exogenous latent variables as α varies, the methodology also reveals whether the formative side and the reflective side of the model are consistently specified, or whether the goals of representing the formative blocks and explainng the reflective manifest variables cannot be jointly achieved. Real data application shows that our methodology produces results matching the causal structure of the model much better than PLS-PM and reveals the existence of causal links even when PLS-PM does not, with comparable predictive power. In fact, PLS-PM expresses endogenous latent variables as linear combinations of reflective manifest variables. This eventually leads to extracting exogenous latent variables without accounting properly for the formative blocks and weakening the causal links from the formative to the reflective side of the model. Differently, our methodology is designed exactly to take into account both sides, resulting in more balanced and interpretable results.

References

Howell R. D., Breivik, E., Wilcox, J.B. (2007), Reconsidering formative measurement, *Psychological Methods*, 12, 205-218.

Vittadini G. (1989), Intederminacy problems in the Lisrel model, *Multivariate Behavioral Research*, 24, 397-414.

Vittadini G., Minotti S., Fattore M., Lovaglio P.G. (2007), On the relationships among latent variables and residuals in PLS path modeling: the formative-reflective scheme, *Computational Statistics & Data Analysis*, 51, 5828-5846.

Wilcox J.B., Howell R.D., Breivik E. (2008), Questions about formative measurement, *Journal of Business Research*, 61, 1219-1228.

Quaderni di Statistica
Vol. 14, 2012

Making classifier performance comparisons when Receiver Operating Characteristic curves intersect

Silvia Figini
Department of Economics, Statistics and Laws, University of Pavia
E-mail: silvia.figini@unipv.it

Chiara Gigliarano
Department of Economics and Social Sciences, Marche Polytechnic University
E-mail: c.gigliarano@univpm.it

Pietro Muliere
Department of Decision Sciences, Bocconi University
E-mail: pietro.muliere@unibocconi.it

Summary: The main objective of this paper is to propose a novel approach for model comparisons when ROC curves show intersections. We investigate in a theoretical framework the relationship between ROC orderings and stochastic dominance and we propose alternative indicators that could substitute the common AUC measure.

Keywords: ROC curve, Classification, Stochastic dominance.

1. Introduction

The receiver operating characteristic (ROC) curve describes the performance of a classification or diagnostic rule, while the area under this curve (AUC) is a common measure for the evaluation of discriminative power; see e.g. Krzanowski et al. (2009). When ROC curves cross each other, the AUC measure can lead to biased results and we are not able to select the best model; see e.g. Hand (2009). Common practise is to compare crossing ROC curves by restricting the performance evaluation to proper subregions of scores (see e.g. Thomas, 2009). In our opinion, however, this issue should be more adequately handled in the statistical literature.

The main objective of this paper is, therefore, to propose a novel approach - based on stochastic dominance - for model comparisons, when ROC curves show intersections.

2. ROC curve and stochastic dominance

Consider a classification tool that gives a real-valued score to classify items into two categories: good or bad. Let the random variable X with c.d.f. F represent the score and $x = (x_1, x_2, ..., x_n)$ be a score profile from X with mean $\mu(x)$ and variance $\sigma^2(x)$. Let $\mathcal{X} = \{x : \mu(x) = \mu\}$ be the set of n-dimensional score profiles with mean μ.

Suppose that for a prespecified cut-off c, item i is labeled as *bad* if $x_i \leq c$ and as *good* otherwise. The true positive rate, or sensitivity, is $F_B(c) = Pr(X \leq c|Bad)$, while the false positive rate, or (1 - specificity), is $F_G(c) = Pr(X \leq c|Good)$.[1]

The ROC curve is obtained representing, for any fixed cut-off value, a point in the cartesian plane having as x-value the false positive rate and as y-value the true positive rate. The best curve is the one that is leftmost, the ideal one coinciding with the y-axis. Then the ROC curve is defined as a plot of $\{(u, ROC_X(u)), u \in (0, 1)\}$, where $ROC_X(u) = F_B(F_G^{-1}(u))$.

For sake of model comparisons, performance indicators based on the ROC curve have been proposed, such as the AUC, which is defined as the integrated sensitivity over all specificity ranges: $AUC = \int_{-\infty}^{+\infty} F_B(s)dF_G(s)$.

If the ROC curves do not cross each other, there is an unambiguous comparison of two diagnostic tests in terms of discriminative power and the AUC index provides consistent results. The ordering induced by the ROC curves is equivalent to the first stochastic dominance: $ROC_X(u) \leq ROC_Y(u)$ if and only if $F_B(F_G^{-1}(u)) \leq H_B(H_G^{-1}(u))$, $\forall u \in (0, 1)$, where X and Y represent the score of two different classifiers, with c.d.f. F and H, respectively. In symbols, we write that $X \geq_{FSD} Y$.

In comparing two score distributions, it is of interest to investigate the transformations by which one distribution is obtained from the other. Saying that $X \geq_{FSD} Y$ means that Y is obtained from X by a *first order performance increasing (FOPI) transfer*, according to which the cumulative proportion of bad individuals, increasingly ordered according to their scores, is always higher in Y than in X.

Let us denote *discriminative power index* any function $I : \mathcal{X} \to \mathbf{R}$. The function I satisfies the $FOPI$ principle of transfers if $I(X) \leq I(Y)$ whenever (X, Y) is a $FOPI$ transfer. Obviously, AUC satisfies this principle.

3. Comparing crossing ROC curves

If two ROC curves intersect each other, the first order stochastic dominance fails and it is not possible to employ the AUC index. Thus we move to the second order stochastic dominance (SSD), according to which X dominates Y (in symbols, $X \geq_{SSD} Y$) if $\int_0^z ROC_X(u)du \leq \int_0^z ROC_Y(u)du \ \forall z \in [0, 1]$.

[1] The sensitivity is the probability of correctly classifying a bad item, while the specificity is the probability of correctly classifying a good item.

The SSD can be obtained from a *second order performance increasing (SOPI) transfer*, according to which Y assigns to bad individuals the smallest scores with higher frequency and the highest scores with smaller frequency than X.[2]

Here we focus on the scenario of one crossing and we say that the ROC curve of distribution X *intersects* that of Y *once from below* if and only if there exists $u^* \in (0,1)$ such that $ROC_X(u) \leq ROC_Y(u)\ \forall\, u \leq u^*$ and $<$ for some $u \leq u^*$, and $ROC_X(u) \geq ROC_Y(u)\ \forall\, u \geq u^*$ and $>$ for some $u \geq u^*$. Figure 1 illustrates an example of intersecting ROC curves.

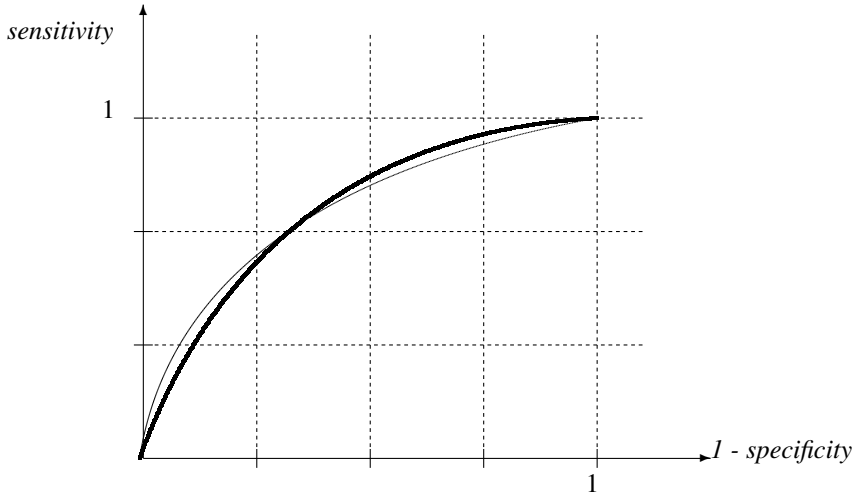


Figure 1. Intersecting ROC curves

Note that if $ROC_X(u)$ intersects once from below $ROC_Y(u)$ and if $\int_0^{u^*}(ROC_Y(u) - ROC_X(u))du \geq \int_{u^*}^1 (ROC_X(u) - ROC_Y(u))du$, then $X \geq_{SSD} Y$.

Since the AUC index may contradict with the criterion of the SSD, alternative measures are required. From Fishburn (1980), we have that the class of indices $I(X) = \int \psi(x)dF_B(x)$, with ψ nondecreasing and concave, is consistent with the SSD. This class of measures provides, therefore, a coherent alternative to the AUC.

If also the SSD is violated, we refer to the third order stochastic dominance: $X \geq_{TSD} Y$ if $\int_0^z \left(\int_0^x ROC_X(u)du\right) dx \leq \int_0^z \left(\int_0^x ROC_Y(u)du\right) dx\ \forall z \in [0,1]$.

The TSD can be obtained from a *third order performance increasing (TOPI) transfer*, according to which in Y a $SOPI$ transfer happens at a higher level of specificity than in X; this criterion thus puts more weigh to smaller false positive rates.

[2] In the income distribution literature, this transfer is called *regressive transfer*.

A discriminative power index I is consistent with the $TOPI$ transfer if and only if $I(Y) \geq I(X)$ with (X, Y) being a $TOPI$ transfer. Note that the AUC index does not satisfy this property.

In case of violation of SSD, it is still possible to compare two crossing ROC curves, provided that the ROC curve corresponding to the score distribution with lower variance intersects once from below the other curve; in particular, if $ROC_X(u)$ intersects once from below $ROC_Y(u)$ and if $\int_0^{u^*}(ROC_Y(u) - ROC_X(u))du \leq \int_{u^*}^1(ROC_X(u) - ROC_Y(u))du$, then $I(Y) > I(X)$ for all $TOPI$ consistent discriminative power indices $I(\cdot)$ if and only if $\sigma^2(y) \geq \sigma^2(x)$.

Following Fishburn (1980), we propose then a class of indices that are consistent with the TSD. More precisely, the class of indicators $I(X) = \int \psi(x)dF_B(x)$, where the function ψ is non-decreasing and concave with a non-negative third derivative, provides an alternative to the AUC measure that is coherent with the $TOPI$ principle of transfers.

4. Concluding remarks

We have provided a novel method for checking for unanimous classifier performance rankings when the ROC curve dominance fails. Our result does not resolve all the ambiguous rankings associated with single crossing ROC curves; it will, however, assist a large number of pairwise comparisons for which the AUC index is not applicable.

Next step of further research will be focused on (i) applying the inverse stochastic dominance theory within the ROC curve framework, (ii) extending the class of discriminative power indices on the basis of the Fishburn (1980)'s results, and finally (iii) providing empirical applications of our methodologies.

References

Fishburn P. (1980), Continua of stochastic dominance relations for unbounded probability distributions, *Journal of Mathematical Economics*, 7, 271–285.

Hand D.(2009), Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Machine Learning*, 77, 103–123.

Krzanowski, W.J. and Hand, D.J. (2009), *ROC curves for continuous data*, CRC/Chapman and Hall.

Lee W. (1999), Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. *Statistics in Medicine*, 18, 455–471.

Thomas L.C. (2009), *Consumer credit models: pricing, profit, and portfolios*, Oxford University Press.

Quaderni di Statistica
Vol. 14, 2012

Latent class mixed effects models for partially ranked preferences – examining changes in postmaterialism over time

Brian Francis
Department of Mathematics and Statistics, Lancaster University
E-mail: B.Francis@Lancaster.ac.uk

Regina Dittrich Reinhold Hatzinger
Institute for Statistics and Mathematics, Vienna University of Economics and Business
E-mail: Regina.Dittrich@wu.ac.at, Reinhold.Hatzinger@wu.ac.at

Summary: This paper discusses the use of mixture models to allow for a random effects structure in the analysis of partially ranked data. Partially ranked data occur in sample surveys, where respondents are invited to choose the preferred and second most preferred (or the most preferred and least preferred) out of a larger set of items. We adopt an paired comparison approach used by Dittrich *et al*. (2007) which treats partially ranked data as originating from underlying Likert scales. The work is motivated by two questions on (post)materialism in consecutive sweeps of the British Household Panel Survey. As part of the survey, respondents were asked to choose the most preferred and the next most preferred out of a set of four items representing preferred priorities for government.

Keywords: Nonparametric maximum likelihood, Latent class analysis, Paired comparisons.

1. Introduction

The method of paired comparisons, popularized by Thurstone(1927), was designed to measure the relative importance or worth of a set of objects. Essentially, with J objects, each pair of objects is taken, and respondents are asked to judge which of the two is most important. In this paper we are concerned with partially ranked objects, where respondents are instead asked to determine the partial ordering of the objects. For ex-

ample, respondents might be asked to identify the most important and the next most important of a set of objects. It is straightforward to transform fully ranked data into paired comparison (PC) form (Dittrich *et al*., 2000), but dealing with partially ranked data is more problematic, as the full rank ordering is unknown. Francis *et al*. (2002) considered a model where the independence of paired comparisons is assumed, but here we consider partially ranked data as a special case of multiple Likert responses (Dittrich *et al*., 2007). The advantage of such an approach is that the problem is placed in the standard framework of generalized linear models and respondent covariates can also be incorporated.

Our model is similar to the Mallows-Bradley-Terry ranking model (Critchlow and Fligner, 1991). In their model the probability of each ranking of the items is taken to be proportional to the product of the probabilities of all pairwise comparisons that are consistent with the ranking. In our approach we use the correspondence between partial rank patterns and the derived PC-patterns. The probability for a single PC response y_{jk} is defined by a reformulation of the Bradley-Terry (BT) model

$$P(y_{jk}) = a_{jk}\ c_{jk}^{(1-y_{jk}^2)} \left(\frac{\sqrt{\pi_j}}{\sqrt{\pi_k}}\right)^{y_{jk}},$$

where y_{jk} takes the value of 1, if object j is preferred to k, takes the value of -1, if object k is preferred to j, and takes the value of zero if no preference is stated. and a_{jk} is a normalising constant. The c_{jk} can be thought of as representing a different probability of no preference for each pair of responses.

The probability of a specific partial rank pattern ℓ is then given by the product over all derived PCs and can be written $P_\ell = P_\ell(y_{12}, y_{13}, \ldots, y_{J-1:J}) = \Delta \prod_{j<k} P(y_{jk})$.

Estimation is based on simple multinomial sampling over the observed partial rank patterns. Let N_ℓ be the number of times that the specific partial rank pattern ℓ is observed, with $N = \sum_\ell N_\ell$. Then the N_ℓ are multinomially distributed, and the likelihood function is

$$L = \Delta^* \prod_u P_\ell^{N_\ell}.$$

The linear predictor of the basic pattern model is $\eta_\ell = \sum_{j<k} y_{jk}(\lambda_j - \lambda_k)$, where the λs (location of the preference parameters) are related to the πs by $\ln \pi = 2\lambda$.

Subject covariates can also be included in the model. We assume that each distinct combination of covariates observed will form a covariate set (Francis *et al*., 2010); assume that there are S such sets ($1 < S <= N$). To model the effect of the covariates, the number of unique response patterns U now become UK response patterns. The number of times the ℓth response pattern occurs within each covariate set s is denoted by $N_{\ell s}$. The linear predictor η becomes

$$\eta_{\ell s} = \sum_{j<k} y_{jk;\ell s}(\lambda_{is} - \lambda_{js}).$$

2. Non-parametric random effects for partial ranks

We now extend this model to allow for repeated observations of responses over time, which will allow us to examine change over time in the ranked responses. The model takes account of attrition over the sweeps of the survey through a full information maximum likelihood approach which assumes an underlying missing at random process. The resulting model uses a nonparametric formulation of the random effects structure (individuals nested within time) fitted using the EM algorithm (Aitkin, 1999). Each discrete mass point is multivalued, with a parameter for each item (Francis *et al.*, 2010). The resultant model is equivalent to a latent class regression model, where the latent class profiles are provided by the mass point components and the covariates act on the class profiles. This provides an alternative interpretation of the fitted model. The linear predictor for this latent class approach with covariates would extend to $\eta_{\ell s} = \sum_{j<k} y_{jk;\ell s}(\lambda_{js} + \delta_{j\ell s} - \lambda_{ks} - \delta_{k\ell s})$ where the location of the preference parameter for item j will be shifted up or down by $\delta_{j\ell s}$ for each response pattern ℓ and covariate set s.

3. An example

We take two questions from the British Household Panel Survey (Buck *et al.*, 1994) which measure materialistic and postmaterialistic values (Inglehart, 1977) in consecutive sweeps of the British Household Panel Survey. As part of the survey, respondents

> In politics it is not always possible to obtain everything one might wish.
> On this card several goals are listed.
> If you had to choose among them, which would be your first choice?
>
> | | Highest Priority |
> |---|---|
> | Maintain order in nation | 1 |
> | Give people more to say in government decisions | 2 |
> | Fight rising prices | 3 |
> | Protect freedom of speech | 4 |
> | Can't choose | 8 |
>
> And which would be your second choice?

Figure 1. The operationalization of the Inglehart Index used in the British Household Panel Survey

were asked to choose the most preferred and the next most preferred out of a set of four items representing preferred political priorities for the individual (Figure 1).

Acknowledgements: This work was partially funded by the UK ESRC National Centre for Research Methods initiative (grant number RES-576-25-0019). The authors would like to thank Walter Katzenbeisser for very helpful discussions.

References

Aitkin M.A. (1999), A general maximum likelihood analysis of variance components in generalized linear models, *Biometrics*, 55, 117–128.

Buck N., Gershuny, J., Rose D., Scott J. (1994), *Changing Households: The British Household Panel Survey 1990-1992*, Colchester: ESRC Research Centre on Micro-Social Change, University of Essex.

Critchlow D., Fligner M. (1991), Paired Comparison, Triple Comparison and Ranking Experiments as Generalized Linear Models and Their Implementation in GLIM. *Psychometrika*, 56, 517-533.

Dittrich R., Katzenbeisser W., Reisinger H. (2000), The analysis of rank ordered preference data based on Bradley-Terry Type Models, *OR Specktrum*, 22, 117-134.

Dittrich R., Francis B., Hatzinger R., Katzenbeisser W. (2007), A paired comparison approach for the analysis of sets of Likert scale responses, *Statistical Modelling*, 7, 3–28

Francis B., Dittrich R., Hatzinger R. (2010), Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: How do Europeans get their scientific knowledge? *Annals of Applied Statistics*, 4, 2181–2202.

Francis B., Dittrich R., Hatzinger R., Penn R. (2002), Analysing ranks using paired comparison methods: an investigation of value orientation in Europe, *Applied Statistics*, 51, 319–336.

Inglehart R. (1977), *The Silent Revolution: Changing Values and Political Styles among Western Publics*, Princeton: Princeton University Press.

Thurstone L.L. (1927), A law of comparative judgement, *Psychological Review*, 34, 273-286.

Quaderni di Statistica
Vol. 14, 2012

Estimating sectoral and smoothed European growth by Generalized Dynamic Factor Model

Antonio Frenda
Doctoral Program in Computational Methods for forecasting and decisions in Economics and Finance, University of Bergamo (Italy)
E-mail: frenda@istat.it

Summary: New Eurocoin (NE) indicator, that is produced by the Bank of Italy, can be described through the projection of the whole Euro bandpassed gross domestic product on a set of regressors - the linear combination of variables contained in the Thomson Financial Datastream. NE provides an index of the current economic situation in the Euro Area, extracting from the Data Source relevant information which represent the main sources of variation, in order to track the entire underlying GDP for the whole Euro area, thus it is an "aggregate indicator". The main aim of this paper is to propose a new theoretical framework for disaggregated business cycle analysis by sectors.

Keywords: Band-pass filters, Sectoral smoothed growth, Real time performance.

1. General remarks

New Eurocoin indicator, that is published monthly by the Bank of Italy and CEPR, provides a summary index of the medium to long-run component (MLRG) of the whole GDP for the aggregate Euro area. The innovation of this research are some procedures, based on Eurocoin methodology, to estimate sectoral MLRG concerning Euro Area. The most important indicator of economic activity is the GDP (gross domestic product) and, unlike the industrial production (IP), it comprises services, agriculture, public sector. Differently from surveys, the GDP does not contain any subjective assessment. And, unlike IP, GDP is defined only quarterly and with a lag. Monthly indicators are commonly used in the prediction of current data on GDP, before the data are available. The main groups of indicators generally used are: Surveys, Financial market data; Labour market data; Monetary aggregates; Industrial production; Prices; Demand Indicators; Foreign Trade. The Eurostat Handbook on Quarterly National Accounts (2000) defines

a flash estimate: "the earliest picture of the economy according to national accounts concepts, which is produced and published as soon as possible after the end of the quarter, using a more incomplete set of information than that used for traditional quarterly accounts". A flash estimate of GDP is released by Eurostat about six weeks after the end of the reference quarter. "As errors in Flash estimates tend to be relatively large, National Statistical Institutes tend to release them at higher level of aggregation than preliminary Quarterly National Accounts estimates". In order to have more disaggregated and timely data, we want to derive in this paper a "Sectoral Eurocoin". A sectoral estimate of GDP is released by Eurostat about ten weeks after the end of the reference quarter. Then, we produce a real-time monthly estimate of growth purified from erratic components, following sectoral breakdown of quarterly gross value added available from Eurostat and European Central Bank. Sectoral composition influences the characteristics of a business cycle, such as its length and amplitude; factor model is used to meet the macroeconomic behavior on the basis of disaggregated data (the sectors). Using generalized dynamic factor model we will outline in this paper a disaggregated analysis of the medium to long run component of GDP (MLRG) in the European economy, studying interrelations and characteristics of sectoral growth rates. The main macro sectors analyzed: Manufacturing, Energy, Construction, Financial, Trade-Transport-Communication. In the dynamic factor model approach, a vector of "n" time series is decomposed into two mutually orthogonal components: a common component characterized by few common factors or latent shocks, and a component "idiosyncratic", led by n specific shocks (one for each variable in the panel). These models allow a net reduction of cross-sectional size of the dataset. It is a well-known result in the literature isolating the business cycle in integrated series that band-pass filter could deteriorate at the end of the sample. Altissimo, Cristadoro, Forni, Lippi, Veronese (2008) show that the same problem arises with application to stationary time series. And, through New Eurocoin Indicator, they develop a method to obtain smoothing of a stationary time series so as to avoid the occurrence of end-of-sample deterioration (Eurocoin is an advancement in estimating of smoothed GDP). New Eurocoin produces real-time monthly estimate of GDP growth, purified from erratic components (short-run fluctuations). Band-pass filters can eliminate erratic components as they are infinite moving averages and are based on past and future values of GDP (see Baxter and King, 1999, pp. 579-80, Christiano and Fitzgerald, 2003, pp. 459-60). However, unlike Eurocoin, they are less reliable for the most recent data, very relevant for economic policy. Eurocoin is an alternative estimate to c_t, the medium to long-run component of GDP. In this paper a sectoral version of the Eurocoin indicator is proposed, because actually Eurocoin is only used to outline aggregate European growth and it provides an estimate of the medium to long-run component (MLRG) of the GDP only for the whole aggregate Euro area. The primary objective of this paper is the one to produce smoothed growth indicators that describe the behaviour of economic activity for a large number of sectors at the monthly frequency, while utilizing a wide range of economic time series in a timely fashion. The innovation of this research are some procedures, based on Eurocoin methodology, to estimate MLRG concerning

European sectoral growth: in fact, the 157 basic variables that now constitute the complete dataset from Thomson Financial Datastream, used to process Eurocoin indicator, belong to different homogeneous data groups. Some series ignore large portions of economic activity (e.g industrial production and export), and all these series exibit heavy short-run fluctuations and could provide not coherent signals, so "there is much diversity and uncertainty about which indicator are to be used" (Zarnowitz and Ozyildirim, 2002). That is why, therefore, it is necessary to outline a more specific analysis of MLRG. The main strategy used has been the projection of sectoral added value on European factors (which are the combination of the 157 variables contained in the Thomson Financial Datastream, and used by the Bank of Italy to build Eurocoin). Our estimates are tested by pseudo real time performance. Following ECB breakdown of gross domestic product we consider that "Industry" business cycle include the following sectors:

- Manufacturing;
- Energy;
- Constructions.

"Service" branch include these sectors:

- Trade;
- Transports;
- Communications;
- Financial.

This paper is based on my PhD Thesis that has as Supervisors Prof. Marco Lippi (University La Sapienza, Rome) and prof. Giovanni Urga (University of Bergamo).

References

Altissimo F., Cristadoro R., Forni M., Lippi M., Veronese G. (2010), New Eurocoin: Tracking Economic Growth in Real Time, *The Review of Economics and Statistics*, MIT Press, vol. 92(4), pages 1024–1034.

Baxter A., King R.G. (1999), Measuring Business Cycle Approximate Band-Pass filters for Economic Time Series, *The Review of Economics and Statistics*, 81, 575–9343.

Christiano L.J., Fitzgerald T.J. (2003), The Band-Pass Filter, *International Economic Review*, 84, 435–465.

Forni M., Hallin M., Reichlin L. (2000), The generalized dynamic factor model: identification and estimation, *The Review of Economics and Statistics* 82, 540–554.

Forni M., Hallin M., Lippi M., Reichlin L. (2001), Coincident and leading indicators for the euro area, *The Economic Journal*, 111, 62–85.

Forni M., Lippi M. (2001), The generalized dynamic factor model: representation theory, *Econometric Theory*, 17, 1113–1141.

Forni M., Hallin M., Lippi M., Reichlin L. (2004), The generalized dynamic factor model: consistency and rates, *Journal of Econometrics*, 119, 231–255.

Forni M., Hallin M., Lippi M., Reichlin L. (2005), The generalized dynamic factor model: one-sided estimation and forecasting, *Journal of the American Statistical Association*, 100, 830–840.

Quaderni di Statistica
Vol. 14, 2012

CCA analysis: A statistical approach applied to ecological process. Relationships among decomposition rates, biological diversity and substrate fractal dimension

Gina Galante
Dept. of Environmental Biology, "Sapienza" University of Rome
E-mail: gina.galante@uniroma1.it

Biancamaria Pietrangeli Domenico Davolos
INAIL- DIPIA
E-mail: biancamaria.pietrangeli@ispesl.it, domenico.davolos@ispesl.it

Oriana Maggi Edoardo Scepi
Dept. of Environmental Biology, "Sapienza" University of Rome
E-mail: oriana.maggi@uniroma1.it, edoardo.scepi@email.it

Rossana Cotroneo
ISTAT
E-mail: cotroneo@istat.it

Stefano Domenico Cicala
University of Sannio
E-mail: sdcicala@gmail.com

Summary: Ecological processes are both influenced by biotic and abiotic factors. Substrate morphology and characteristics may also influence benthic decomposers abundances and distribution. Fractal dimension (Taniguchi & Tokeshi, 2004) can give a measure of substrates surfaces complexity and may also be related to both water turbulence and macroinvertebrates clinging. In this study the functional relationships among macroinvertebrates, bacteria diversity and abundances, the water chemical and physical parameters, the rocks, pebbles and stones numbers and dimension and fractal dimension of substrate were investigated in the river Sacco during the spring season.

Keywords: Fractal dimension, Random forest, CCA analysis.

1. General remarks

Three sampling sites were selected in Sacco area along 5 km stretch of the rive.The study started on 21 May 2009 and run over 3 weeks using litter bags technique, both coarse and fine mesh size bags were used to detect respectively macroinvertebrates and bacteria (Graca et al, 2007). A total of 150 bags were sealed and randomly distributed at the sampling sites. Bags were positioned both in riffles and pools areas, well submerged, tied to rocks and stones with fishing nylon wire. Dissolved oxygen and water temperature were relieved from each sampling station and bag position point. Triplicate of fine and coarse mesh bags were retrieved from each sampling site weekly. Leaf breakdown rate (k) was estimated by fitting the amount of remaining leaf material data to the exponential model, $Y_t = Y_0 e^{-kt}$, where Y_t is the AFDM remaining at time t (days), and Y_0 the AFDM at the beginning of the experiment (Petersen & Cummins 1974). Leaves mass losses have been estimated for each litter bag using the relation $ln(Y_t/Y_0) = -Kt$. In decomposition rates curve fitting AFDM has been expressed like percent of the remains mass at any sampling date. Substrate characteristics were detected by overlaying a one meter wide plastic square upon each litter bag and counting number and dimensions of stones, rocks and pebbles inside the square, then substrate selected areas were photographed (pixel resolution 10mm). Five classes of rocks, stone and pebbles were identified: rocks $> 25cm$; rocks$= 20cm$; stones$= 15cm$, pebbles$= 10cm$ and number of pebbles$< 10cm/cm^2$.The acquired images were both transformed and processed to eliminate water reflex, then exported a and elaborated with Fractal-3 software to calculate fractal dimension (FD) using the box-counting method, a quantitative analysis of perimeter convolution. This algorithm, commonly known as Hausdorff Dimension (H.D.) is :

$$H.D. = \frac{logN}{logN(s)} \tag{1}$$

and estimate the aggregate perimeter as FD that describes the complexity of an object (Carr & Benzer, 1991)(fig.1).

The leaves retrieved from coarse mesh bags were rinsed into a 250 μm mesh screen to retain the associated macroinvertebrates, which were sorted and collected in ethanol (70%v/v) until identification and counting. Macroinvertebrates were identified by stereomicroscope and diversity indices were elaborated. A correlation matrix was elaborated to relate decomposition rates to: substrate characteristics, fractal dimension of substrate, SHDI (Shannon Diversity index), SHEI (Shannon Diversity Index) diversity indices, S-species richness, total abundances, dissolved oxygen and water temperature. Rocks, stones, and both pebbles dimension and number of small pebbles in a 10 cm^2 width surfaces was also considered as variables. The same parameters were used to perform a CCA (Canonical Correspondence Analysis) that is a multivariate method widely used in

ecology field. The method is designed to extract synthetic environmental gradients from ecological datasets.The result is that the axes of the final ordination, rather than simply reflecting dimensions of the greatest variability in the species data, are restricted to the linear combinations of the environmental variables and the species data. We underline that, before proceeding to apply the CCA model, we also applied a supervised data pre-processing to reduce redundant variables. The variables selection was carried out using a data mining techniques:random forest (Breiman, 2004).

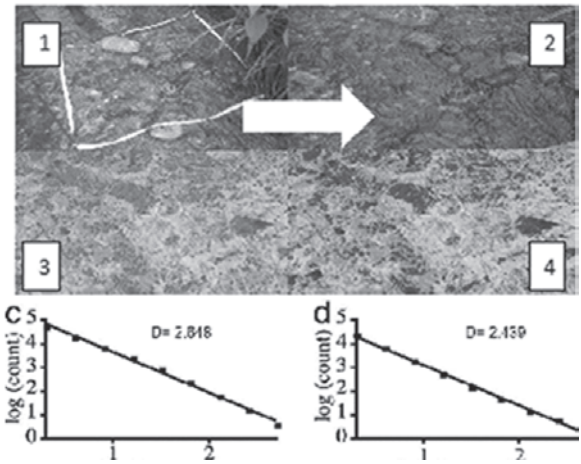


Figure 1. Procedure of acquisition and elaboration of substrate FD. 1: plastic square positioned in sampling site; 2, 3, 4: elaboration steps. In the bottom two graphical examples of box counting

2. Results and Discussions

CCA analysis (AFDM target variable) highlights that pebbles dimension and FD of substrate are affected directly by decomposition rates and macroinvetebrates abundance. For the first axe there is a high statistical meaningful value for abundances' total and pebbles. AFDM decrease at the total abundances increase and in presence of pebbly substrate. As regard the second axe, has been found that stones = 15 cm, fractal dimension of substrate and temperature are negative correlate with AFDM, while total abundances and pebbles are positive correlate. Decrements of both temperature and fractal dimension of substrate influences the decomposition process. In other words, low temperature and low fractal dimension of substrate inhibit the decomposition process(fig.2).

Diversity indices seem to highlight that decomposition process is driven mainly by abundance, (relative and total). Species abundances are linked to temperature and substrates morphology and both seems to influence breakdown rates. In particular, few big

rocks contribute to reduce FD of substrate while small and medium pebbles and stones obviously increment this parameter. The macroinvertebrates assemblage in relation to substrate heterogeneity in a stream ecosystem was recently investigated (Boyero, 2003), but the fractal dimension of substrate was not taken into account, and the effects of it on litter breakdown processes were ignored. FD is linked to abundances and decomposition rates probably due to its contribution to the availability of refuges and hanging and feeding surfaces for invertebrates and to water turbulence (physical fragmentation). The more complex is substrate, the more abundant are species and faster K. Species diversity seems to be not influenced by FD. Many local factors could influence benthic fauna diversity, both temperature and substrate characteristics appear to be important: the high temperature variation during the experimental time could have influenced species diversity due to the specie-specific temperature limitation. While FD-K and abundances seems to be inferred, the relationship between FD and diversity needs to be deeply investigate.

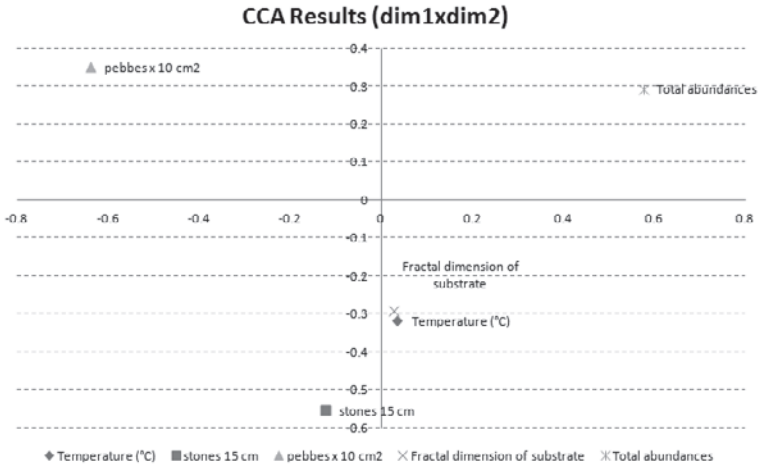


Figure 2. CCA results

References

Breiman L. (2004), Random forests, *Machine learning*, 45, 5–32.

Graca M.A.S., Barlocher F., Gessner M.O. (2007), *Methods to study bitter decomposition: a practical guide*, Springer-Verlag, New York.

Taniguchi H., Tokeshi M. (2004), Effects of habitat complexity on benthic assemblages in a variable environment, *Freshwater Biology*, 49, 1164–1178.

Quaderni di Statistica
Vol. 14, 2012

A three-way analysis for compositional data: insights on Arno river (Tuscany, central Italy) water chemistry

Michele Gallo
Department of Human and Social Sciences, University of Naples L'Orientale
E-mail: mgallo@unior.it

Antonella Buccianti
Department of Earth Sciences, University of Florence
E-mail: antonella.buccianti@unifi.it

Summary: Based on the availability of an extensive three-way array regarding a water sampling carried out in the Arno river basin (central-northern Apennines, Italy), we are proposing the Tucker3 model for compositional data in order to study the complex system affected by several physical-chemical processes, either natural or attributable to anthropogenic phenomena. Furthermore, several graphical procedures are given to show the geochemical results of the Tucker3 model when it is applied to the compositions.

Keywords: CoDa, Tucker3 model, Joint biplot.

1. General remarks

Compositional data (CoDa) have important and particular properties that preclude the application of standard statistical techniques in their raw form (%, ppm, mg/L and so on), thus affecting the results obtained by monitoring survey in space and time (Buccianti and Pawlowsky-Glahn, 2005). Changes in the chemical composition of the river water chemistry observed in their development in space and time, can be analysed by using a three-way array. In these cases, three-way models, such as Tucker models (Tucker, 1966), can be used to explore the complex relationships among chemistry, time and space, also requiring that the particular properties of compositional data are satisfied, i.e., scale invariance and subcompositional coherence. The scale invariance here merely reinforces the intuitive idea that compositional data provides information only about

relative and not absolute values. On the other hand, the subcompositional coherence requires that a result obtained by investigating a full composition or some of its subcompositions present the same relations within the common parts. Gallo (2011a) has shown how the Tucker3 analysis can be correctly applied to the CoDa, and here a practical application is presented and discussed thanks to the availability of an extensive three-way array. The data are related to a sampling of surficial waters carried out in the Arno river basin between May 2002 and October 2003 (Nisi et al, 2008). This case study, represented by a complex natural system affected by several physical-chemical processes, either natural or attributable to anthropogenic phenomena, allowed us to explore how a CoDa Tucker3 analysis works in a low-dimensional framework.

2. Theory

2.1. Compositional data and preprocessing

Let $v'_1, ..., v'_J$ be positive quantities with the same measurement scale, the vector $\mathbf{v}'= (v'_1, ..., v'_J)$ is the basis of compositional data and $\mathbf{v} = \mathbf{v}'||\mathbf{v}'||$ is a composition vector, where $||.||$ is the norm of the vector. The natural sample space for compositional data is the unit simplex S^J defined by $S^J = \{v_1, ..., v_J : v_1 > 0, ..., v_J > 0; \sum_{j=1}^{J} v_j = \delta\}$. According to the principle that compositional data give information about relative, not absolute values of components, it is appropriate to consider the logarithms of the ratios, called log-ratios, to transform the compositional data before using standard unconstrained multivariate analysis. Aitchison (1986) proposed the use of the log-ratios between the part of a composition and the relative geometric mean. Through this transformation, a direct association between the simplex sample space and the real space is found. In this way, it is possible to work in real space, where it is easier to apply statistical methods, and later, by using inverse functions, the results can be projected back to the simplex space.
When compositional vectors are observed on several occasions they can be arranged in a three-way array as rows. Thus, the three-way array $\mathbf{V}$ (I x J x K) has I compositional vectors with J-dimensions, and observed on K different occasions. Moreover, we define $\mathbf{V}_k$, known as frontal slice, the matrix of data observed at the occasion k obtained by fixing the third mode of $\mathbf{V}$. By considering the transformations proposed by Aitchison (1986) and Egozcue et al. (2003) (see Gallo 2011a for more detail), the logarithmic transformation can be applied to a three-way array, thus $\mathbf{L}(I$ x J x $K)$ is an array with typical element $log(v_{ijk})$, and $\mathbf{L}_k$ is the kth frontal slice $(k = 1, ..., K)$. The kth frontal slice of the centred log-ratios can be written as $\mathbf{L}_k\mathbf{P}_J^{\perp}$, where $\mathbf{P}_J^{\perp} = (\mathbf{I}_J - \mathbf{1}_J\mathbf{1}_J^t/J)$ is a symmetric and idempotent centring matrix. To ensure that log-ratios are centred with respect to column means that each frontal slice is premultiplied by the symmetric and idempotent centring matrix $\mathbf{P}_I^{\perp} = (\mathbf{I}_I - \mathbf{1}_I\mathbf{1}_I^t/I)$. Thus, the kth frontal slice of centred log-ratios is $\mathbf{Y}_k = \mathbf{P}_J^{\perp}\mathbf{L}_k\mathbf{P}_I^{\perp}$. These frontal slices can be concatenated between them

obtaining the following matricizing of the three-way array as $\mathbf{Y}_A = [\mathbf{Y}_1|...|\mathbf{Y}_k|...|\mathbf{Y}_K]$.

2.2. Tucker3 for CoDa

Let $\mathbf{G}$ (P x Q x R) and $\mathbf{E}$ (I x J x K) be the core and the residual arrays, with ($P < I$), ($Q < J$) and ($R < K$), respectively. In the Kronecker product notation, the Tucker3 model can be written as

$$\mathbf{Y}_A = \mathbf{A}\mathbf{W}_A = \mathbf{A}\mathbf{G}_A(\mathbf{C}^t \otimes \mathbf{B}^t) + \mathbf{E}_A \tag{1}$$

where $\mathbf{A}$ (I x P), $\mathbf{B}$ (J x Q) and $\mathbf{C}$ (K x R) are the loadings matrices for $first$-, $second$- and $third$-modes; $\mathbf{G}_A$ is the matricized core-array where the generic element gives the interaction between factors, and $\mathbf{E_A}$ is the matricized residual array. In order to estimate the parameters of the model, several algorithms are proposed in literature (see Gallo 2011a, Smilde et al., 2004). These algorithms give a base $\mathbf{W}_A = (\mathbf{C} \otimes \mathbf{B})\mathbf{G}_A^t$ that it is just orthogonal and not orthonormal: $\mathbf{W}_A^t\mathbf{W}_A = \mathbf{G}_A\mathbf{I}_{QR}\mathbf{G}_A^t$. In this form, the representation of the I points in the subspace do not preserve the Euclidean distances. Fortunately, in the three-way Tucker analysis, the solution obtained is not uniquely determined and any arbitrary non-singular transformation of loadings matrices may be compensated by the inverse transformation applied to the core arrays (Kroonenberg, 2008). Thus, let $\mathbf{T_A}$ be a transformation matrix so that $\mathbf{W}'_A = \mathbf{W}_A\mathbf{T_A} = (\mathbf{C} \otimes \mathbf{B})\mathbf{G}_A^t\mathbf{T}_A$ is columnwise orthonormal, i.e. $\mathbf{T}_A$ is given by the Gram-Schmidt orthonormalization, therefore we have $\mathbf{Y}'_A = \mathbf{A}'\mathbf{W}_A^t + \mathbf{Y}_A = \mathbf{A}\mathbf{W}_A^t + \mathbf{Y}_A$ with $\mathbf{A}' = \mathbf{A}(\mathbf{T}_A^t)^{-1}$. In this case, $\mathbf{W}'_A$ is an orthonormal base for the principal coordinates $\mathbf{A}'$. Of course, due to the symmetry of the Tucker analysis, the same procedure can be used for $\mathbf{B}$ and $\mathbf{C}$.

3. Case study: results and discussion

The main chemical composition of the main Arno river water (Na^+, K^+, Mg^{2+}, $Ca^{2+}, Cl^-, SO_4^{2-}, HCO_{3^-}, SiO_2, NH_{4^+}, NO_{2^-}, NO_{3^-}$ mg/L or ppm) has been analyzed by using the Tucker3 model for compositional data. The samples have been collected from 22 spatial coordinates (distance from the spring) over four different periods (from May 2002 to October 2003). In accordance with the approach discussed in Section 2, the data have been analyzed with the aim of studying the complex system affected by several physical-chemical processes, either natural or attributable to anthropogenic phenomena. The results obtained by the application of the proposed procedure can be summarized by using several graphical tools as joint biplots, one-mode plots and pre-component plots (Gallo, 2011b). Interesting results were obtained by using the joint biplot where the values of the first and second log-contrast are related to vectors whose lengths are associated with the variance of the ratio between compositional variables. In our case it was possible to verify that vectors characterised by higher lengths were con-

stantly associated with the presence of Cl^-, Na^+, SO_4^{2-}, as well as with the presence of oxidized and reduced forms of the nitrogen species, which are variables highly affected by pollution processes, that are able to perturb the natural range of values attributable to natural phenomena (Berner and Berner, 1996). By using these graphical tools it was consequently possible to visualize, immediately, that the natural relationships among the components of the composition have been modified by anthropogenic processes in the observed period of time due to weathering natural processes, attributable to the lithology and geology of the basin, and therefore, revealing a framework similar to that found in other polluted rivers of the European area.

References

Aitchison J. (1986), *Statistical Analysis of Compositional Data*, Chapman and Hall, London.

Berner E., Berner R. (1996), *The Global Environment: Water, Air and geochemical Cycles*, Prentice Hall, New Jersey.

Buccianti A., Pawlowsky-Glahn V. (2005), Water Chemistry and Compositional Data Analysis. New perspectives of investigation, *Geology*, 37, 703–727.

Buccianti A. (2011), Natural laws governing the distribution of the elements in geochemistry: the role of the log-ratio approach, in: Pawlowsky-Glahn and Buccianti (eds.), *Compositional Data Analysis. Theory and Applications*, John Wiley Sons, 255–266.

Egozcue J.J., Pawlowsky-Glahn V., Mateu-Figueras G., Barcelo-Vidal C. (2003), Isometric logratio transformations for compositional data analysis, *Mathematical Geology*, 35, 279–300.

Gallo M. (2011a), Tucker3 analysis of compositional data, *Submitted*.

Gallo M. (2011b), Log-ratio and parallel factor analysis: an approach to analyze three-way compositional data, in: Squillante, Proto and Kacprzyk (eds.), *Advances in intelligent and soft computing*, Springer.

Kroonenberg P.M. (2008), *Applied multiway data analysis*, Wiley, Hoboken.

Nisi B., Vaselli O., Buccianti A., Minissale A., Delgado-Huertas A., Tassi F., Montegrossi G. (2008), Geochemical and isotopic investigation of the dissolved load in the running waters from the Arno valley: evaluation of the natural and anthropogenic input, in Nisi B. (eds.), *Memorie Descrittive della Carta Geologica d'Italia*, LXXIX, 160 pp.

Smilde K.A., Bro R., Geladi P. (2004), *Multi-way analysis: applications in the chemical sciences*, Wiley, Chichester.

Tucker L.R. (1966), Some mathematical notes on three-mode factor analysis, *Psychometrika*, 31, 279–311.

Quaderni di Statistica
Vol. 14, 2012

A semiparametric approach to source separation using Independent Component Analysis

Sujit Ghosh
Department of Statistics, North Carolina State University
E-mail: sujit.ghosh@ncsu.edu

Ani Eloyan
Department of Biostatistics, Johns Hopkins University
E-mail: aeloyan@jhsph.edu

Summary: Data processing and source identification using lower dimensional hidden structure plays an essential role in many fields of applications where large datasets are often encountered. One of the common methods for source separation using lower dimensional structure involves the use of Independent Component Analysis (ICA), which is based on a linear representation of the observed data in terms of independent hidden sources. This paper first presents a set of sufficient conditions to establish the identifiability of the sources and the mixing matrix using moment restrictions of the hidden source variables. The consistency of the proposed estimate is established under additional mild regularity conditions. The proposed method is illustrated and compared with existing methods using simulated data scenarios and using data sets on brain imaging.

Keywords: Constrained EM-algorithm, Density Estimation, Source Identification.

1. Introduction

The problem of finding a representation of multivariate random variables which maintains its essential distributional structure using a set of lower dimensional random variables has been of interest to researchers in statistics, signal processing and neural networks. Such representations of higher dimensional random vector using a lower dimensional vector provide a statistical framework to the identification and separation of the sources. Since the linear transformations of data are computationally and conceptually easier to implement, most of the methods are based on finding a linear transfor-

mation of the data. Some of the major approaches for solving this problem include principal component analysis (PCA), factor analysis (FA), projection pursuit (PP) and independent component analysis (ICA). A distinguishing feature of the ICA compared with other source separation methods is that the lower dimensional random variables are extracted as independent sources in contrast to uncorrelated random variables (e.g., as in PCA). Some of the early approaches to ICA are based on estimating the mixing matrix of the linear transformation by the maximization of the mutual information or the negentropy function (see Comon (1994) for details). Other methods for estimating the mixing matrix are based on gradient algorithms or cumulant functions which are described in detail by Hyvarinen et al. (2001), Cardoso and Souloumiac (1993) and the references therein. A general formulation of the source separation problem can be presented as follows. Given a random sample $X_1, \ldots, X_T$, where $X_i = (x_{i1}, \ldots, x_{in})^T$ are independent and identically distributed random vectors, can we find a unique transformation $g : \mathbb{R}^n \to \mathbb{R}^m$ for some $m \leq n$ and densities $f_1, \ldots, f_m$ such that $X_i \stackrel{d}{=} g(S)$ where $S = (s_1, \ldots, s_m)^T$ is the vector of independent sources, in other words, $s_j \stackrel{ind}{\sim} f_j(\cdot)$ for $j = 1, \ldots, m$ and $i = 1, \ldots, T$. A special case emerges when the relationship is assumed to be linear. Then $g(S) = AS$ and the problem reduces to the estimation of the mixing matrix A and the probability densities $f_1, \ldots, f_m$. In matrix notation, a model for ICA can be written as,

$$X = AS + E, \tag{1}$$

where $X = (x_1, \ldots, x_n)^T, S = (s_1, \ldots, s_m)^T, A = (a_{ij})_{n \times m}$ and $E = (e_1, \ldots, e_n)^T$ is an $n \times 1$ vector of independent gaussian noise variables each with mean 0. Writing $B = (A \;\; I)$ and $Y = (S^T \;\; E^T)^T$ we can equivalently express (1) as $X = BY$. In most of the early literature on likelihood based solutions to the ICA, the densities of the independent sources are prespecified parametrically and are chosen based on the fact that these densities should be nongaussian. Boscolo (2004) proposed a pseudo-likelihood based method for ICA using model (1) where the densities of the sources are estimated nonparametrically by using the kernel density estimate. It was shown that simultaneously estimating the densities of the sources along with the mixing matrix improves the estimation compared to some of the parametric approaches. A more recent nonparametric approach to the linear ICA model (1) proposed by Chen and Bickel (2006) is based on score functions. However, most of these previous methods for ICA are difficult to compare mutually, even based on simulated data sets, when the mixing matrix A and the source densities $f_1, \ldots, f_m$ are not uniquely identified.

In most of the commonly used algorithms for ICA the fact that a model for ICA is not fully identifiable is often completely ignored (e.g., FastICA, JADE). In this paper we derive the conditions for the uniqueness of the linear representation (1) by imposing a set of minimal moment constraints on the distributions of the independent sources. We then use a newly proposed semi-parametric density estimation method based on a suitable class of mixture densities that allows to conserve the moment restrictions needed for identifiability and establish consistency of our proposed estimator of W. Next, we present an iterative method for computing the proposed estimate of the unmixing matrix

W and the source densities $f_1, \ldots, f_m$ simultaneously. Finally, we present empirical analysis based on simulated data and compare the performance of our method to three existing competitive methods for which software are available and an illustrative example based on a real dataset.

2. Parameter Identifiability of the ICA

The statistical estimation of the mixing matrix A (or its inverse W) and the source densities $f_1, \ldots, f_m$ remains an ill-posed problem until the 'true parameters,' the mixing matrix and the source densities are uniquely defined in the statistical model given by (1). The following result provides a set of sufficient conditions under which the ICA model has a solution and it is unique.

Theorem 1. *Suppose the mixing matrix A in noisy ICA model (1) is of full column rank and the independent sources are all nongaussian. Further, suppose that all of the third moments of the sources exist and they satisfy the following conditions,*

$$\begin{gathered} E(s_j) = 0, \;\; E(s_j^2) = 1, \textit{ for } j = 1, \ldots, m \textit{ and} \\ 0 < E(s_1^3) < E(s_2^3) < \ldots < E(s_m^3) \end{gathered} \tag{2}$$

then the ICA model is fully identifiable, in the sense if X has two representations given by $X = AS + E = BG + \widetilde{E}$, where A and B are each of full column rank and both S and G satisfy the conditions in (2), then $A = B$ and $S \stackrel{d}{=} G$.

In some applications, we often have subject matter knowledge that the original sources are positive valued random variables. Since the third moment of a positive valued random variable is necessarily positive, the following result can be obtained immediately from Theorem 1.

Theorem 2. *Suppose $s_1, \ldots, s_m$ in the ICA model (1) are positive valued random variables with $E(s_j) = 1$, for $j = 1, \ldots, m$. Then the model is fully identifiable if* $\mathrm{var}(s_1) < \ldots < \mathrm{var}(s_m)$.

Next we show that the sufficient conditions stated in Theorem 1 are minimal if we assume that skewness of the source variables are distinct.

Theorem 3. *The conditions (2) given in Theorem 1 are minimal if the sources $s_1, ..., s_m$ in the ICA model (1) are assumed to have third order moments and further assuming that the skewness measures of the densities are distinct.*

3. A Semiparametric ICA Model

By Theorem 1 a set of sufficient conditions requiring the existence of the third moments of densities $f_1, \ldots, f_m$ in ICA model makes the mixing matrix A identifiable.

Eloyan and Ghosh (2011a) developed a flexible class of models based on a mixture of densities (for instance gaussian kernels) to estimate a univariate density subject to moment constraints. We extend their method to the multivariate case for estimating the densities of the independent components of S. Following their work, we propose to estimate each of the source densities f_j by the following mixture of densities

$$f_j(s) = \sum_{k=1}^{N_j} \theta_{jk} \phi \left(\frac{s - \mu_{jk}}{\sigma_{N_j}} \right) \frac{1}{\sigma_{N_j}}, \tag{3}$$

where $\mu_{j1} < \mu_{j2} < \ldots < \mu_{jN_j}$ is a suitable sequence of known numbers (knots) and $\sigma_{N_j} > 0$ is chosen as a function of μ_{jk}'s and N_j, $\phi(\cdot)$ is a kernel density function satisfying a set of regularity conditions. Given the μ_{jk}, σ_{N_j} and N_j, the weights θ_{jk} are estimated subject to a set of restrictions implying that $f_j, j = 1, \ldots, m$ satisfy a set of sufficient conditions for identifiability (e.g., as in (2) in Theorem 1). In particular, in order to satisfy the set of three conditions given in (2), we estimate the θ_{jk}'s subject to the following necessary conditions: (i) $\sum_{k=1}^{N_j} \theta_{jk}\mu_{jk} = 0$, $\sum_{k=1}^{N_j} \theta_{jk}\mu_{jk}^2 = 1 - \sigma_{N_j}^2$ and $\sum_{k=1}^{N_j} \theta_{jk}\mu_{jk}^3 > 0$; and (ii) $\sum_{k=1}^{N_j} \theta_{jk} = 1$ and $\theta_{jk} \geq 0$. Detailed proofs of the theorems and emipirical illustrations are available in the technical report Eloyan and Ghosh (2011b).

References

Boscolo R., Pan H., Roychowdhury V. (2004), Independent component analysis based on nonparametric density estimation, *IEEE Transactions on Neural Networks*, 15, 55–65.

Cardoso J.F., Souloumiac A. (1993), Blind beamforming for non gaussian signals, *IEE-Proceedings-F*, 140, 362–370.

Chen A., Bickel P. (2006), Efficient independent component analysis, *The Annals of Statistics*, 34, 2825-2855.

Common P. (1994), Independent component analysis, a new concept?, *Signal Processing*, 36, 287–314.

Eloyan A., Ghosh S.K. (2011a), Smooth density estimation with moment constraints using mixture densities, *Journal of Nonparametric Statistics*, 23, 513–531.

Eloyan A., Ghosh S.K. (2011b), A Semiparametric Approach to Source Separation Using Independent Component Analysis, NC State University Technical Report# 2635.

Hyvarinen A., Karhunen J., Oja E. (2001), *Independent Component Analysis* John Wiley and Sons, New York.

Quaderni di Statistica
Vol. 14, 2012

Evaluate the magnitude of non uniform DIF for Rasch model: the polytomous case

Silvia Golia
Department of Quantitative Methods, University of Brescia
E-mail: golia@eco.unibs.it

Summary: Differential Item Functioning (DIF) is understood to be present when something about the characteristics of a test taker interferes with the relationship between ability and item responses. Non uniform DIF exists when there is interaction between the subject ability level. The present study proposes a way to assign a severity grade to DIF when non uniform DIF is present.

Keywords: DIF magnitude, Non Uniform DIF, Rating Scale Model.

1. Introduction

The assessment of differential item functioning (DIF) is become an important component in the study of latent variables done making use of data coming from tests. DIF is understood to be present when something about the characteristics of a test taker interferes with the relationship between ability and item responses. The presence of DIF represents a violation of the measurement invariance condition, but in a restricted form that pertains to the conditional dependence of group membership and item response. In a DIF analysis, the population is divided in two subgroups; the group of primary interest, or focal group, is compared to a reference group.

Two types of DIF can be identified: uniform and non uniform. Uniform DIF (UDIF) occurs when the relative advantage of one group over another on a test item is uniform, favoring only one group consistently across the entire scale of ability. Non uniform DIF (NUDIF) exists when there is interaction between ability level and group membership and the relative advantage of one group over another is not uniform across the entire ability continuum. In the case of UDIF, Penfield (2007) proposed a system for categorizing the severity of uniform DIF in polytomous items analogous to the Educational Testing Service system for characterizing dichotomous items as items with negligible,

moderate or large level of UDIF. The aim of the present study is to identify a way to assign a severity grade to DIF when NUDIF is present. In order to do this, a simulation study is arranged.

2. Simulation Study

This study takes into account one of the models for polytomous items which belongs to the Rasch family of models, that is the Rating Scale Model (RSM) (Andrich, 1978). This model theorizes that the log-odds ratio of two adjacent categories $j-1$ and j is given by the difference between the ability of person n (β_n), the difficulty of item i (δ_i) and the threshold j (τ_j). The data generating mechanism that allows to simulate data affected by NUDIF is based on the following manipulation of the RSM equation:

$$\ln\left[\frac{P_{nij}}{P_{ni(j-1)}}\right] = \beta_n - \delta_i - \tau_j + b_i * \beta_n * group_n \tag{1}$$

where the variable $group_n$ is a dummy variable coded as 1 if the subject n belongs to the focal group and 0 otherwise. The extra term $b_i * \beta_n * group_n$ emphasizes the interaction between the ability level and group membership for the item i and the coefficient b_i specifies the strength of this interaction.

In the simulation procedure a sample of 1000 abilities, which represent the target or true abilities β_n, was drawn from a standardized normal distribution and attributed at random to reference and focal group so that the two groups have almost the same size ($n_{Ref} = 492$; $n_{Foc} = 508$). The test is made by 15 items with corresponding difficulty parameters δ_i fixed equal to: [-1.7684, -1.4726, -0.8373, -0.5323, -0.3092, -0.1237, 0.9029, 1.0733, 1.3682, 1.8274, 0.7783, -0.2662, -0.9496, 0.3092, 0.0]. The set of threshold parameters τ_j is [-1, -0.5, 0, 0.5, 1], which implies six response categories. The simulation design considers increasing values of b_i, reported in Table 1, and increasing number of NUDIF items that is two, three, four and five items that correspond to a percentages of DIF items in the test equal to 13.3%, 20%, 26.7% and 33.3% respectively. In the case of two items, the DIF items are 0.0 and 0.3092. If there are three DIF items, these are 0.0, 0.3092 and -0.9496. When there are four DIF items, these are 0.0, 0.3092, -0.9496 and -0.2662. Finally, in the case of five DIF items, these are 0.0, 0.3092, -0.9496, -0.2662 and 0.7783. For each combination of b_i and number of DIF items, 200 datasets are generated and 200 sets of estimated abilities are computed.

3. Results and discussion

In order to study a possible categorization of the severity of NUDIF, the percentage of rejection, based on the 200 simulated datasets, of the null hypothesis underlying the two-sample Kolmogorov-Smirnov test (K-S test) applied to the real and estimated

Table 1. Values of the b_i coefficient

Item	0.0	0.3092	-0.9496	-0.2662	0.7783
Case 1	0.250	0.350	0.450	0.350	0.375
Case 2	0.375	0.525	0.675	0.525	0.563
Case 3	0.500	0.700	0.900	0.700	0.750
Case 4	0.625	0.875	1.125	0.875	
Case 5	0.750	1.050	1.350	1.050	
Case 6	0.875	1.225	1.575	1.225	
Case 7	1.000	1.400	1.800	1.400	
Case 8	1.125	1.575	2.025		
Case 9	1.250	1.750	2.250		
Case 10	1.500	2.100			
Case 11	1.750	2.450			

abilities and the value of the first eigenvalue of PCA on Rasch residuals for the focal group are used. The K-S test represents the tool utilized to assess the possible biasing effect of the DIF items on the abilities estimation and the rejection percentage of the null hypothesis underlying the K-S test, that is the percentage of times that the hypothesis of absence of a biasing effect of the DIF items on the ability estimation is rejected, can be considered as an index of DIF severity. Given that for NUDIF items the subjects belonging to the focal group experience a different ability, it is likely to find a second dimension in the data referred to the focal group. If this extra dimension is strong enough to show up, then the tool used to detect this dimension (first eigenvalue of PCA on Rasch residuals for the focal group) can be considered as an index of DIF severity. The evaluation of both the rejection percentage of the null hypothesis underlying the K-S test and the first eigenvalue of PCA on Rasch residuals for the focal group, joined to the value of the coefficients b_i, allows a possible categorization of the severity of NUDIF. Table 2 reports, for each combination of number of DIF items and values of b_i, the rejection percentage of the null hypothesis underlying the K-S test and the mean first eigenvalue of PCA on Rasch residuals computed from the 200 simulated datasets.

The analysis of the results reported in Table 2 shows that the absence of a second dimension in the Rasch residuals for the focal group, highlighted by the first eigenvalue of PCA about 1.4 (Brentari and Golia, 2007), is associated with a modest rejection percentage of the K-S test (about or less than 30%). Then, it is possible to classify as negligible level of NUDIF the cases of two items and average b_i less than 0.9, three items and average b_i less than 0.53 and four items and average b_i less than 0.35. When the test contains two NUDIF items and the average b_i is bigger than 0.9, the two items show a moderate to large level of NUDIF. If the questionnaire contains three NUDIF items, it is possible to classify the level of NUDIF of these items as moderate when the average b_i lies between 0.9 and 1.224 (first eigenvalue of PCA on Rasch residuals between 1.5-1.9 and rejection percentage of the K-S test between 40%-70%) and large when the aver-

Table 2. Rejection percentage of the K-S test and average first eigenvalue of PCA on Rasch residuals

	2 Items	3 Items	4 Items	5 Items
Case 1	12.4% - 1.38	19.4% - 1.38	33.0% - 1.38	45.0% - 1.37
Case 2	19.0% - 1.37	29.6% - 1.40	55.0% - 1.46	74.2% - 1.50
Case 3	23.4% - 1.39	42.0% - 1.48	77.0% - 1.60	90.0% - 1.66
Case 4	28.5% - 1.41	52.2% - 1.61	87.0% - 1.76	
Case 5	32.4% - 1.45	60.6% - 1.71	91.0% - 1.92	
Case 6	34.4% - 1.49	64.4% - 1.82	96.8% - 2.06	
Case 7	38.8% - 1.54	70.4% - 1.92	97.0% - 2.22	
Case 8	38.8% - 1.59	78.0% - 2.03		
Case 9	42.6% -1.62	80.2% - 2.11		
Case 10	46.2% - 1.72			
Case 11	51.6% - 1.79			

age b_i is bigger than 1.224 (first eigenvalue of PCA on Rasch residuals bigger than 1.9 and rejection percentage of the K-S test bigger than 70%). When the test includes four NUDIF items it is possible to classify the level of NUDIF of these items as moderate when the average b_i lies between 0.53 and 0.7 (first eigenvalue of PCA on Rasch residuals between 1.5-1.6 and rejection percentage of the K-S test between 50%-80%) and large when the average b_i is bigger than 0.7 (first eigenvalue of PCA on Rasch residuals bigger than 1.6 and rejection percentage of the K-S test bigger than 80%). When the test contains five NUDIF items, a negligible level of NUDIF is associated with an average b_i less than 0.355 (no second dimension in Rasch residuals and rejection percentage of the K-S test less than 50%). Average b_i bigger than 0.355 implies moderate to large level of NUDIF.

The estimation of b_i for NUDIF items can be made making use of the Generalized Partial Credit Model evaluated separately on subjects belonging to reference and focal group. A careful review of the questionnaire with the exclusion of some of the NUDIF items could be needed when there are items with no negligible level of NUDIF.

References

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

Brentari E., Golia S. (2007), Unidimensionality in the Rasch model: How to detect and interpret, *Statistica*, 6, 253–261.

Penfield R.D. (2007), An approach for categorizing DIF in polytomous Items, *Applied Measurement in Education*, 20, 335–355.

Quaderni di Statistica
Vol. 14, 2012

University admission test and students' careers: evidence from the School of Economics in Florence

Leonardo Grilli Carla Rampichini
Dipartimento di Statistica "G. Parenti", Università di Firenze
E-mail: grilli@ds.unifi.it, rampichini@ds.unifi.it

Roberta Varriale
ISTAT
E-mail: varriale@istat.it

Summary: In the academic year 2008/2009, the School of Economics of the University of Florence introduced a compulsory test to evaluate the background of the students wishing to enrol in a degree program. In this paper, we assess the predictive power of the test score in terms of number of gained credits, making comparisons with the predictive power of variables recorded in administrative data, such as the type of high school and the high school final grade. To disentangle direct and indirect effects, the result of the admission test is treated as an intermediate variable in a regression chain graph. About 20% of the enrolled student did not gain any credit at the end of the first year, thus we consider a two-part (hurdle) model, in order to deal correctly with excess zeros.

Keywords: Hurdle model, Regression chain graph, University credits.

1. Introduction

In the academic year 2008/2009, the School of Economics of the University of Florence introduced a compulsory test to evaluate the background of the students wishing to enrol in a degree program. The test is based on 40 multiple-choice items covering 3 areas: Logic (12 items, 30%), Reading (10 items, 25%) and Mathematics (18 items, 45%). For each item, one out of 5 alternatives is correct, with the following scoring system: 1 if correct, 0 if blank, -0.25 if wrong. Thus the total score ranges from -10 to 40, and the threshold for passing the test is fixed at 9: candidates with a lower total score are advised against enrollment.

We consider the participants to the first edition of the test (September 2008). The data set is obtained by merging data collected at the test with the administrative data of the School of Economics. After deleting 68 foreign students (due to missing information), the data set has 1057 observations. The available students' variables are listed in the following. *Pre-test*: Female, Far-away resident (indicator for residence in the provinces of Massa-Carrara and Grosseto or in a province out of Tuscany), Type of high school (Scientific, Humanities, Technical, Other), High school irregular career (indicator for age at high school diploma > 19), High school grade (from 60 to 100, centered at 80). *Test*: Total test score, Partial test scores (Logic, Reading, Mathematics), Test passed (indicator for total test score ≥ 9). *University career*: Delay in enrollment (indicator for being enrolled one or more years after high school diploma), Degree program (Management, Economics, Tourism, Marketing and Statistics), Credits gained during the first year (from 0 to 60), Second year enrollment at the School of Economics.

The test was passed by 853 candidates (80.7%). The test result is not mandatory for enrollment, but it influences the probability of enrollment: the enrollment rates were 65.3% overall, 67.9% for candidates who passed the test and 54.4% for candidates who did not pass the test.

The analysis is based on 690 students who took the test and then enrolled at the School of Economics. After one year (December 2009), the number of gained credits was quite low: only 77.0% gained credits, with a mean of 29.8 (out of 60). The result is even worse for the subset of students who did not pass the test: only 58.6% gained credits, with a mean of 23.5. The association between gained credits and dropout is strong: the dropout rate is 77% among students without credits, as compared to 11% among students that gained credits. Therefore, the number of credits gained after one year is a very informative outcome variable as it is related to both dropout and speed of progression.

2. Methods

The analysis aims at evaluating if the university admission test is a good predictor of students' careers. To disentangle direct and indirect effects of students background characteristics on the number of gained credits, the result of the admission test is treated as an intermediate variable in a regression chain graph (Cox and Wermuth, 1996; Wermuth and Sadeghi, 2012). The specified chain graph model has tree blocks: (*i*) pre-test (exogenous) variables, (*ii*) standardized test scores, and (*iii*) gained credits after one year (outcome). The test result could be summarized by the total score, but this would obscure some interesting aspects of the phenomenon: first of all, the three areas (Logic, Reading and Math) have different numbers of items; moreover, we wish to evaluate the relationships of each of the three partial scores with pre-test variables and the outcome. Therefore, we consider the Logic, Reading and Math scores as distinct variables, using standardized values to eliminate the effect of the different numbers of items. In

principle, the three standardized partial scores should be jointly regressed on pre-test covariates with a multivariate model. However, the same point estimates, and quite the same standard errors, are obtained by regressing each partial score on pre-test variables.

The model for the gained credits is complicated by excess zeros (23% of freshmen did not gain any credit). Therefore, we use a two-part or hurdle model (Cameron and Trivedi, 2005). Let Y_i denote the number of gained credits after one year for the *i*-th student, taking integer values in the set $\{0, 1, 2, \ldots, 60\}$. The model for the credits has two components: a logit model for the probability of gaining at least one credit $P(Y_i > 0 \mid \mathbf{z}_i)$, and a linear model for the expected number of gained credits $E(Y_i \mid Y_i > 0, \mathbf{x}_i)$. The linear model is fitted on the subset of students who gained at least one credit. The covariates of the two sub-models, $\mathbf{z}_i$ and $\mathbf{x}_i$, are distinct in principle, but they can even be the same.

Exams have different credits, usually 6, 9 or 12, thus the distribution of positive credits is quite irregular: the main peaks are at 6, 15, 24, 36 and 45 credits, depending on the path followed by the student. The median value is 30 credits, while only 0.75% of the students gained the maximum value 60. No parametric distribution appropriately describes this pattern: therefore, in order to model the gained credits without imposing distributional assumptions, we estimate the parameters via OLS and compute robust standard errors.

3. Results

The estimated parameters of the regression chain graph model are reported in Table 1. The pre-test variables explain only a small part of the variability of the standardized test scores, with R^2 ranging from 0.129 for Logic to 0.312 for Math. On average, females have worse results in all three areas, while candidates with a higher grade have better results. The type of high school has a significant effect: compared to candidates from a high school focusing on science, candidates from humanities have a worse performance in Math and a better performance in Reading, whereas candidates from technical and other high schools have a worse performance in all three areas. Candidates with an irregular career, as well as candidates residing far away, have worse results in Math.

The last two columns of Table 1 refer to the estimated two-part model for gained credits. Even controlling for pre-test covariates, the standardized partial test scores have a significant effect on credits: students with a higher score on Reading have a higher probability of gaining credits, $P(Y > 0)$, whereas students with a higher score on Math on average gain a higher number of credits during the first year, $E(Y|Y > 0)$. Thus, students with difficulties in Reading have lower chances to start-up their university career, while students with difficulties in Math tend to proceed slowly in the first year, likely for problems encountered in Math and Statistics (which are often the hardest exams). The score on Logic does not contribute to predict the acquisition of credits once the scores on Reading and Math are known.

Table 1. Chain graph model estimates

	Standardized test scores			*Gained credits (Y)*	
Variable	*Logic*	*Reading*	*Math*	$P(Y>0)$	$E(Y\|Y>0)$
Constant	0.500***	0.432***	0.864***	1.726***	31.128***
Female	-0.230**	-0.442***	-0.328***	0.130	1.459
Far-away resident	-0.221	-0.186	-0.412**	-0.192	-0.862
HS type (ref: Scientific)					
Humanities	0.008	0.482***	-0.672***	-0.140	-0.942
Technical	-0.565***	-0.341***	-0.876***	-0.337	-1.844
Other	-0.769***	-0.331**	-1.138***	-0.325	-4.766*
HS irregular career	0.059	0.030	-0.404***	-1.011***	-5.048*
HS grade	0.017***	0.027***	0.023***	0.030**	0.427***
Std test scores					
Logic				0.121	0.122
Reading				0.390***	0.064
Mathematics				0.167	2.682***
RMSE	2.369	2.219	3.003		12.831
R^2	0.129	0.163	0.312		0.243

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

The effects of pre-test covariates are mediated by the test scores, with the notable exceptions of high school grade (positive effect) and irregular career (negative effect). Therefore, these covariates are proxies of abilities and attitudes of the students that are not captured by the admission test.

The results of our analysis show that students' careers are difficult to predict since a large part of variability in gained credits remains unexplained. Anyway, the admission test gives indications useful for student tutoring: a low Reading score is related to a difficult start-up, while a low Math score is related to a slow progression.

References

Cameron A.C., Trivedi P.K. (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press, Cambridge.

Cox D.R., Wermuth N. (1996), *Multivariate dependencies: models, analysis and interpretation*, Chapman & Hall, London.

Wermuth N., Sadeghi K. (2012), Sequences of regressions and their independences, to appear in *Test*, http://arxiv.org/abs/1103.2523.

Quaderni di Statistica
Vol. 14, 2012

CUBE models for interpreting ordered categorical data with overdispersion

Maria Iannario
Department TEOMESUS, University of Naples Federico II
E-mail: maria.iannario@unina.it

Summary: In this paper we introduce a new probability distribution for ordinal data by using a mixture of discrete random variables in order to take overdispersion into account. After discussing parsimony of the model and related inferential issues, we check on a real data set the effectiveness of the proposal for capturing feeling, uncertainty and overdispersion.

Keywords: Ordinal data, CUBE models, Beta-Binomial distribution.

1. Introduction and background

For the analysis of ordinal data, the standard approach of CUB models considers feeling and uncertainty as the main components of the response (Iannario and Piccolo, 2012). Indeed, it seems useful to consider also overdispersion in ordinal scores caused by a possible variability among the individual feeling of respondents. Generally, a different feeling among subjects is the main motivation for the inclusion of covariates in CUB models. However, it is necessary to distinguish among two sources of variability in this kind of mixture distributions.

If a covariate affects the respondents' feeling, it induces a shift in the location of the responses which may be interpreted with a dummy covariate, for instance: in such case, if we consider aggregate data, we observe a higher variability. Instead, if there is an extra variability induced by different responses, mainly located around the same feeling, we get an high dispersion in the scores. This overdispersion cannot be explained by the Binomial component since this random variable implies a strong relationship among its expectation and variance. To take this heterogeneity into account, it is possible to consider the parameter in the binomial model as a random variable drawn from a Beta distribution. The resulting compound distribution (Beta-Binomial) is the key issue for

introducing the new mixture distribution proposed in this paper. Indeed, we introduce a shift Beta-Binomial random variable in the CUB mixture in order to explicitly characterize the role of a possible overdispersion in ordinal data.

The paper is organized as follows: in the next section we specify the new model and briefly discuss the main inferential issues related to estimation, testing and validation of such model. In section 3, a case study should validate the proposal on a real data set. Some concluding remarks end the paper.

2. *Specification of* CUBE *models and inferential issues*

For a given number $m > 4$ of categories, let r denote the realization of a discrete random variable R whose probability mass function is defined by:

$$Pr\,(R = r) = \pi\,\beta_r(\xi, \delta) + (1 - \pi)\,U_r\,, \quad r = 1, 2, \ldots, m\,, \tag{1}$$

where $U_r = \frac{1}{m}$ is the discrete Uniform distribution and $\beta_r(\xi, \delta)$ is a shifted Beta-Binomial distribution specified by:

$$\beta_r(\xi, \delta) = \binom{m-1}{r-1} \frac{\prod_{k=1}^{r} [1 - \xi + \delta(k-1)] \prod_{k=1}^{m-r+1} [\xi + \delta(k-1)]}{[1 - \xi + \delta(r-1)]\ [\xi + \delta(m-r)] \prod_{k=1}^{m-1} [1 + \delta(k-1)]}\ .$$

Distribution (1) will be denoted as a CUBE model since it is a ***C***ombination of ***U***niform and shifted ***BE***ta-binomial random variables. It is characterized by the parameter vector $\boldsymbol{\theta} = (\pi, \xi, \delta)' \in \Omega(\boldsymbol{\theta})$ where the parametric space:

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

is the positive octant in R^3 bounded over the unit square.

This class of models includes CUB models as special case ($\delta = 0$) but allows also for distributions with two opposite modes at $R = 1$ and $R = m$ (if $\delta > 0.5$). In addition, it is worth to say that a substantial overdispersion may be obtained with small values of δ parameter ($\delta < 0.3$, say).

More interestingly, CUBE distribution encompasses several random variables as listed in Table 1. Specifically, it lists some nested models implied by the CUBE probability distribution.

From an inferential point of view, given a sample of ordinal data $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)'$, the log-likelihood function is expressed by:

$$\ell\,(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \pi \left[\beta_{r_i}(\xi, \delta) - \frac{1}{m} \right] + \frac{1}{m} \right\}\ . \tag{2}$$

Thus, Maximum Likelihood (ML) estimates, parameter testing and fitting measures have been obtained by similar steps as in Iannario and Piccolo (2012). Specifically, the EM algorithm has been used for the estimation procedure.

Table 1. Probability distributions implied by CUBE *models.*

Models	π	ξ	δ
CUBE	$\pi \in (0,1]$	$\xi \in [0,1]$	$\delta > 0$
CUB	$\pi \in (0,1]$	$\xi \in [0,1]$	$\delta = 0$
Shifted Binomial	$\pi = 1$	$\xi \in [0,1]$	$\delta = 0$
Inverse HyperGeometric	$\pi = 1$	$\xi = \frac{\beta}{1+\beta}$	$\delta = \frac{1}{1+\beta}$
Discrete Uniform	$\pi = 0$	ξ undefined	$\delta = 0$

Finally, for testing the significance of the overdispersion parameter we perform a likelihood ratio test by halving the p-value of a χ^2 distribution with $g = 1$ degree of freedom (because of the borderline nature of the null hypothesis). Among fitting measures, we quote $\mathcal{F}^2$, that is a normalized measure which compares observed and expected relative frequencies.

3. Empirical analysis

To validate the effectiveness of the proposal we check the possible presence of an overdispersion in the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy since 1965. Data collection is entrusted to a specialized company and the interview stage is preceded by a series of meetings at which officials from the Bank of Italy and representatives of the company give instructions directly to the interviewers.

From the inspection of the information provided by the interviewer, an interesting aspect is the analysis of the context for submitting questionnaires. At the end of each interview, the interviewer answers to several topics by synthesizing his opinions about the quality of the responses. Thus, we examine the answer to "Global comprehension of questions" (ranging from 1=*low comprehension* to 10=*high comprehension*) by observing the differences between the 2006 and 2008 waves.

Generally, we observe a high level of perceived comprehension with a low level of uncertainty.

Table 2. Estimated models for Respondent Interview Comprehension.

Waves	$\hat{\pi}_{\text{CUB}}$	$\hat{\pi}_{\text{CUBE}}$	$\hat{\xi}_{\text{CUB}}$	$\hat{\xi}_{\text{CUBE}}$	$\hat{\delta}$	LRT	ΔBIC	$\Delta\mathcal{F}^2$
2006	0.911	0.905	0.186	0.227	0.065	222.470	-214.22	0.0603
2008	0.905	0.964	0.227	0.237	0.062	216.955	-208.68	0.0596

Table (2) lists the main estimation results for CUB and CUBE models, respectively: we also report the LRT of CUBE versus CUB model and the measures ΔBIC and $\Delta\mathcal{F}^2$

(defined as the difference of BIC and $\mathcal{F}^2$ indices of the two models, respectively).

All estimated parameters are highly significant and the fitting is acceptable ($\mathcal{F}^2$ range from 0.8875 to 0.8969). A sensible overdispersion is detected for both waves (BIC reduction is noticeable). For a global picture of the modification induced by the extra parameter, we plot over the unit square all estimated models in Figure 1 by enlarging the points in proportion to the estimated overdispersion parameter δ.

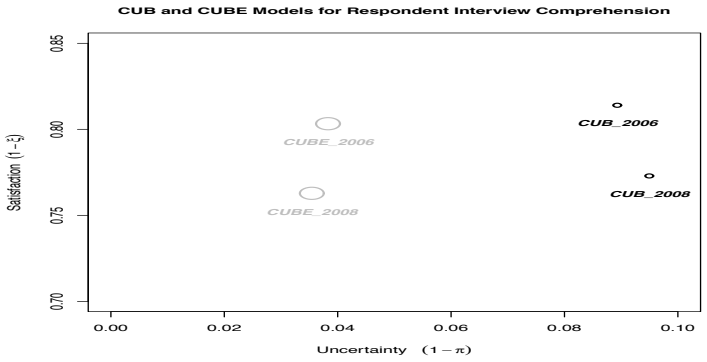


Figure 1. CUB *and* CUBE *models for Respondent Interview Comprehension.*

The role of δ is clearly depicted in the plot since this extra parameter lowers uncertainty in both waves and, moderately, the assessed perception of the feeling parameters.

The introduction of CUBE models enriches the class of CUB probability structures and the introduction of an extra parameters generates distributions which are closer to observed data both from fitting and interpretative point of view. Further improvements may be introduced if we will consider CUBE model with covariates, hierarchical contexts and further generalizations.

Acknowledgements: This work has been realized with the partial support of the PRIN2008 project: "Modelli per variabili latenti basati su dati ordinali" (CUP n.E61J10000020001) at University of Naples Federico II.

References

Iannario M., Piccolo D. (2012), CUB models: Statistical methods and empirical evidence, in: Kenett R. S. and Salini S. (eds.), *Modern Analysis of Customer Surveys: with applications using R*, J. Wiley & Sons, Chichester, 231–258.

Quaderni di Statistica
Vol. 14, 2012

Parameter estimation for a copy number variation detection based on log R ratio and B allele frequency using latent variable methods

Gun Ho Jang
Department of Biostatistics and Epidemiology, University of Pennsylvania
E-mail: gunjang@upenn.edu

Summary: High-density single nucleotide polymorphisms (SNP) array made it possible to achieve more accurate copy number variation (CNV) detection. Methods are developed for CNV calling using prespecified parameter values. While a few parameter estimation methods are available. In this paper, we proposed a method to estimate parameter values, which takes into account inter-marker correlation parsimoniously using hidden Markov model. A computationally efficient EM algorithm is developed. Finally, we demonstrated accuracy of our method in detecting various CNV regions.

Keywords: Mixture model, Latent variable, Copy number variation.

1. Introduction

Copy number variation (CNV) is a widespread characteristic of the human genome related to various diseases. There are some methods for CNV detection using raw signal intensity produced by SNP-array. The high variability of raw signal intensity made it difficult to interpret. The log R ratio (LRR) and B allele frequency (BAF) are transformations of raw signal intensities which are introduced in Peiffer et al. (2006) to have intuitive and straightforward interpretation. Some models are proposed on LRR and BAF like PennCNV (Wang et al., 2007) using prespecified parameters. Although PennCNV has a parameter estimation method, it discourages using it. Wang et al. (2009) tried to estimate parameters for an extended model. But Wang et al (2009)'s method fails for noisy data. Simple Gaussian distributions are considered for LRR which are not realistic for the real data, particularly in tail area. Similar phenomena happen for mixture models. For example, normal mixture models can be fitted for the data having t-distribution. Mostly three component mixture normal model fits well: one component

centered at zero and the other two are symmetric around zero. While two symmetric components have more chance to be considered as copy number variation. Since noisy data produce fat-tail, we consider student's t-distributions which are known to be robust in parameter estimation rather than normal distributions. Models and estimation methods are described in Section 2, and simulation results are shown in Section 3.

2. Model Specification and Parameter Estimation

Allele-specific copy numbers are denoted by $c = (c_A, c_B)$, that is, copy number polymorphisms (CNP) at a bi-allelic marker locus. The sum $\text{CN}(c) = c_A + c_B$ is the copy number at the maker locus. The normal states have copy number 2, i.e., $(2,0), (1,1)$ and $(0,2)$. If the copy number is less than (or greater than) 2 at a locus, there are copy number deletion (or duplication) at the locus. Because the duplications with four or more copies are rare and virtually indistinguishable (Wang et al., 2007), the maximum copy number is set to 4. Hence only 15 CNPs are considered, say $\mathcal{C}$ the set of 15 CNPs.

The raw signal intensities X_A and X_B for allele A and B have big variability locus-by-locus which hinders interpretation of each value. As a normalization, we consider two measures log R ratio and B allele frequency, denoted as r and b. For the derivation, the total signal intensity $R = X_A + X_B$ and the relative allelic intensity ratio $\theta = \arctan(X_B/X_A)/(\pi/2)$ are adjusted by "cannoical genotype clusters" using reference values as in Peiffer et al. (2006), that is, $r = \log_2(R/R_{\text{ref}})$ and b is a linear stretch of θ to have values $0, 1/2, 1$ at $\theta_{\text{AA}}, \theta_{\text{AB}}, \theta_{\text{BB}}$, respectively. Then, b is truncated on $[0,1]$ because $b < 0$ or $b > 1$ has no more information than $b = 0$ or $b = 1$.

2.1. Model Specification

Given each CNP, we assume the conditional probability of r_k given copy number state c_k at locus k is the student's t distribution with degree of freedom ν, mean $\mu_{\text{CN}(c_k)}$ and variance $\sigma^2_{\text{CN}(c_k)}$, that is,

$$r_k \mid c_k = (c_{k,A}, c_{k,B}) \sim t(\nu, \mu_{\text{CN}(c_k)}, \sigma^2_{\text{CN}(c_k)}). \tag{1}$$

Due to the truncation and linear transformation of θ used for the b calculation, we model b conditional on each CNP, that is, a mixture distribution of δ_0 (a degenerate distribution at 0 and truncated normal $N(0, \sigma_{b,\text{CN}(c_k)})$ on $[0,1]$ when $c_{k,B} = 0$, symmetrically, a mixture of δ_1 and truncated normal $N(1, \sigma_{b,\text{CN}(c_k)})$ on $[0,1]$ for $c_{k,A} = 0$, and otherwise truncated normal $N(c_{k,B}/(c_{k,A} + c_{k,B}), \sigma_{b,\text{CN}(c_k)})$ on $[0,1]$ where $0/(0+0)$ is treated as $1/2$ when $\text{CN}(c_k) = 0$.

The latent variables $\mathbf{c}$ have the inter-marker correlation. Although there exists short or long range correlation, we only consider the correlation between adjacent markers. A homogeneous Markov model depending on inter-marker distance d_k (the distance

between $(k-1)$-th and k-th markers) is considered, that is,

$$P(c_k \mid c_{k-1}) = t(\text{CN}(c_{k-1}), \text{CN}(c_k), d_k)\; G(c_{k-1}, c_k) \tag{2}$$

where $t(i,j,d)$ is a transition among copy numbers given by $t(i,j,d) = 1-(1-\gamma_{ij})(1-e^{-\eta d})$ if $i=j$ and $t(i,j,d) = \gamma_{ij}(1-e^{-\eta d})$ otherwise, and $G(\cdot,\cdot)$ is the transition between the allele-specific copy number states given copy numbers.

2.2. Parameter Estimation

We used the following expectation maximization (EM) algorithm to maximize the marginal likelihood given by, for given data $\mathbf{b} = (b_1, \ldots, b_M)$ and $\mathbf{r} = (r_1, \ldots, r_M)$,

$$P(\mathbf{b}, \mathbf{r}) = \{\prod P(r_k \mid c_k) P(b_k \mid c_k)\} \times P(c_1, \ldots, c_M). \tag{3}$$

The E-step is simply the computation of $Q(\mathcal{E} \mid \mathcal{E}^\ell) = \mathbb{E}[\log P(\mathbf{b}, \mathbf{r}, \mathbf{c} \mid \mathcal{E}) \mid \mathbf{b}, \mathbf{r}, \mathcal{E}^\ell]$ given the observed data and the parameter estimates $\mathcal{E}^\ell$ at the current step ℓ, that is,

$$\begin{aligned} Q(\mathcal{E} \mid \mathcal{E}^\ell) = &\sum_{s \in \mathcal{C}} P(c_1 = s)\rho_1(s) + \sum_{k \geq 2} \sum_{s,t \in \mathcal{C}} \xi_k(s,t) \log P(c_k = t \mid c_k = s, \mathcal{E}) \\ &+ \sum_k \sum_{s \in \mathcal{C}} \rho_k(s)[\log P(b_k \mid c_k = s, \mathcal{E}) + \log P(r_k \mid c_k = s, \mathcal{E})] \end{aligned} \tag{4}$$

where $\rho_k(s) = P(c_k = s \mid \mathbf{b}, \mathbf{r}, \mathcal{E}^\ell)$ and $\xi_k(s,t) = P(c_k = t \mid c_{k-1} = s, \mathbf{b}, \mathbf{r}, \mathcal{E}^\ell)$ which can be easily computed using forward-backward algorithm (Baum et al., 1970).

In M-step, parameters are estimated part by part since all three parts (BAF, LRR, CNP) are separated in (4). Obviously each part of (4) is weighted log-likelihood and any numerical algorithms can be hired for maximization. MM algorithms, however, are used for LRR and CNP parts in order to achieve stability and to reduce computing memory usage and time. For example, μ_j parameter in LRR part is estimated by minimizing a majorization of concerned terms in $-Q(\mathcal{E} \mid \mathcal{E}^\ell)$, i.e., $\sum_k \tilde{\rho}_k(j)(b_k - \mu_j)^2/[\sigma_j^2 \nu] \geq \sum_k \tilde{\rho}_k(j) \log(1 + (b_k - \mu_j)^2/[\sigma_j^2 \nu])$ where $\tilde{\rho}_k(j) = \sum_{s \in \mathcal{C}} \rho_k(s) I(\text{CN}(s) = j)$.

Finally, the latent variables $\mathbf{c}$ is inferred to have the maximum likelihood using Viterbi's algorithm (Viterbi, 1967).

3. Simulations

For the performance evaluation, we simulated several sets of the LRR and BAF for 1000 unrelated individuals following three steps: First, 2000 independent chromosomes were generated containing 55,860 SNP loci using inter-marker correlation on chromosome 4 in the HapMap CEU population (The International HapMap Consortium, 2005).

Secondly, 10 single deletions and 1 single duplication regions with fixed lengths of 5, and 10 loci (a total of 20 and 2 for two CNV groups) were randomly chosen along the chromosome so that all these regions are disjoint. Lastly, the LRR and BAF values were generated according to the distribution discussed in Section 2.

Table 1 presents that our method has the best accuracy which are bigger than 99% in general. For the length of estimated CNV region, our method outperforms than the others in general.

Table 1. Correct call rate and Mean (s.d.) Length of Estimated CNV Regions

CNV	True	Methods		
Length	CN	Our proposal	Wang et al. (2007)	Wang et al. (2009)
$L = 5$	1	99.7% 4.99 (0.96)	70.5% 5.26 (1.53)	99.6% 5.25 (1.32)
	3	78.6% 4.89 (0.86)	47.5% 5.31 (1.19)	71.3% 4.86 (1.14)
$L = 10$	1	100.0% 9.99 (1.04)	99.8% 10.07 (1.46)	100.0% 10.26 (1.32)
	3	99.2% 9.67 (1.40)	93.5% 9.97 (1.69)	95.0% 9.63 (1.77)

References

Baum L. E., Petrie T., Soules G. and Weiss N. (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Ann. Math. Statist.*, 41, 164–171.

Peiffer D. A., Le J. M., Steemers F. J., Chang W., Jenniges T., Garcia F., Haden K., Li J., Shaw C. A., Belmont J., S S. C., R R. S., Barker D., and Gunderson K. (2006), High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping, *Genome Res.*, 16, 1136–1148.

The International HapMap Consortium (2005), A Haplotype Map of the Human Genome, *Nature*, 437, 1299-1320.

Wang H., Veldink J. H., Blauw H., van den Berg L. H., Ophoff R. A., and Sabatti C. (2009), Markov models for inferring copy number variations from genotype data on Illumina platforms, *Hum. Hered.*, 68, 1–22.

Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S. F., Hakonarson H., and Bucan M. (2007), PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data, *Genome Res.*, 17, 1665–1674.

Viterbi A. J. (1967), Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Trans. Inf. Theory*, 13, 260–269.

Quaderni di Statistica
Vol. 14, 2012

A two-step procedure for neural network modeling

Michele La Rocca Cira Perna
Department of Economics and Statistics, University of Salerno
E-mail: larocca@unisa.it, perna@unisa.it

Summary: In this paper a novel procedure for model selection in neural network modeling is discussed. Identification of hidden layer size is based on a predictive accuracy criterion, in order to avoid overfitting while identification of the number and type of input nodes is based on a multiple testing scheme, in order to avoid data snooping.

Keywords: Neural Networks, Multiple Testing, Subsampling.

1. Introduction

In this paper a two-step procedure for model selection in neural network modeling is presented and discussed. The basic idea of the novel approach is that hidden neurons and input neurons play a different role in neural network modeling and they should be selected by using different criteria. Particularly, input neurons are related to the explanatory variables, and so they have a very clear interpretation in terms of relevance of a given set of variables, while the hidden layer size has no apparent interpretation, but it models the nonlinearity degree of the functional relationship. The selection strategy acts as follows. The hidden layer size is considered as a smoothing parameter which takes into account the trade-off between bias and variability and, as a consequence, it is selected to maximize the predictive accuracy of the network, avoiding overfitting, On the contrary, the number and the type of input neurons are selected by means of a formal test procedure, based on appropriate measures of relevance of a given input variable to the model. In this latter step, in order to avoid the data snooping problem, familywise error rate is controlled by using a multiple testing scheme (Romano and Wolf, 2007). Clearly, analytical derivation of the sampling distribution of the test statistic involved is difficult but it can be approximated by using the subsampling, which is able to deliver consistent results under very weak assumptions (Politis *et. al.*, 2001).

The paper is organized as follows. In section 2 neural network modeling is briefly reviewed and the model selection procedure is discussed while in section 3 some results on simulated data are reported.

2. The model selection procedure

Let $\left\{\mathbf{Z}_i = \left(Y_i, \mathbf{X}_i^T\right)^T\right\}$ be *iid* random vectors of order $(d+1)$. It is usually of interest the relationship between Y_i and $\mathbf{X}_i$. If $\mathbf{E}\left(Y_i\right) < \infty$, then $\mathbf{E}\left(Y_i \,|\mathbf{X}_i\,\right) = g\left(\mathbf{X}_i\right)$ and we can write

$$Y_i = g\left(\mathbf{X}_i\right) + \varepsilon_i \tag{1}$$

where $\varepsilon_i \equiv Y_i - g\left(\mathbf{X}_i\right)$ and g is a function satisfying general regularity conditions. Clearly, by construction, the error term ε_i is such that $\mathbf{E}\left(\varepsilon_i \,|\mathbf{X}_i\,\right) = 0$.

The function g can be approximated by using the output of the network:

$$f\left(\mathbf{x}, \mathbf{w}\right) = w_{00} + \sum_{j=1}^{r} w_{0j}\psi\left(\tilde{\mathbf{x}}^T\mathbf{w}_{1j}\right) \tag{2}$$

where $\mathbf{w} \equiv \left(w_{00}, w_{01}, \ldots w_{0r}, \mathbf{w}_{11}^T, \ldots, \mathbf{w}_{1r}^T\right)^T$ is a vector of size $r(d+2)+1$ of network weights, $\mathbf{w} \in \mathbf{W}$ with $\mathbf{W}$ compact subset of $\mathcal{R}^{r(d+2)+1}$, and $\tilde{\mathbf{x}} \equiv \left(1, \mathbf{x}^T\right)^T$ is the input vector augmented by a bias component 1. The network (2) has d input neurons, r neurons in the hidden layer and identity function for the output layer. The (fixed) hidden unit activation function ψ is a sigmoidal function.

Given a sample of size n and a value λ for the weight-decay, the estimated parameter vector is given by:

$$\hat{\mathbf{w}}_n = \underset{\mathbf{w}\in\mathbf{W}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left[Y_i - f\left(\mathbf{X}_i, \mathbf{w}\right)\right]^2 + \lambda \sum_{j=1}^{r(d+2)+1} w_j^2 \right\}$$

In the single hidden feedforward neural network class, model selection involves the choice of the number and type of input neurons and the number of hidden neurons.

The hidden layer size is treated as a smoothing parameter which takes into account the degree of nonlinearity of the function $g(\cdot)$. In order to fix r, K-fold cross-validation is an effective tool. Once the dataset has been split into K groups of equal-sized parts, the hidden layer size and the weight decay are selected as those values which minimize

$$CV(\hat{\mathbf{w}}_n, \lambda) = \frac{1}{n}\sum_{i=1}^{n} L\left(Y_i, f(\mathbf{X}_i, \hat{\mathbf{w}}_{-\tau(i)})\right)$$

where L is a proper chosen loss function and $\hat{\mathbf{w}}_{-\tau(i)}$ are network weights estimated with the $\tau(i)$ part of the data removed, with $\tau : \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$ being an indexing function denoting the partition of the observation i.

Once the hidden layer size is fixed (along with the weight-decay value), a proper set of input variables can be selected. Following White and Racine (2001) and La Rocca and Perna (2005, 2009), a measure of relevance of variable X_j to the model is given by $\theta_j = \mathbf{E}\left[f_j^2\left(\mathbf{x}, \mathbf{w}_0\right)\right]$, where f_j is the partial derivative of neural network function with

respect to the variable X_j and $\mathbf{w}_0$ is the true parameter vector. Therefore, the hypothesis that a given set of variables has no effect on Y can be formulated as a multiple test:

$$H_j : \theta_j = 0 \quad vs \quad H_j' : \theta_j > 0, \quad j = 1, 2, \ldots, d.$$

Each null H_j can be tested by using the statistic,

$$\hat{T}_{n,j} = n\hat{\theta}_{n,j} = \sum_{i=1}^{n} f_j^2 \left(\mathbf{X}_i, \hat{\mathbf{w}}_n\right)$$

where $\hat{\theta}_{n,j}$ is the sample counterpart of θ_j. Clearly, large values of the test statistics indicate evidence against the null H_j. The problem here is how to decide which hypotheses to reject, accounting for the multitude of tests. In order to control the familywise error rate, defined as the probability of rejecting at least one of the true null hypotheses, we use a multistep proposal by Romano and Wolf (2007), suitable for joint comparison of multiple misspecified models. The algorithm runs as follows:
1. Relabel the hypothesis from H_{r_1} to H_{r_d} in redescending order with respect to the value of the test statistics $\hat{T}_{n,j}$, that is $\hat{T}_{n,r_1} \geq \hat{T}_{n,r_2} \geq \ldots \geq \hat{T}_{n,r_d}$.
2. Set $j = 1$ and $R_0 = 0$.
3. For $R_{j-1} + 1 \leq s \leq S$, if $0 \notin [T_{n,r_s} - c_j, \infty)$, reject the null H_{r_s}.
4.a If no further null hypotheses are rejected, stop.
4.b Else, let R_j be the number of hypotheses rejected, let $j = j + 1$ and go to step 3.

The quantiles $c_1, c_2, \ldots$ are estimated by using the subsampling. This choice can be justified for several reasons: (*i*) the method does not require any knowledge of the specific structure of the data and so it is robust against misspecifications, a key property when dealing with artificial neural network models; (*ii*) the procedure delivers consistent results under very weak assumptions; (*iii*) the subsampling algorithm does not change dramatically when moving from *iid* to dependent data.

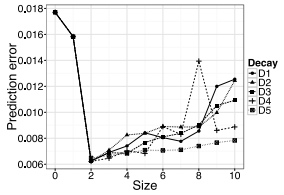
3. Numerical examples

Simulated datasets are generated from two models:

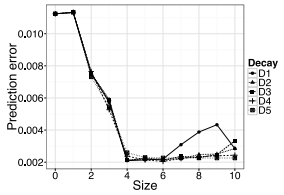
M1. $Y = 3\psi\left(2X_3 + 4X_4 + 3X_5 + 3X_6\right) + 3\psi\left(2X_3 + 4X_4 - 3X_5 - 3X_6\right) + \varepsilon$

M2. $Y = \left(10\sin\left(\pi X_3 X_4\right) + 20\left(X_5 - 0.5\right)^2 + 10X_6 + 5X_7 + \varepsilon\right)/25$

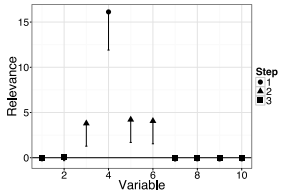
In both cases, we assume that Y depends on 10 explicative variables, but just variables $\{X_3, X_4, X_5, X_6\}$ are relevant to the Model 1 and $\{X_3, X_4, X_5, X_6, X_7\}$ are relevant to Model 2. In Model 1, $\varepsilon \sim N(0, 0.7)$ and ψ is the logistic activation function, $\mathbf{X} = (X_3, X_4, X_5, X_6)^T$ is a vector of multivariate Gaussian random variables with zero mean, unit variance and pairwise correlation equal to 0.5. In Model 2, $\varepsilon \sim N(0, 1)$ and $\mathbf{X} = (X_3, X_4, X_5, X_6, X_7)^T$ is drawn randomly from the unit hypercube.



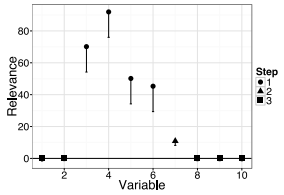
(a) Hidden layer size, Model 1



(b) Hidden layer size, Model 2



(c) Variable selection, Model 1



(d) Variable selection, Model 2

Figure 1. Hidden layer size choice and variable selection for datasets of size $n = 300$ *from Model 1 and 2. Decay values: D1=0, D2=5E-6, D3=5E-5, D4=5E-4, D5=5E-3. Size of the test:* $\alpha = 0.01$. *Quantiles estimated by* 1999 *subsamples of size* 140.

From the examples reported in Figure 1, multiple step procedures appear clearly necessary since in both cases, in the first step, some variables were incorrectly classified as not relevant to the model.

References

La Rocca M., Perna C. (2005), Variable selection in neural network regression models with dependent data: a subsampling approach, *Computational Statistics & Data Analysis*, 48 (2), 415–429.

La Rocca M., Perna C. (2009), Neural Network Modelling with Applications to Euro Exchange Rates, in: Kontoghiorghes E. J., Rustem B., Winker P. (eds.), *Computational Methods in Financial Engineering*, 163-189.

Politis D.N., Romano J. P., Wolf M. (2001), On the asymptotic theory of subsampling, *Statistica Sinica*, 11(4), 1105-1124.

Romano J.P., Wolf M. (2007), Control of generalized error rates in multiple testing, *Annals of Statistics*, 35(4), 1378-1408.

White H., Racine J. (2001), Statistical inference, the bootstrap, and neural-network modeling with application to foreign exchange rates, *IEEE Transactions on Neural Networks*, 12(4), 657-73.

Quaderni di Statistica
Vol. 14, 2012

Record Linkage between Large Datasets: Evidence from the 15th Italian Population Census

Luca Mancini, Luca Valentino, Francesco Borrelli, Luigi Marcone
Italian National Institute of Statistics
E-mail:lmancini@istat.it, luvalent@istat.it, borrelli@istat.it, lmarcone@istat.it

Summary: Record linkage with imperfect key identifiers has serious dimensionality constraints when records belong to large datasets. Aggressive search space reduction strategies are needed in order to limit the number of comparisons. However, achieving a lower space and time complexity usually comes at the expense of accuracy, particularly when the blocking or sorting attributes are affected by error. The purpose of the paper is to identify suitable algorithms which help minimize this trade–off. The simulations are run on real data from the latest Italian population census. The results show that the *simhashing* techniques typically used for near-duplicate detection of web documents offer a promising alternative to more traditional *blocking* or *sorting*.

Keywords: Record linkage, Census, Simhash.

1. Introduction

Probabilistic record linkage (*RL*) assigns to a given pair of records a probability of identifying the same statistical unit (Jaro, 1989). When the size of the data sources to be linked is large a subset of pairs has to be selected before matching probabilities can be estimated. Suitable space reduction strategies like *blocking* and *sorted neighbour methods* (*SNM*) are typically used to reduce dimensionality. In the case of *SNM*, for instance, records are first sorted by the same key and then scanned using a fixed window of size w so that only those pairs falling inside w are compared.[1] The success of search algorithms like *SNM* crucially depends on the quality of the sorting key(s) of choice. If the sorting is done on imperfect keys a number of potential matching pairs will go undetected because one of the records will fall outside the scanning window. The purpose of the paper is to identify strategies which help reduce the number of pairwise comparisons

[1] Assuming two data sets of size n_1 and n_2 records respectively, the number of comparisons is of order $\mathrm{O}(wn)$ i.e. exponentially lower than the cross product $\mathrm{O}(n^2)$ between all records.

on one hand and are not too sensitive to linkage key errors, on the other.[2] We evaluate a *hash*–based algorithm known as *simhash* (*SH*) (Charikar, 2002). Although this technique has been successfully applied to near–duplicate detection of web crawls (Manku et al., 2007; Sood and Loguinov, 2011) and short text messages (Pi et al., 2009) we are not aware of applications to *RL* tasks when the entities of interest are individuals. *SH* is evaluated in relation to *SNM* against two sets of performance metrics: a) the search space reduction ratio and b) the accuracy of the estimated links.

2. The model

Let $B = 1, 2, \ldots, n$ be the set of all unique and contiguous two-character substrings (or *bigrams*) of a chosen alphanumeric attribute across the entire dataset and $B(s) \subseteq B$ be the subset of *bigrams* of string s referring to a specific individual record. Each token i has a weight $w_i \subseteq \Re$ which measures its contribution to s. The combination of $B(s)$ and w_i represents the *feature vector* of string $s, W_j(s)$. Define V as the set of all feature vectors and assume a similarity measure $m : V^2 \rightarrow \{0, 1\}$ that maps pairs of vectors to real numbers between 0 and 1. Simhash is a one–way mapping of $V \rightarrow \{0, 1\}^j$ that converts feature vectors to j–bit binary strings. To obtain $W_j(s)$ each bigram i of $B(s)$ is hashed into a j–bit uniformly random binary *hash* ϕ_i, such that

$$W_j(s) = \sum_{i \in B(s)} (2\phi_{ij} - 1)w_i = \begin{cases} w_i & \text{if} \quad \phi_{ij} = 1 \\ -w_i & \text{if} \quad \phi_{ij} = 0 \end{cases} \tag{1}$$

where ϕ_{ij} is the j^{th} bit of ϕ_i. The j–bit *SH* fingerprint ς_j of s is finally obtained as

$$\varsigma_j(s) = \begin{cases} 1 & \text{if} \quad W_j(s) > 0 \\ 0 & \text{if} \quad W_j(s) \leq 0 \end{cases} \tag{2}$$

It has been shown by Charikar (2002) that the probability of collision between two fingerprints on a given bit *j* is proportional to the *cosine distance* between the projections of the corresponding strings on a hyperplane. More formally,

$$Pr[\vec{\varsigma} = \vec{\varsigma}'] = 1 - \theta(\vec{s}, \vec{s}') \approx (cos\vartheta + 1)/2 \tag{3}$$

where $\theta \in [0, \pi]$ is the angle between the two vectors. As bits in ς are pairwise independent, string similarity can be easily measured using the *Hamming distance* (*HD*) between their corresponding ς_j.[3] Record pairs are regarded as potential links and included in the search space if their ς_j are less than k bits apart. Unlike *SNM*, the *SH*

[2] This research was inspired by dimensionality issues encountered during the *RL* between large population registers used in the 2011 Italian census.

[3] The bitwise *HD* between $\varsigma_5 = [01101]$ and $\varsigma_5{}' = [10100]$ is equal to 3.

scanning window w is not fixed but its size depends on the *HD* between consecutive ordered fingerprints. The search for candidate matches of record i will stop as soon as $HD[\ \varsigma_j^i, \varsigma_j^{i+w}\ ] \geq k$. It can happen that two fingerprints whose $HD < k$ end up too far apart in the sorted array. In such cases it is possible to rotate head-to-tail b–bit blocks of ς_j without altering the *HD* between the fingerprints and repeat the search on the newly sorted array. This ensures that pairs previously missed are included in the space.

3. Results

We used data on foreign residents from the *Liste anagrafiche comunali* (Municipal Population Registers) and the *Permessi di soggiorno* (Permits to Stay Archive). Each individual record name was combined into one single string, shingled into *bigrams* and hashed into 128-bit binary sequences to create ς_j. Two conventional metrics like *precision* and *recall* were used to assess the accuracy of the linkage.[4] Table 1 shows the results of three different *RL* scenarios.[5] With $k = 45$ and only one 64–bit block permutation *SH*'s recall was nearly perfect and significantly higher than *SNM*'s.[6] Moreover this result was achieved starting from a search space less than half the size of its *SNM*'s counterpart. An analysis of the genuine links missed by *SNM* but not by *SH* revealed that the latter is much less sensitive to errors such as first and last names inversions or omitted middle names. The impression is that *SNM* needs to generate a much larger space in order to approach the performance of *SH*. With the exception of the first scenario, the results also suggest that *recall* and space reduction ratio both increase when either k or b are reduced. A marginally lower k yields huge savings in terms of space with no significant loss of accuracy. On the other hand, a higher number of rotations of smaller b–blocks improves accuracy with little cost in terms of space. Record linkage between large datasets in the absence of error-free primary key identifiers is seriously constrained by dimensionality problems. The solution is to find suitable space reduction strategies which limit the number of potentially linkable pairs on one hand and guarantee a good accuracy level of the returned links, on the other hand. The paper evaluates the performance of *simhashing* techniques against traditional *sorted neighbour* methods. Although *hash*–based algorithms have been successfully used for the detection of near-duplicates in large electronic document repositories, we are not aware of applications to *RL* tasks when the entities of interest are individuals. The performance of *SH* was tested

[4] *Precision* and *recall* measure the proportion of predicted links. Whereas the former is calculated on the overall number of returned links, the latter is defined in terms of the total number of true matches. When this is unknown a 100% benchmark recall value can be assigned to the method returning the highest number of true links.

[5] The last case is technically a de-duplication as file$_A$ is compared to itself. The same *RL* strategy in terms of variable selection, metrics and threshold settings was used for all methods. Only in the first *RL* was the file size small enough to create the $n_a \times n_b$ search space.

[6] The choice of k in the paper was heuristic, i.e. the parameter was finally set after observing the *HD* between "marginal" links.

Table 1. Results

file$_A$ (n)	file$_B$ (n)	Method	Space (n)	true links (n)	precision (%)	recall (%)
2,440	1,955	$n_a \times n_b$	$\approx$ 4.7 ML	536	99.6	100
2,440	1,955	SNM (10)	18,970	503	99.8	93.8
2,440	1,955	SH (45/64)	7,446	532	99.6	99.2
2,440	1,955	SH (40/64)	2,917	520	99.6	97.1
2,440	1,955	SH (40/16)	5,412	523	99.6	97.6
34,416	41,122	SH (40/16)	245,318	9,512	99.8	100
34,416	41,122	SH (45/64)	473,659	9,456	99.9	99.4
34,416	41,122	SH (40/64)	182,089	9,379	100	98.6
34,416	41,122	SNM (50)	$\approx$ 1.6 ML	9,323	99.7	98.0
70,314	–	SH (40/16)	$\approx$ 1.7 ML	8,251	98.1	100
70,314	–	SH (40/64)	$\approx$ 1.5 ML	8,208	98.4	99.5
70,314	–	SNM (15)	984,921	7,943	98.0	96.2

Notes: *SNM*(w), *SH*(k/b).*Precision* and *recall* were estimated using *Relais 2.3*.

in a context of linkage keys affected by errors using data from population registers. The findings are encouraging and show that *SH* clearly outperforms traditional fixed-window *SNM* both in terms of space reduction ratio and accuracy. Further research is needed to find more universal criteria for optimal parameter setting such as fingerprint length, *HD* inclusion thresholds and bit permutation strategies.

References

Charikar M.S. (2002), Similarity estimation techniques from rounding algorithms, *Proceedings ACM STOC '02*, May 19-21 2002, Montreal, Quebec, Canada, 380–388.

Jaro, M.A. (1989), Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. In *Journal of the American Statistical Society*, 64(406), 414–420.

Manku G.S., Jain A., Sarma A.D. (2007), Detecting Near-Duplicates for Web Crawling, *Proceedings WWW '07*, May 8-12, 2007, Banff, Canada, 141–149.

Pi B., Fu S., Wang W., Han S. (2009), Simhash-based Effective and Efficient Detecting of Near-Duplicate Short Messages, *Proceedings ISCSCT '09*, 26-28 December, 2009, Huangshan, China, 20–25.

Sood, S., Loguinov D. (2011), Probabilistic Near-Duplicate Detection Using Simhash, *Proceedings CIKM '11*, Oct. 20-24, 2011, Glasgow, UK, 1117–1126.

Quaderni di Statistica
Vol. 14, 2012

Nonparametric testing for agreement among several judges

Marco Marozzi
Department of Economics and Statistics, University of Calabria
E-mail: mmarozzi@unical.it

Summary: The problem of whether the rankings of some objects given by a set of judges show any agreement or are more or less independent is addressed. The most familiar measure for concordance is the Kendall W coefficient. Classical tests for concordance are the Friedman and F tests. Legendre (2005) showed via simulation that the Friedman test is too conservative and less powerful than its permutation version but his study was very limited. In this paper, the study of Legendre is deeply extended. It is shown that the Friedman test is too conservative and less powerful than both the F test and the permutation test for concordance which always have a correct size and very similar power. The F test should be preferred because it is computationally much easier.

Keywords: Nonparametric Testing, Concordance, Latent Variables.

1. Introduction

Rankings of n objects are very often considered in human resources management, education, marketing, politics, finance when job applicants, new products, political parties, public services, investment alternatives are ranked by executives, head hunters, focus groups, investors or even an automated algorithm (like that considered in Marozzi 2009). An important question is whether the rankings given by a set of p judges/criteria show any agreement or are more or less independent. The most familiar measure for concordance is the Kendall W coefficient. Classical tests for concordance between several judges are the Friedman test and the F test. Legendre (2005) showed via simulation that the Friedman test is too conservative and less powerful than its permutation version which always has a correct size. Unfortunately, his study is very limited because considers only the normal distribution and does not consider the F test. In this paper, the simulation study of Legendre is deeply extended by considering also the uniform (as a

light tailed model) and the Cauchy (as a heavy tailed model) distributions and the F test, with the aim at finding out if the conclusions drawn by Legendre (2005) for the normal case and for the Friedman and the permutation test apply also to other distributions and to the F test.

2. Notations and methods

The Kendall coefficient is defined as

$$W = \frac{12\sum_{i=1}^{n}(R_i - \overline{R})^2}{p^2(n^3 - n) - pT} = \frac{12\sum_{i=1}^{n} R_i^2 - 3p^2n(n+1)^2}{p^2(n^3 - n) - pT} \tag{1}$$

where $R_i = \sum_{j=1}^{p} R_{ij}$, R_{ij} is the rank of the ith object ($i = 1, ..., n$) given by the jth judge ($j = 1, ..., p$), $\overline{R} = \frac{1}{n}\sum_{i=1}^{n} R_i$ and T is a correction factor for tied ranks $T = \sum_{k=1}^{m}(t_k^3 - t_k)$, where t_k is the number of tied ranks in each of m groups of ties. W is an estimate of the variance of the sums of ranks R_i divided by its maximum which occurs when all judges are perfectly concordant (ie the p rankings are identical), therefore $W \in [0, 1]$. The Kendall W can be applied to data collected for estimating the same general property of the objects that cannot be directly observed. In finance, think about a set of financial ratios that act as criteria for assessing profitability, efficiency, activity, liquidity of a set of firms (the objects), they are computed by financial analysts (the judges) to evaluate how sound would be an investment in equity or debt issued by a certain firm. The soundness of an investment is a latent variable.

Classical tests for concordance are the Friedman C and the F test. The null hypothesis H_0 is the independence of the rankings given by the judges. The alternative hypothesis H_1 is that at least one judge is concordant with at least one of the other judges (one sided H_1). The Friedman statistic is computed as $C = p(n-1)W$ and under H_0 is asymptotically distributed as the χ^2 distribution with $n - 1$ df. The F test is based on $F = \frac{W(p-1)}{1-W}$ that under H_0 is asymptotically distributed as the F distribution with $\nu_1 = n - 1 - 2/p$ and $\nu_2 = \nu_1(p-1)$ df. Both the C and F tests are guaranteed to have the exact level asymptotically. A test that is guaranteed to have the exact level also for small sample sizes is the permutation test provided that the object rankings are exchangeable under H_0. Since H_0 is the independence of the object rankings, exchangeability holds and the permutation of all rank vectors independently of one another is justified. Note that $S = \sum_{i=1}^{n} R_i^2$, W, C, and F statistics are permutationally equivalent and then it is not important which one is used as pivotal statistic. Let $\underline{R}_j = (R_{1j}, ..., R_{nj})$ be the vector of object ranks given by judge j. To perform the permutation P test of concordance: (i) compute $S_0 = S$ as the observed value of the test statistic; (ii) randomly permute (independently of one another) $\underline{R}_j$ for $j = 1, ..., p$ and compute $S_1 = \sum_{i=1}^{n}(R_i^*)^2$, where $R_i^* = \sum_{j=1}^{p} R_{ij}^*$ and R_{ij}^* is the permuted rank of object i given by judge j; (iii) repeat step 2 for $B - 1$ times; (iv) compute the p-value as $L = \frac{1}{B}\sum_{b=1}^{B} \#(S_b \geq S_0)$; (v) reject (accept) H_0 if $L \leq (>)\alpha$.

3. Comparison study and conclusion

Under H_0 only independent judges are generated. We consider $n = 5, 10, 20, 50, 100$ and $p = 2, 3, 4, 5, 10, 20, 25, 30$. Let $\underline{X}_j = (X_{1j}, ..., X_{nj})$ be the n x 1 vector of values for judge j. Note that $\underline{R}_j$ contains the rank of $\underline{X}_j$. p independent judges are simulated by independently drawing X_{ij} for $i = 1, ..., n$ and $j = 1, ..., p$ from the standard normal distribution. This is the N distribution setting under H_0. By drawing X_{ij} from the uniform distribution and the standard Cauchy distribution we obtain respectively the Un and Cau distribution settings. Under H_1 p_{ind} independent judges and p_{sim} partly similar judges are generated with $p = p_{ind} + p_{ind}$. We consider $n = 20$ and all (p_{ind}, p_{sim}) such that $p_{sim} = 1, ..., 5$ with $p = 5$, and such that $p_{sim} = 1, ..., 10$ with $p = 10$. Note that when $p_{sim} = 1$ H_0 is true. The values of partly similar judges are generated as $\underline{X}_j = \underline{Y} + \sigma\underline{Z}_j$, $j = 1, ..., p_{sim}$ where $\underline{Y}$ is a n x 1 vector drawn from the standard normal distribution and $\underline{Z}_j$ is a n x 1 vector of deviates drawn from the standard normal distribution. We call $\underline{Y}$ and $\underline{Z}_j$ the common and the uncommon vector respectively, because the partly similar judges have $\underline{Y}$ in common and $\underline{Z}_j$ different. $\underline{Z}_j$ elements are multiplied by $\sigma = 0.5, 1, 2$ to simulate deviates with different scale. This is the (N, N) distribution setting under H_1. By drawing $\underline{Y}$ and $\underline{Z}_j$ from the standard normal as well as the uniform and the standard Cauchy distribution we obtain other 8 distribution settings under H_1. 10000 Monte Carlo simulations and 1000 permutations were considered. The nominal significance level was set to $\alpha = 0.05$.

Table 1 displays part of the results about power. The results under H_0 are not shown, they confirm those of Legendre (2005) for the N setting and apply equally to the Un and Cau settings: both the F and the P test have size very close to 0.05 and the C test is very conservative when $p \leq 5$. Under H_1 and the (N, N) distribution setting, the F and the P test have practically the same power whereas the C test is less powerful due to its excessive conservativeness under H_0. Without surprise, as the number of judges increases, the difference in power between the C test and the other tests decreases; and as σ increases, the power of all the tests decreases. These results confirm those of Legendre (2005) about the C and the P test. The results for the (N, N) and (N, Un) settings are practically equal. Under the (N, Cau) setting the power of all tests is lower than for the (N, N) and (N, Un) settings because an uncommon vector drawn from the Cauchy gives rise to partly similar judges that are less similar than when this vector is from the normal or uniform distribution.

Comparing the (N, N) setting results with those for (Un, N) and (Cau, N), we see that the results for (N, N) and (Un, N) are practically equal. Under the (Cau, N) setting the power of all tests is greater than for the (N, N) and (Un, N) settings because a common vector drawn from the Cauchy generates partly similar judges that are more similar than when the common vector is from the normal or uniform.

The study of Legendre (2005) considered only the normal distribution and concluded that the permutation test for concordance should be preferred to the Friedman test because it has a size very close to α and is more powerful. By extending this study with

Table 1. Power of the tests with 5 judges

	$\sigma = 0.5$			$\sigma = 1$			$\sigma = 2$		
p_{sim}	C	F	P	C	F	P	C	F	P
				(N, N) setting					
1	0.032	0.049	0.049	0.032	0.048	0.048	0.034	0.053	0.052
2	0.196	0.256	0.254	0.119	0.163	0.163	0.058	0.087	0.086
3	0.796	0.849	0.846	0.436	0.516	0.515	0.137	0.184	0.182
4	0.999	1.000	1.000	0.897	0.926	0.924	0.330	0.405	0.403
5	1.000	1.000	1.000	0.996	0.997	0.998	0.624	0.689	0.686
				(N, Cau) setting					
1	0.033	0.051	0.050	0.033	0.053	0.053	0.032	0.049	0.050
2	0.092	0.130	0.125	0.054	0.079	0.079	0.038	0.060	0.058
3	0.310	0.380	0.379	0.126	0.172	0.169	0.058	0.085	0.084
4	0.735	0.792	0.789	0.291	0.365	0.363	0.091	0.133	0.133
5	0.960	0.972	0.971	0.565	0.634	0.631	0.148	0.200	0.199
				(Cau, N) setting					
1	0.031	0.048	0.047	0.031	0.050	0.051	0.032	0.051	0.051
2	0.238	0.305	0.301	0.197	0.260	0.256	0.133	0.178	0.176
3	0.889	0.922	0.921	0.784	0.840	0.838	0.557	0.637	0.635
4	1.000	1.000	1.000	0.998	0.999	0.999	0.940	0.957	0.957
5	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.997	0.997
				(Cau, Cau) setting					
1	0.031	0.049	0.050	0.032	0.049	0.048	0.029	0.046	0.046
2	0.156	0.207	0.206	0.112	0.153	0.152	0.079	0.114	0.113
3	0.620	0.692	0.689	0.425	0.502	0.500	0.243	0.306	0.305
4	0.978	0.986	0.985	0.848	0.883	0.881	0.568	0.633	0.629
5	1.000	1.000	1.000	0.983	0.988	0.988	0.836	0.869	0.869

the consideration of a light tailed as well as a heavy tailed distribution we obtain similar conclusions. The most interesting result of our study is that the F test, not considered by Legendre, and the permutation test have the same behavior and then we suggest to use the F test because it is computationally much easier than the permutation test.

References

Legendre P. (2005), Species Associations: the Kendall Coefficient of Concordance Revisited, *Journal of Agricultural, Biological, and Environmental Statistics*, 10, 226–245.

Marozzi M. (2009), A Composite Indicator Dimension Reduction Procedure with Application to University Student Satisfaction, *Statistica Neerlandica*, 63 258–268.

Quaderni di Statistica
Vol. 14, 2012

Identification of causal effects in the presence of nonignorable missing outcome values

Alessandra Mattei Fabrizia Mealli
Department of Statistics "G. Parenti", University of Florence
E-mail: mattei@ds.unifi.it, mealli@ds.unifi.it

Barbara Pacini
Department of Statistics and Mathematics, University of Pisa
E-mail: barbara.pacini@sp.unipi.it

Summary: We will study how the presence of an instrument for nonresponse can help to sharpen bounds on the average treatment effect for the whole population. We will also provide new sufficient conditions for partial identification of causal effects for specific subpopulations of units defined by their nonresponse behavior in all possible combinations of treatment and instrument values.

Keywords: Bounds, Instrumental variables, Principal stratification.

1. Introduction

Inference on causal effects may be invalidated due to the presence of missing outcome values. In this paper, we investigate the use of a binary instrument for nonresponse in the potential outcome approach to causal inference (Rubin, 1974,1978), developing a novel approach to deal with nonignorable missing outcome values without imposing any restriction on treatment effect heterogeneity.

Under the stable unit treatment value assumption (SUTVA; Rubin, 1980), if unit i in the study ($i = 1, \ldots, N$) is assigned to treatment $T_i = t$ ($t = 1$ for treatment and $t = 0$ for no treatment), we denote with $Y_i(T_i = 1) = Y_i(1)$ and $Y_i(T_i = 0) = Y_i(0)$ the two potential outcomes, either of which can be observed depending on the value taken by T. We also denote with $S_i(t)$ the post-treatment potential variable, which represents a response indicator for $Y_i(t)$: the observation of $Y_i(t)$ is missing if $S_i(t) = 0$. The observed outcomes can be defined as $S_i^{obs} = T_i S_i(1) + (1 - T_i)S_i(0)$ and $Y_i^{obs} = T_i Y_i(1) + (1 - T_i)Y_i(0)$ if $S_i^{obs} = 1$ and Y_i^{obs} is missing if $S_i^{obs} = 0$. Throughout

the paper, we will maintain the assumption that the treatment is randomly assigned: $T_i \perp S_i(0), S_i(1), Y_i(0), Y_i(1)$. To simplify the notation, we will drop the i subscript in the sequel.

2. Causal Inference in the Presence of Unintended Missing Outcomes

Suppose that $Y(t)$ is bounded within some known interval $[L_t, U_t]$, where $-\infty < L_t \leq U_t < +\infty$, $t = 0, 1$. Define $E_{t1}(Y^{obs}) = E(Y^{obs} \mid T = t, S^{obs} = 1)$, and $P_{1|t} = Pr(S^{obs} = 1 \mid T = t), t = 0, 1$. Then, under randomization of the treatment, the following bound on the average causal effect of the treatment for the whole population can be established, using results derived by Manski (2003, Proposition 1.1, page 9):

$$\begin{aligned} &\Big(E_{11}(Y^{obs})P_{1|1} + L_1(1 - P_{1|1})\Big) - \Big(E_{01}(Y^{obs})P_{1|0} + U_0(1 - P_{1|0})\Big) \\ &\qquad \leq E(Y(1) - Y(0)) \leq \\ &\Big(E_{11}(Y^{obs})P_{1|1} + U_1(1 - P_{1|1})\Big) - \Big(E_{01}(Y^{obs})P_{1|0} + L_0(1 - P_{1|0})\Big) \end{aligned} \tag{1}$$

The sampling process is completely uninformative regarding the causal effects for units who would respond under only one or none of the two treatment regimes. As an alternative, we can focus on the local average treatment effect. Using the framework of principal stratification (Frangakis and Rubin, 2002), units under study can be stratified into four latent groups, named Principal Strata, according to the joint values of the two potential response indicators $(S(0), S(1))$: $11 : S(0) = S(1) = 1$; $01 : S(0) = 0, S(1) = 1$; $10 : S(0) = 1, S(1) = 0$; and $00 : S(0) = S(1) = 0$. Let $G \in \{11, 10, 01, 00\}$ be the principal stratum membership indicator. In stratum 11, we can observe $Y(1)$ for some respondent units under treatment and $Y(0)$ for some other respondent units under control. For this type of units, the sampling process provides information on both $Y(0)$ and $Y(1)$, therefore we can relax the restriction that the support of Y is finite, and use the empirical distribution of the two potential outcomes to identify the causal estimands of interest (e.g., Zhang and Rubin, 2003).

3. Identifying Causal Effects with an Instrumental Variable for Nonrespose

In addition to treatment T, whose causal effect on Y is still our primary interest, suppose that units are exposed to an additional treatment Z which is related to nonresponse S but unrelated to the outcome Y. The assignment of two binary treatments, T and Z, implies that four potential outcomes can be defined for the primary outcome, Y, and the response indicator, S: $S(t, z), Y(t, z)$ for $t = 0, 1$ and $z = 0, 1$. Both treatments are assumed randomized, so that:

Assumption 1 $T, Z \perp \{S(t, z), Y(t, z)\}_{t,z=0,1}$

Table 1. Principal strata with a binary treatment and a binary instrument for nonresponse

G	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
$S(0,0)$	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
$S(0,1)$	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
$S(1,0)$	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
$S(1,1)$	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

In order to characterize Z as an instrument, we propose the following assumptions:

Assumption 2 $Y(0,0) = Y(0,1) \text{ and } Y(1,0) = Y(1,1).$

Assumption 3 $E(S(0,1) - S(0,0)) \neq 0, \text{ and } E(S(1,1) - S(1,0)) \neq 0.$

Then, the following proposition holds.

Proposition 1 *Suppose that $Y(t,z)$ is bounded within some known interval $[L_{tz}, U_{tz}]$, where $-\infty < L_{tz} \leq U_{tz} < +\infty$, $t = 0,1$ and $z = 0,1$. Under Assumptions 1 through 3, $L_{t0} = L_{t1} \equiv L_t$ and $U_{t0} = U_{t1} \equiv U_t$, and the following bound can be derived:*

$$\begin{aligned}
&\max\Big\{E_{111}(Y^{obs})P_{1|11} + L_1(1-P_{1|11}); E_{101}(Y^{obs})P_{1|10} + L_1(1-P_{1|10})\Big\} - \\
&\quad \min\Big\{E_{011}(Y^{obs})P_{1|01} + U_0(1-P_{1|01}); E_{001}(Y^{obs})P_{1|00} + U_0(1-P_{1|00})\Big\} \\
&\qquad \leq E(Y(T=1) - Y(T=0)) \leq \qquad (2) \\
&\min\Big\{E_{111}(Y^{obs})P_{1|11} + U_1(1-P_{1|11}); E_{101}(Y^{obs})P_{1|10} + U_1(1-P_{1|10})\Big\} - \\
&\quad \max\Big\{E_{011}(Y^{obs})P_{1|01} + L_0(1-P_{1|01}); E_{001}(Y^{obs})P_{1|00} + L_0(1-P_{1|00})\Big\}.
\end{aligned}$$

where $E_{tz1}(Y^{obs}) = E(Y^{obs} \mid T = t, Z = z, S^{obs} = 1)$ and $P_{s|t,z} = \Pr(S^{obs} = s \mid T = t, Z = z)$, $s = 0,1$, $t = 0,1$ and $z = 0,1$

We can easily show that the bound in Equation (2) is tighter than the bound in Equation (1) if the following assumption holds:

Assumption 4 $S(t,0) \leq S(t,1) \quad \forall\, t.$

We now focus on local average treatment effects. In the presence of two binary treatments, T and Z, principal strata are defined according to the 16 joint values of $S(0,0), S(0,1), S(1,0)$, and $S(1,1)$ (see Table 1).

In many studies it may be plausible to assume that subjects who are willing to respond under control, would be willing to respond also when treated. Formally,

Assumption 5 $S(0,z) \leq S(1,z) \quad \forall\, z.$

Assumptions 4 and 5 imply that only principal strata 1, 2, 4, 6, 8 and 16 are not empty, therefore the estimand of interest is the causal effect for the union of strata 6, 8 and 16. Let $E^{\leq\alpha}_{tz1}(Y^{obs})$ and $E^{\geq\alpha}_{tz1}(Y^{obs})$ be the conditional expectations of Y^{obs} in the α ($0 < \alpha < 1$) fraction of the observed respondents ($S^{obs} = 1$) assigned to $T = t$ and $Z = z$ with the smallest and largest values of the outcome variable, Y, respectively. The following proposition holds.

Proposition 2 *If Assumptions 1–5 hold, then the following bound on the average treatment effect for the union of strata 6, 8 and 16 can be derived:*

$$E^{\leq\pi_{6,8,16|111}}_{111}(Y^{obs}) - E_{011}(Y^{obs}) \leq \tag{3}$$
$$E(Y(T=1) - Y(T=0) \mid G \in \{6,8,16\}) \leq E^{\geq\pi_{6,8,16|111}}_{111}(Y^{obs}) - E_{011}(Y^{obs})$$

where $\pi_{6,8,16|111} = \Pr(G \in \{6,8,16\} \mid T = 1, Z = 1, S^{obs} = 1) = \frac{P_{1|0,1}}{P_{1|1,1}}$.

The benefit of using an instrument for nonresponse is due to the fact that more information can be extracted from the data about the causal effects of the treatment: strata 6, 8 and 16, generally include a larger proportion of units than the group of the always respondents without instrument (stratum 11, see Section 2).

References

Frangakis C.E., Rubin D.B. (2002), Principal stratification in causal inference, *Biometrics*, 58, 191–199.

Manski C.F. (2003), *Partial Identification of Probability Distributions*, Springer-Verlag.

Rubin D.B. (1974), Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psycology*, 66, 688–701.

Rubin D.B. (1978), Bayesian Inference for Causal Effects, *Annals of Statistics*, 6, 34–58.

Rubin D.B. (1980), Discussion of "Randomization Analysis of Experimental Data: the Fisher Randomization Test" by D. Basu, *Journal of the American Statistical Association*, 75, 591–593.

Zhang J., Rubin D.B. (2003), Estimation of causal effects via principal stratification when some outcomes are truncated by death, *Journal of Educational and Behavioral Statistics*, 28, 353–368.

Quaderni di Statistica
Vol. 14, 2012

A comparison of multidimensional IRT models for assessing test dimensionality

Mariagiulia Matteucci Stefania Mignani
Department of Statistical Sciences, University of Bologna
E-mail: m.matteucci@unibo.it, stefania.mignani@unibo.it

Summary: In educational and psychological measurement, the study of test dimensionality is a fundamental issue. The aim of the paper is to compare item response theory (IRT) models with different structures in the latent abilites, in the case of a test consisting of different subscales. In fact, a unidimensional model may not properly fit the data while multidimensional models, also assuming the existence of general and specific traits, may be more accurate. IRT models with a multidimensional ability structure are compared through simulation studies in order to understand their effective capability of recovering different test structures. Model estimation is conducted via Markov chain Monte Carlo (MCMC) methods, adopting a fully Bayesian approach. An application is also conducted on real data on Italian standardized student assessments.

Keywords: Item response theory, Multidimensional models, Test dimensionality.

1. Introduction

When a test consists of different subscales, it may not be appropriate to apply a model for unidimensional data. For this reason, the study of test dimensionality has become a fundamental issue in educational and psychological measurement. Too often, models such as item response theory (IRT) models are applied under the assumption of unidimensionality, *i.e.* the presence of a single or at least one predominant latent ability, while more complex structures could be underlying the response process. On the other hand, IRT models with a multidimensional ability structure (see *e.g.* Reckase, 2009; Sheng and Wikle, 2007; 2008) could be able to describe the manifest item responses with a increased degree of accuracy. One may argue that separate unidimensional models could be fitted to the subtests, but this approach is not efficient, since information from responses of correlated abilities is not taken into account. Among

the possible alternatives, three main approaches can be distinguished. The first is represented by the use of multi-unidimensional models (Sheng and Wikle, 2007), where each latent trait measures a single subscale, and abilities may be correlated. The second approach is given by the classical multidimensional IRT models (Béguin and Glas, 2001; Reckase, 2009), where each item may potentially load on all the latent trait, analogously to explanatory factor analysis for the case of continuous manifest variables. A third reasonable approach involves the existence of a general trait, denoting the overall ability, and specific traits which are related to the different subscales. To implement this approach, models with an additive or hierarchical structure can be used (Jennrich and Bentler, 2011; Sheng and Wikle, 2008; 2009).

In this paper, IRT models with different structures in the latent abilites are compared through simulation studies in order to understand their effective capability of recovering different test structures. Model estimation is conducted via Markov chain Monte Carlo (MCMC) methods, adopting a fully Bayesian approach (see *e.g.* Béguin and Glas, 2001; Sheng and Wikle, 2008). This approach has the advantage of estimating item parameters and individual abilities jointly. Moreover, uncertainties about item parameters and abilities can be incorporated into their respective prior distributions. Finally this approach has been applied to real data on Italian standardized student assessments.

2. Multidimensional IRT models

Multidimensional models are derived by extending models for unidimensional data. In particular, normal ogive models with two parameters, which are easy to be treated within a fully Bayesian approach, are reviewed here. Given a test consisting of k binary items measuring a single ability θ, and denoting y_{ij} the response of the candidate i, with i=1,...,n, to the item j, with j=1,...,k, the unidimensional two-parameter normal ogive model can be expressed as

$$P(y_{ij} = 1|\theta_i, \alpha_j, \delta_j) = \Phi(\alpha_j\theta_i - \delta_j) = \int_{-\infty}^{\alpha_j\theta_i - \delta_j} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt, \qquad (1)$$

where α_j and δ_j are the item discrimination and item threshold, respectively.

An explorative extension of model (1) to the presence of $\theta_1, \ldots, \theta_v, \ldots, \theta_m$ concurrent abilities is given by the following multidimensional model

$$P(y_{ij} = 1|\boldsymbol{\theta}_i, \boldsymbol{\alpha}_j, \delta_j) = \Phi(\sum_{v=1}^{m} \alpha_{vj}\theta_{vi} - \delta_j) = \int_{-\infty}^{\sum_{v=1}^{m} \alpha_{vj}\theta_{vi} - \delta_j} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt, \quad (2)$$

where $\boldsymbol{\theta}_i$ is the ability vector of dimension m for individual i, and $\boldsymbol{\alpha}_j$ is the vector of discrimination parameters for item j. In order to be identified, this model requires the imposition of strong constraints on the item parameters (Béguin and Glas, 2001). Moreover, due to its compensatory nature, several paradoxical results have been observed in the recent literature when applying scoring algorithms.

Assuming that the k-item test is divided into m subtests, each containing k_v items measuring a single ability, the probability function associated to the multi-unidimensional model becomes

$$P(y_{vij} = 1|\theta_{vi}, \alpha_{vj}, \delta_{vj}) = \Phi(\alpha_{vj}\theta_{vi} - \delta_{vj}) = \int_{-\infty}^{\alpha_{vj}\theta_{vi}-\delta_{vj}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt, \quad (3)$$

where all model parameters are referred to the ability dimension v. Model 3 has a confirmatory nature, and it was compared to the unidimensional model (1) by Sheng and Wikle (2007).

When it is possible to assume the concurrent presence of general and specific abilities, hierarchical models or additive models (Sheng and Wikle, 2008; 2009) may be appropriate in order to assign both an overall score and a set of subscores to individuals. In the first case, denoting with θ_0 the general trait, the measurement model is identified by (3) but the underlying latent structure is given by

$$\theta_{vi} = \beta_v \theta_{vi} + \epsilon_{vi}, \quad (4)$$

with $\epsilon_{vi} \sim N(0, 1)$.

Unlike the hierarchical model, the general ability directly affects the candidate's responses to the items in the test in the additive model, described by

$$P(y_{vij} = 1|\theta_{0i}, \theta_{vi}, \alpha_{0vj}\alpha_{vj}, \delta_{vj}) = \Phi(\alpha_{0vj}\theta_{0i} + \alpha_{vj}\theta_{vi} - \delta_{vj}). \quad (5)$$

3. Preliminary results and discussion

Estimation of models described in Section 2 is conducted adopting a fully Bayesian approach so that all dependencies among variables and sources of uncertainty could be incorporated. In particular, the estimation is based on Gibbs sampler (Geman and Geman, 1984) within the Markov chain Monte Carlo (MCMC) methods. An important advantage of this method is that the estimation is free from the limitations of using Gaussian quadrature in marginal maximum likelihood estimation (Béguin and Glas, 2001). The model performances were compared by using simulation studies and resorting to Bayesian model choice techniques, such as Bayes factors (BF) and information criteria based on Bayesian deviance. Two specific latent abilities were assumed, besides a general latent trait. Simulations were conducted by manipulating the ability correlation matrix, the sample size and the test length.

The main results show that models with more complex structures (hierarchical and additive models) outperform unidimensional and multi-unidimensional models. However, limitations due to the size of the sample should be considered carefully. In the application to Italian data on standardized student assessment, the responses to the INVALSI test of Italian language in the upper secondary school were analyzed according

to different models. Even if the test is scored on a single scale, it suggests at least a bidimensional structure. In fact, the test is divided in reading comprehension and grammar items. The results suggest that a multidimensional ability structure should be taken into account, and the assumption of the existence of a general trait is also satisfied.

This work is not without limitations. First of all, it represents only an attempt to compare multidimensional models under different conditions and within a fully Bayesian estimation. More complex ability structures should be evaluated, in terms of number of specific traits and hierarchical structures. Moreover, the sensitivity of the results to the inclusion of different prior distributions for the model parameters and hyperparameters should be studied in more detail. Finally, all the approaches considered have a confirmatory nature. Exploratory models were not taken into account because they are more difficult to be treated, due especially to model identification issues. However, in the analysis of real response data, these models should be considered as well.

Acknowledgements: This work was partially supported by a research grant from the Italian Ministry of University and Research (MIUR), PRIN 2008 on "Latent structure analysis: new boundaries in statistical methods and models".

References

Béguin A.A., Glas C.A.W. (2001), MCMC estimation and some model-fit analysis of multidimensional IRT models, *Psychometrika*, 66, 541–562.

Geman S., Geman D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Jennrich R. I., Bentler P. M. (2011), Exploratory bi-factor analysis, *Psychometrika*, 76, 537–549.

Reckase M. (2009), *Multidimensional Item Response Theory*, Springer-Verlag, New York.

Sheng Y., Wikle C. K. (2007), Comparing multiunidimensional and unidimensional item response theory models, *Educational and Psychological Measurement*, 67, 899–919.

Sheng Y., Wikle C.K. (2008), Bayesian multidimensional IRT models with an hierarchical structure, *Educational and Psychological Measurement*, 68, 413–430.

Sheng Y., Wikle C.K. (2009), Bayesian IRT models incorporating general and specific abilities, *Behaviormetrika*, 36, 27–48.

Quaderni di Statistica
Vol. 14, 2012

Assessing perceived food chemical risk by means of an educational project analysed with permutation tests

Federico Mattiello
Department of Statistics, University of Bologna
E-mail: federico.mattiello2@unibo.it

Mario Bolzan
Department of Statistics, University of Padova
E-mail: mario.bolzan@unipd.it

Licia Ravarotto
Health Awareness and Communication Department,
Istituto Zooprofilattico Sperimentale delle Venezie
E-mail: lravarotto@izsvenezie.it

Summary: In this study two *media* were given to students from 4th and 5th year of some high schools in Italy (the target population), they were asked to fill one questionnaire before (*ex-ante*) and one after (*ex-post*) seeing them. These forms aim to evaluate three different aspects: the perceived food chemical risk, the knowledge of the concept "chemical risk" and the attractiveness of the two *media*. We wanted to test if the differences before/after were different between the two instruments (comic book and video) and if the two *media* were different themselves in terms of attractiveness (hence only looking at *ex-post* questions). To reach this goal we made use of permutation tests on matching items between *ex-ante* and *ex-post* forms, and nonparametric combination methodology (NPC) to gather information into an overall test.

Keywords: Educational project, Permutation tests, NonParametric combination.

1. The project

This educational project was carried out by the Health Awareness and Communication department of "Istituto Zooprofilattico Sperimentale delle Venezie": a public veteri-

nary Institute which conducts prevention, control and research activities in animal health and welfare and food safety and it was funded by the Italian Ministry of Health. The research group in charge for this project is involved in several studies and activities on risk perceptions and communication (Trifiletti *et al*. (2012)). The project was designed to evaluate three different aspects of interests for public health and social welfare: **(i)** the chemical risk perception related to daily food consumption, with items that directly ask how much students feel threatened by some topics (such as OGMs, use preservatives or the lack of food safety controls); **(ii)** the knowledge of the concept "chemical risk" itself, with items that ask for specific definition or bad behaviour related to food; **(iii)** the better tool to convey information about these topics by means of the storytelling method, with items that directly ask which aspects of the instruments the students appreciated and which not. These features along with the fact that the target population was directly involved in the whole project (from designing to experimenting), makes it quite original and somewhat unique especially in this field. Moreover, also topics of interest are quite complex in the sense that are difficult to measure because they are latent and depend on a number of unknown information (*e.g*. personal beliefs or attitudes); this reflects on the questionnaires that are tailored as precisely as possible to the specific step of the study and hence they are mostly exploratory. These reasons led us to keep the analysis as simple as possible in order to maintain comprehension of results. This is the reason why we kept into account only items that matches between the *ex-ante* and *ex-post* forms. More in detail, we considered two questions for the perceived risk index: R_1) *how much do you feel exposed to contaminants that can be present in foods* (on a scale from 0 to 3), R_2) *how much do you feel worried by use of pesticides, preservatives, quality of food you eat out of home* (3 items on a scale from 1 to 10); two questions for the knowledge index: K_1) *the definition of chemical food risk* (right or wrong answer, taking values 0 or 1), K_2) *composed by* 6 *items about true or false statements mainly related to food behaviour* (so this answer will be on a scale from 0 to 6). In comparing the two tools instead, we obviously have considered only the corresponding 5 questions of *ex-post* forms: T_1) *how much did you like the tool* (on a scale from 1 to 10), T_2) *how many times did you read it?* (on a scale from 0 to 4), T_3) *how much did you appreciate graphical aspects?* (5 items on a scale from 1 to 10), T_4) *how much effective do you think the tool is?* (6 items on a scale from 1 to 10), T_5) *do you think the tool is able to: catch the attention, convey scientific contents, emotionally involve the person interviewed* (3 items on a scale from 0 to 3). Note that for the knowledge index we used $K_1 + K_2$ in order to have a single "item" that measure the number of correct answers, hence it is on a scale from 0 to 7; this is not a problem in the sense that the distribution of K_2 is not so influenced by K_1 and it seemed reasonable to gather together statements concerning the knowledge of "chemical food risk" (statements in K_2 are also quite heterogeneous though).

2. Analysis

The main focuses here are to test which *medium* is more effective in increasing the perceived chemical food risk and the knowledge of it. This means that we are testing if the distributions of the numeric differences of matching items $d_{Mi} = x_{Mi}^{\text{post}} - x_{Mi}^{\text{ante}}$, $M = \{\text{comic}, \text{video}\}$ are symmetric or not, and more precisely if the distribution of D_{Mi}, $M =$ "comic" is different from that of D_{Mi}, $M =$ "video"[1]. In view of this the global null hypothesis can be written as $\{H_0^G := H_0^R \bigcap H_0^K\}$, where exponents means respectively Global Risk and Knowledge, but here we are interested in the marginal hypotheses rather than the global one. The global alternative is simply $\{H_1^G := H_1^R \bigcup H_1^K\}$ where both H_1^R and H_1^K are two-sided alternatives, in the sense that become active if the observed test statistic happens to lie on one of the tails of its (permutation) distribution, *i.e*.

$$H_1^K := \left\{ D^K_{\text{comic}} \overset{d}{<} D^K_{\text{video}} \right\} \bigcup \left\{ D^K_{\text{comic}} \overset{d}{>} D^K_{\text{video}} \right\},$$

where D^K_{video} is the random variable for knowledge index. The procedure we used to obtain the perceived risk and the knowledge indexes can be described with the following steps: calculate $d_{Mi}, M = \{\text{comic}, \text{video}\}$, perform a permutation test separately for each item with a proper test statistic (in this case a permutation version of the Anderson-Darling, see Pesarin F. and Salmaso L. (2010) for details about this and other test statistics useful in this framework), combine together the tests using a combining function (*e.g*. Fisher's p-values combining function) and calculate the (permutation) p-values for the hypotheses of interest. In particular we used an *R* package named "SOUP" (Stochastic Ordering using Permutation tests and pairwise comparisons)[2], that is a substantial improvement of Mattiello's master thesis (Mattiello (2010) and Arboretti *et al*. (2010)).

3. Results

Hereafter we report the matrices of p-values resulting from the analysis: both should be read as "row$_i$ > column$_j$" (the distribution of) in the alternative and equivalence in the null hypothesis. We used $10{,}000$ random permutation of the pooled sample which is composed by $n = 117$ "video" data and $n = 72$ "comic" data[3].
We can see that the only active hypothesis is the one related to the knowledge index with $H_1 :=$ "video" > "comic" because the associated test is rejected at all reasonable levels, so we can at least deduce that (with the given the data) the video tool is more effective than the comic book in increasing the knowledge about chemical food risk and hence in conveying information about the topic. Regarding the risk index instead,

[1] D_{Mi} is the random variable associated to d_{Mi}

[2] not yet submitted to CRAN

[3] so we had a permutation space cardinality of 2.095833×10^{53}

we cannot say there is any significant difference between the two instruments but this is not necessarily a drawback, in fact one interpretation could be that the perceived threat related to food is changed (or not changed) in the same way by the two *media*. Concerning the tool evaluation index instead, the comic book seems to have been more appreciated by students because the associated test is rejected at all usual value. One possible explanation is suggested by answers that students gave to free text questions (not reported here): comic book is easier to use than the video because it needs no laptop nor internet connection (apparently a problem for some students) to be seen and can be easily shared with others.

Table 1. P-values matrices of the tests.

	$\geqq$	comic	video
risk index	comic	–	0.3673
	video	0.6337	–
knowledge index	comic	–	0.9947
	video	0.0054	–
tool evaluation	comic	–	0.0004
	video	1	–

Acknowledgements: The authors would like to thank the Health Awareness and Communication Department of I.Z.S.Ve., especially Stefania Crovato and Giulia Mascarello.

References

Pesarin F., Salmaso L. (2010), *Permutation Tests for Complex Data*, Wiley & Sons, Chichester.

Arboretti R.G., Corain L., Gomiero D. and Mattiello F. (2010), Nonparametric multivariate ranking methods for global performance indexes, *Quaderni di Statistica*, 12, 79–106.

Mattiello F. (2010), Some resampling-based procedures for ranking of multivariate populations, University of Padua, Faculty of Statistical Sciences, *Master Thesis*.

Trifiletti E., Crovato S., Capozza D., Visintin E.P., Ravarotto L. (2012), Evaluating the effects of a message on attitude and intention to eat raw meat: Salmonellosis prevention, *Journal of Food Protection*, 75, 394–399.

Quaderni di Statistica
Vol. 14, 2012

A multilevel model with time series components for the analysis of tribal art prices

Lucia Modugno Simone Giannerini Silvia Cagnone
Department of Statistics, University of Bologna
E-mail: lucia.modugno@unibo.it, simone.giannerini@unibo.it, silvia.cagnone@unibo.it

Summary: In the present work, we extend the classic multilevel model to include time series components at the second level. In order to show its potentials, we perform an econometric analysis of the Tribal art market. In literature, art prices are modelled through the hedonic regression model, a classic fixed-effect model. We use, instead, a multilevel model for the analysis of Tribal art prices that takes into account the influence of time on artwork prices. Since we do not have repeated measurements of the same items over time, we propose to treat the data as two-level structured in that items are grouped in time points. Hence, the proposed model copes with the time dependence of random effects.

Keywords: Multilevel model, Hedonic regression model, Dependent random effects.

1. Introduction

In the present work, we extend the classic multilevel model to cope with autocorrelated random effects. In order to show its potentials, we apply the new model to a real database. The relevant data come from the first world database of Tribal art prices, built by a team of researchers of the University of Bologna, Faculty of Economics – Rimini (see Modugno and Giannerini, 2008). The database contains more than 20000 records of items sold from 1998 to 2011 by the most important auction houses.

Among the existing methods to build indexes for prices of artworks, the most suitable for fitting our data is the hedonic regression model (Rosen, 1974). Thus, this work has two innovative contributions. First, it extends the multilevel model to include time series components at the second level, and shows its application to a real database (Section 3). Moreover, it proposes the new multilevel model as a new approach for the analysis

of art prices that will give a substantial advantage over the traditional fixed-effect model (section 2). For this reason, we chose to start by fitting the traditional hedonic regression model upon the Tribal art data, fitting a classic multilevel model and then extending it to cope with the case under investigation.

2. Hedonic regression model and multilevel model for Tribal art prices

In literature, art prices are modelled through the hedonic regression model, a fixed-effect model that takes into account the heterogeneity of artworks by explaining prices through object features and constructs a price index by neutralizing the effect of quality. Our idea, instead, is to consider the influence of time effects on prices through a different approach. Since we observe different artworks sold at every auction, Tribal art data do not constitute either a panel or a time series. Rather, they have a two-level structure in that items, level-1 units, are grouped in time points, level-2 units. Hence, we propose the multilevel model to explain heterogeneity of prices among, in particular, semesters.

Consider an hedonic regression model for the price of artworks, that we call "FE-hedonic", with the vector $\mathbf{x}_{it}$ containing k object features as covariates:

$$\log_{10}(y_{it}) = \beta_{0t} + \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{it}, \quad \epsilon_{it}|\mathbf{x}_{it}^{\mathrm{T}} \sim \mathrm{NID}(0, \sigma^2) \tag{1}$$

for the semester $t = 1, \ldots, T$ and the item $i = 1, \ldots, n_t$, where n_t is the total number of items sold in the semester t. The dependent variable is the logarithm of the observed price, and β_{0t} represents the mean price of the semester t.

The corresponding two-level model, that we call "RE-hedonic", has the same form as (1) but the time-specific intercept is modelled as (Skrondal and Rabe-Hesketh, 2004)

$$\beta_{0t} = \beta_0 + u_t, \quad u_t|\mathbf{X}_t \sim \mathrm{NID}(0, \sigma_u^2), \quad u_t \perp \epsilon_{it} \tag{2}$$

where β_0 is the overall mean price and u_t is a random intercept for the semester t.

The RE-hedonic model produces very similar estimates and residuals to the FE-hedonic model but with more parsimony. However, the assumption of normality for the (first level) errors of both models are not valid. On the other hand, the predicted random effects are normally distributed but they are not independent for different groups as the multilevel model assumes. In particular, the correlograms of level-2 residuals point at an autoregressive-like structure, similar to that of an AR(1) process.

Improving the classical multilevel model, for this case, requires relaxing the assumption of independence among random effects. Since they represent time effects, the inclusion of such correlation implies treating them as a time series. To our knowledge, in literature, the applications of multilevel models to longitudinal data consider occasions, that is the points in time, as the lowest level units and individuals as higher units. Therefore, any time dependence structure is assessed at the first level. The main theoretical contribution of this work is the derivation of a multilevel model with time series components at the second level, which the next section is devoted to.

3. A multilevel model with time series components

Consider a random intercept model with k level-1 covariates:

$$y_{it} = \beta_{0t} + \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{it}, \quad \epsilon_{it}|\mathbf{x}_{it}^{\mathrm{T}} \sim \mathrm{NID}(0, \sigma^2) \tag{3}$$

for $i = 1_t \ldots, n_t$ and $t = 1, \ldots, T$. The slopes in $\boldsymbol{\beta}$ are fixed; the intercepts β_{0t}, instead, are group-specific and random, and they are modeled as

$$\beta_{0t} = \beta_0 + u_t,$$

where u_t represents the deviation of the group-specific intercept β_{0t} from the overall mean, β_0. The usual assumption of independence for the random effects in (2) is relaxed by assuming an autoregressive process of order 1 for the level-2 errors:

$$u_t = \rho u_{t-1} + \eta_t, \qquad \eta_t|\mathbf{X}_t \sim \mathrm{NID}(0, \sigma_\eta^2),$$

with $|\rho| < 1$, that guarantees stationarity, $\eta_t \perp u_s$ and $\eta_t \perp \epsilon_{it}$ for all $s < t$ and for all i.

We then proceed to the parameter estimation. We chose the full maximum likelihood estimation method and implemented it through the E-M algorithm.

In order to assess the finite sample properties of the found estimators and to validate the practical implementation of the algorithm, we have performed a Monte Carlo study that has confirmed good finite sample performances of the estimators (Modugno, 2012).

4. The new model and Tribal art prices

In this section, we fit the new model upon the Tribal art dataset, by including the same covariates as in the previous models.

The estimates and the predicted random effects are quite close to both competing models. However, the proportion of variability explained by the between-semesters variance (ICC) is bigger for the AR-RE-hedonic model than for the the RE-hedonic model. This confirms that, at least in part, the structure at the second level has been taken into account. Also in this case, the hypothesis of normality for the level-1 residuals is rejected but that for the random effects is not rejected. The autocorrelation functions (global and partial) of level-2 residuals do not reveal any residual dependence structure. Hence, the autoregressive behaviour of the level-2 residuals observed for the RE-hedonic model has been completely absorbed by an AR(1) process for the level-2 errors.

Finally, we compare the prediction capability of the AR-RE-hedonic and RE-hedonic models for both level-1 and level-2 responses. On the one hand, the models predict responses of out-of-sample level-1 units belonging to existing groups with similar performances. On the other hand, the AR-RE-hedonic model allows to forecast better the effect of a 1-lagged out-of-sample semester through the AR(1) process, and, therefore, the prices of objects sold in that semester.

Table 1. Akaike Information Criterion, Bayesian Information Criterion and Root Mean Square Error for prediction of responses of units in the semester $T + 1$

	FE-hedonic	RE-hedonic	AR-RE-hedonic
AIC	15273	15365	15307
BIC	16013	15916	15866
RMSE	0.371	0.368	0.325

Table 1 sums up the comparison among the three models fitted on Tribal art data. The AR-RE-hedonic model has smaller AIC and BIC than the RE-hedonic model, thus it provides a better fit. Moreover, with respect to the FE-hedonic model, although the new model presents similar estimates and similar interpretation of results, it has less parameters to be estimated resulting in a smaller BIC and provides a decomposition of the total variability of the response (as the classic multilevel model). Finally, the AR-RE-hedonic allows a better forecasting of the responses of units in a 1-lag-ahead group, that are the prices of objects that will be sold one semester later, as shown by the reduced RMSE. Therefore, the proposed model improves considerably the fit of the Tribal art data with respect to both the hedonic regression model and the classic multilevel model providing a more flexible framework.

References

Modugno L., Giannerini S. (2008), La prima banca dati dedicata all'arte etnica: prime evidenze empiriche, *Sistema economico*, 2, 45–57.

Modugno L. (2012), A Multilevel Model with Time Series Components for the Analysis of Tribal Art Prices, *PhD dissertation*, University of Bologna.

Rosen S. (1974), Hedonic prices and implicit markets: product differentiation in pure competition, *Journal of Political Economy*, 82, 34–55.

Skrondal A., Rabe-Hesketh S. (2004), *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*, Chapman & Hall/CRC, New York.

Quaderni di Statistica
Vol. 14, 2012

Modeling count panel data with a Zero-Inflated Poisson model

Anthea Monod
Department of Mathematics, Swiss Federal Institute of Technology (EPFL)
E-mail: Anthea.Monod@epfl.ch

Summary: In the present work, we model and analyse count data (nonnegative integers) in a panel (longitudinal) data framework, where the responses comprise an abundance of zero-valued observations. In particular, we adapt the zero-inflated Poisson model to the context of generalised linear models to address the presence of excess zero values in the responses. By deriving a conditional probability specification utilising a cross-sectional random effect, we address the problem of serial correlation in the panel data and simultaneously pacify the problem of overdispersion. The model is fitted via the method of maximum likelihood and the finite sample performance is tested via a simulation study.

Keywords: Overdispersion, Generalised Linear Models, Random Effects Modeling.

1. Introduction

Data taken for several subjects over several time periods taking the form of nonnegative integers arise in many fields of science, and occur in the particular case when measurements are recorded by counting. In counting the occurrence of a particular phenomenon, there is also the possibility that there is no occurrence to count, in which case, this measurement would be zero.

Statistical and econometric modeling of count panel data has been previously addressed by Hausman, Hall & Griliches (1984), most notably via the introduction of cross-sectional-specific effects; excess zeros are not addressed in this work. Meanwhile, the presence of abundant zero observations in count data has been addressed by Lambert (1992) via the construction of a class of zero-inflated regression models, though mixed-effect and panel-data issues were not considered.

We consolidate these approaches to model count panel data consisting of excess zeros by considering a zero-inflated Poisson model in the context of generalised linear

models and introduce a cross-sectional random effect, which addresses the problem of serial correlation in the panel data and simultaneously appeases the problem of overdispersion that commonly arises in data with excess zeros, where the variance exceeds that stipulated by the parametric model. From this, we derive a conditional zero-inflated Poisson probability; we fit the model via maximum likelihood and perform simulation studies to test the finite sample performance.

2. The Zero-Inflated Poisson Generalised Linear Model with Cross-Sectional Random Effects

The technique of zero-inflated Poisson (ZIP) regression presented by Lambert (1992) gives rise to a new class of regression models for count data with an abundance of zero observations by assuming that the integer-valued response $Y \sim \mathrm{ZIP}(\alpha, \lambda)$ is distributed as a mixture of a Poisson distribution $\mathcal{P}(\cdot)$ with parameter λ, and a distribution with a point mass of one at the value zero, with mixing probability α:

$$Y \sim \begin{cases} 0 & \text{with probability } \alpha; \\ \mathcal{P}(\lambda) & \text{with probability } 1-\alpha. \end{cases} \tag{1}$$

To study count panel data in the context of generalised linear models, a log-linear link $\log \lambda_{it} = \mathbf{x}_{it}^{\top}\boldsymbol{\beta}$ for the Poisson parameter λ_{it}, assumed to vary with cross-sectional unit $i = 1, 2, \ldots, N$ and time period $t = 1, 2, \ldots, T_i$ is specified, which relates linearly to a matrix of given covariates $\mathbf{x}_{it}$, and the vector of parameters to be estimated, $\boldsymbol{\beta}$.

While the repetition of repeated measurements for each unit i taken over several time periods t presents the advantage of obtaining more precise parameter estimates, it also presents the disadvantage that the observations are not independent so standard regression methods can not be applied. Serial correlation is expected in panel data due not only to the time-series aspect, but also due to an implicit dependence between covariates since all measurements are taken on the same unit i, which can be difficult to specify precisely. To address this correlation, the existence of a linear, unobservable, cross-sectional-specific random effect $\zeta_i = Z_0 + Z_i$ is often assumed; these independent random effects account for autocorrelation across all time periods t and simultaneously maintain noncorrelation across individual units i. Integrating this effect into the Poisson generalised linear model gives

$$\log \tilde{\lambda}_{it} = \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + \zeta_i \iff \tilde{\lambda}_{it} = e^{\mathbf{x}_{it}^{\top}\boldsymbol{\beta}} e^{\zeta_i} = \lambda_{it} e^{\zeta_i}, \tag{2}$$

rendering the modified Poisson parameter $\tilde{\lambda}_{it}$ random, as opposed to λ_{it}, a deterministic function of $\mathbf{x}_{it}$. Indeed, the independent e^{ζ_i} ensure cross-sectional noncorrelation between $\tilde{\lambda}_{it}$ and $\tilde{\lambda}_{jt}$ for $i \neq j$, and autocorrelation between $\tilde{\lambda}_{it}$ and $\tilde{\lambda}_{is}$ for $t \neq s$.

Incorporating the now-random Poisson parameter $\tilde{\lambda}_{it}$ given by Equation (2) in the

ZIP model in Equation (1), we obtain the following conditional probability specification

$$\begin{aligned}\text{Prob}(Y_{it}=y_{it}|\mathbf{x}_{it},\zeta_i) &= \begin{cases}\alpha(1)+(1-\alpha)\text{Prob}(Y_{it}=y_{it}|\mathbf{x}_{it},\zeta_i) & \text{if } y_{it}=0;\\ (1-\alpha)\text{Prob}(Y_{it}=y_{it}|\mathbf{x}_{it},\zeta_i) & \text{otherwise.}\end{cases}\\ &= \begin{cases}\alpha+(1-\alpha)\exp(-\lambda_{it}e^{\zeta_i}) & \text{if } y_{it}=0;\\ (1-\alpha)\dfrac{(\lambda_{it}e^{\zeta_i})^{y_{it}}\exp(-\lambda_{it}e^{\zeta_i})}{(y_{it})!} & \text{otherwise.}\end{cases}\end{aligned} \tag{3}$$

Effectively, by posing a gamma distribution with parameters (u,u) for the cross-sectional random effects $e^{\zeta_i}\sim\Gamma(u,u)$, Equation (3) gives

$$Y_{it}\sim\begin{cases}0 & \text{with probability } \alpha;\\ \text{Neg.Bin.}\left(u,\dfrac{u}{\lambda_{it}+u}\right) & \text{with probability } 1-\alpha,\end{cases} \tag{4}$$

since recall for some $X\sim\mathcal{P}(\lambda\xi)$ where $\xi\sim\Gamma(\varphi,\psi)$, X is then distributed as a negative binomial random variable with parameters φ and ψ; see Monod (2007) for more details. Asymptotically, this precisely gives us the ZIP model with parameter λ_{it}.

2.1. Overdispersion

For a standard parametric Poisson model, $E[Y_{it}]=\text{Var}(Y_{it})=\lambda_{it}$, a constraint that is usually too restrictive and in practical applications, often it is the case that $E[Y_{it}]<\text{Var}(Y_{it})$, *i.e.* that overdispersion is present.

In using a ZIP model, we have already allowed the variance to exceed the mean, since for $Y_{it}\sim\text{ZIP}(\alpha,\lambda_{it})$ with a fixed $\alpha>0$, we have $E[Y_{it}]=(1-\alpha)\lambda_{it}$ and $\text{Var}(Y_{it})=(1-\alpha)\lambda_{it}(1+\alpha\lambda_{it})$ so indeed $E[Y_{it}]<\text{Var}(Y_{it})$. Additionally, from Equation (4), for the first and second moments, we find $E[Y_{it}]=\lambda_{it}$ and $\text{Var}(Y_{it})=\lambda_{it}(1+\lambda_{it}/u)$, contributing to the variance of the model by an amount of λ_{it}/u.

3. Estimation & Simulation Studies

The parameters of interest in our model to be estimated by maximum likelihood are those of the Poisson component in the parameter vector $\boldsymbol{\beta}$, that of the gamma component for the cross-sectional random effects u, and the mixing probability of the ZIP model α. In general, this gives a $(2+\max\{T_1,\ldots,T_N\})$-dimensional parameter vector. Through the cross-sectional random effects given in Equation (2), we indeed have *i.i.d.* observations over i which indeed allows a likelihood to be written; the serial correlation must be taken into account in the computation of the density. As the conditional probability depends on whether the value of an observation is zero or not, three cases are distinguished to construct components of the likelihood function: all t for a given i are nonzero; all t for a given i are zero; the t for a given i are a mixture of zero and nonzero values.

In simulation studies, we assumed a univariate covariate x_{it} generated from a normal $N(0,1)$ distribution, and set $T_i = T = 2$ for all i and our parameter vector is $\boldsymbol{\vartheta} = (\alpha, \beta_1, \beta_2, u) = (0.22, 0.5, 0.5, 1.5) \in \mathbf{R}^4$. Numerical optimization is carried out by a grid search method. Two simulation studies were done, comprising of 1000 observations each; one with a coarser partition of 10 and 100 simulations executed, the other with a finer partition of 20 and 500 simulations executed. The average maximum likelihood estimates $\widehat{\boldsymbol{\vartheta}}$ and their variances of the parameter values recovered across all simulations are given below.

Table 1. Results of Simulation Studies

			Simulation 1		Simulation 2	
$\boldsymbol{\vartheta}$		$\widehat{\boldsymbol{\vartheta}}$	Mean	Variance	Mean	Variance
α	0.22	$\hat{\alpha}$	0.171000	0.000736	0.170100	0.001460
β_1	0.5	$\hat{\beta}_1$	0.531999	0.024218	0.544101	0.022585
β_2	0.5	$\hat{\beta}_2$	0.529999	0.024343	0.542200	0.022363
u	1.5	$\hat{u}$	1.346999	0.111809	1.346400	0.094496

4. Conclusion

Values recovered for the β_t are close to the true values, however the results for the α and u parameters were somewhat less accurate. These two parameters seem to be negatively related: if the algorithm estimates that α should be small, then the value of u should also be small so that the zero values are due to the Poisson contribution of the model; the nature of the problem essentially renders it difficult to distinguish between two parameters that both contribute to the same effect in the generation of the data. Nevertheless, the model performs well for the estimation of the cross-sectional random effects, which remains effective in dealing with serial correlation in panel data.

References

Hausmann J., Hall B. H., Griliches Z. (1984), Econometric Models for Count Data with an Application to the Patents-R& D Relationship, *Econometrica*, 52(4), 909–938.

Lambert D. (1992), Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34(1), 1–14.

Monod A. (2007), *An Analysis on Count Panel Data Using a Zero-Inflated Posison Model*, M.Sc. Thesis, Institute of Statistics, University of Neuchâtel, Switzerland.

Quaderni di Statistica
Vol. 14, 2012

On the Stationarity of Threshold Models with Multiple Variables

Marcella Niglio
Department of Economics and Statistics, University of Salerno
E-mail: mniglio@unisa.it

Cosimo Damiano Vitale
Department of Economics and Statistics, University of Salerno
E-mail: cvitale@unisa.it

Summary: In our contribution the weak stationarity (stationarity in the following) is faced for a nonlinear multivariate time series model called Threshold Vector Autoregressive Moving Average (TVARMA). Its stochastic structure is based on k VARMA regimes whose switching among them is based on a univariate threshold variable. After the presentation of the model, the stationarity conditions are depicted generalizing some results given in the univariate domain for the threshold models.

Keywords: Nonlinearity , Stationarity, Threshold models.

1. The Threshold Vector ARMA model

Since their introduction in the late seventies, threshold models have raised the attention of the literature for their flexibility to catch different features of the data generating process. Their impact in nonlinear time series domain has been enormous and after more than 30 years new results and further developments on this class of models are proposed to face problems that are still open (for a recent review on the threshold models see Tong (2011)). In our contribution the attention will be focused on the so called Threshold Autoregressive Moving Average model, proposed in the univariate domain in Tong (1983) and here generalized to the multivariate context. More precisely, let $\boldsymbol{y}_t$ a K-variate time series, it is said to follow a Threshold Vector Autoregressive Moving

Average (TVARMA) model if:

$$\begin{aligned} \boldsymbol{y}_t &= \sum_{k=1}^{\ell} \Big[\boldsymbol{A}_1^{(k)} \boldsymbol{y}_{t-1} + \ldots + \boldsymbol{A}_p^{(k)} \boldsymbol{y}_{t-p} + \\ &\quad \boldsymbol{u}_t + \boldsymbol{M}_1^{(k)} \boldsymbol{u}_{t-1} + \ldots + \boldsymbol{M}_q^{(k)} \boldsymbol{u}_{t-q} \Big] I(x_{t-d} \in \mathcal{R}_k), \end{aligned} \tag{1}$$

where $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{K,t})', \boldsymbol{u}_t = (u_{1,t}, \ldots, u_{K,t})', \{\boldsymbol{u}_t\}$ is a sequence of independent white noise with $E[\boldsymbol{u}_t] = \boldsymbol{0}$ and covariance matrix $\Sigma_{\boldsymbol{u}}$, $\boldsymbol{A}_i = [a_{s,r}]$, $\boldsymbol{M}_i = [m_{s,r}]$ are the $(K \times K)$ matrices of coefficients, for $i = 1, \ldots, p$ and $j = 1, \ldots, q$. In model (1) x_t is the univariate strictly stationary and ergodic *threshold process*, d is the threshold delay, $\mathcal{R}_k = [r_{k-1}, r_k)$ such that $-\infty = r_0 < r_1 < \ldots < r_\ell = \infty, \bigcup_{k=1}^{\ell} \mathcal{R}_k = \mathcal{R}, r_k$ is the so called *threshold value* (for $k = 1, 2, \ldots, \ell$).

Note that model (1) has a local linear structure characterized by k VARMA regimes where the switching among them is regulated by a latent variable (related to x_t) which assumes two values: 0 and 1.
Further model (1) can be written equivalently as TVAR(ℓ; 1) process:

$$\boldsymbol{Y}_t = \boldsymbol{A}^{(k)} \boldsymbol{Y}_{t-1} + \boldsymbol{U}_t, \qquad \text{if} \quad x_{t-d} \in \mathcal{R}_k, \tag{2}$$

with

$$\underset{[K(p+q)\times 1]}{\boldsymbol{Y}_t} = \begin{bmatrix} \boldsymbol{y}_t \\ \vdots \\ \boldsymbol{y}_{t-p+1} \\ \boldsymbol{u}_t \\ \vdots \\ \boldsymbol{u}_{t-q+1} \end{bmatrix}, \qquad \underset{[K(p+q)\times 1]}{\boldsymbol{U}_t} = \begin{bmatrix} \boldsymbol{u}_t \\ \underset{[K(p-1)\times 1]}{\boldsymbol{0}} \\ \boldsymbol{u}_t \\ \underset{[K(q-1)\times 1]}{\boldsymbol{0}} \end{bmatrix},$$

and matrix $\boldsymbol{A}^{(k)}$ so partitioned

$$\underset{[K(p+q)\times K(p+q)]}{\boldsymbol{A}^{(k)}} = \begin{bmatrix} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{A}_{12}^{(k)} \\ \boldsymbol{A}_{21}^{(k)} & \boldsymbol{A}_{22}^{(k)} \end{bmatrix},$$

where

$$\underset{(Kp\times Kp)}{\boldsymbol{A}_{11}^{(k)}} = \begin{bmatrix} \boldsymbol{A}_1^{(k)} & \ldots & \boldsymbol{A}_{p-1}^{(k)} & \boldsymbol{A}_p^{(k)} \\ \boldsymbol{I}_K & \ldots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \vdots & \boldsymbol{I}_K & \boldsymbol{0} \end{bmatrix}, \qquad \underset{(Kq\times Kq)}{\boldsymbol{A}_{12}^{(k)}} = \begin{bmatrix} \boldsymbol{M}_1^{(k)} & \ldots & \boldsymbol{M}_{q-1}^{(k)} & \boldsymbol{M}_q^{(k)} \\ \boldsymbol{0} & \ldots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \ldots & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

$$\underset{(Kp\times Kp)}{\boldsymbol{A}_{21}^{(k)}} = \boldsymbol{0}, \qquad \underset{(Kq\times Kq)}{\boldsymbol{A}_{22}^{(k)}} = \begin{bmatrix} \boldsymbol{0} & \ldots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{I}_K & \ldots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \ldots & \boldsymbol{I}_K & \boldsymbol{0} \end{bmatrix}.$$

Given the one to one relation between model (1) and model (2), the results given in the following are presented for the latter model and can be extended, without further developments, to model (1).

2. TVARMA stationarity

To discuss the stationarity conditions of the TVARMA process (1) it is needed to distinguish between the so called *local* and *global* stationarity.

Definition 1. *Let* $\boldsymbol{y}_t \sim$*TVARMA(*$\ell; p, q$*), it is said to be locally stationary if the dominant eigenvalues of* $\boldsymbol{A}^{(k)}$ *in model (2), for* $k = 1, \ldots, \ell$*, are all less than 1.*

It is interesting to note that in presence of models with switching regimes, the local stationarity could not imply the stationarity of the entire process (even called *global stationarity*). In the multivariate domain it has been clearly stated in Francq and Zakoïan (2001) that discuss, among the others, the stationarity conditions of the Markov-switching model where the VARMA coefficients are allowed to change according to a Markov chain.
The stationarity conditions of model (2) can be obtained extending some results recently given in Niglio and Vitale (2012) for the univariate Threshold ARMA models.
To simplify the presentation, the stationarity conditions are proposed for a TVARMA(2;p, q) model, with $k = 2$ regimes. The generalization to the case with $k > 2$ regimes can be easily obtained.
Starting from model (2), the TVARMA structure under analysis can be represented as:

$$\boldsymbol{Y}_t = \boldsymbol{A}^{(1)}\boldsymbol{Y}_{t-1}I_{t-d} + \boldsymbol{A}^{(2)}\boldsymbol{Y}_{t-1}(1 - I_{t-d}) + \boldsymbol{U}_t,$$

with $I_{t-d} = I(x_{t-d} \in \mathcal{R}_1)$ and $R_1 = (-\infty, r_1]$.
After m iterations $\boldsymbol{Y}_t$ becomes:

$$\begin{aligned}\boldsymbol{Y}_t = \quad & \prod_{j=0}^{m-1}\left[\boldsymbol{A}^{(1)}I_{t-d-j} + \boldsymbol{A}^{(2)}(1 - I_{t-d-j})\right]\boldsymbol{Y}_{t-m} + \\ & \sum_{w=1}^{m-1}\prod_{j=0}^{w-1}\left[\boldsymbol{A}^{(1)}I_{t-d-j} + \boldsymbol{A}^{(2)}(1 - I_{t-d-j})\right]\boldsymbol{U}_{t-w} + \boldsymbol{U}_t. \qquad (3)\end{aligned}$$

Note that model (3) is stationary if $\prod_{j=0}^{m-1}\left[\boldsymbol{A}^{(1)}I_{t-d-j} + \boldsymbol{A}^{(2)}(1 - I_{t-d-j})\right]$ converges (at least in probability) to a null matrix, as m grows.
Using arguments similar to those presented in Niglio and Vitale (2012), it can be shown the following theorem.

Theorem 1. *Let* $\boldsymbol{y}_t \sim TVARMA(2; p, q)$ *process defined in (3) with* $\boldsymbol{A}^{(1)}$ *and* $\boldsymbol{A}^{(2)}$ *having eigenvalues* $\lambda_{i,1} < \lambda_{i,2} < \ldots < \lambda_{i,K}$*, for* $i = 1, 2$*. Then* $\boldsymbol{y}_t$ *is stationary if*

$$\rho(\boldsymbol{A}^{(1)})^{p_1} \rho(\boldsymbol{A}^{(2)})^{1-p_1} < 1, \tag{4}$$

with $0 < p_1 = E[I_{t-d}] < 1$ *and* $\rho(\boldsymbol{A}^{(i)})$ *the dominant eigenvalue of* $\boldsymbol{A}^{(i)}$*, for* $i = 1, 2$.

What stated in Theorem 1 can be extended to the more general case with $k > 2$ regimes whereas when $p_1 = 0$ or $p_1 = 1$ the condition (4) equals the stationarity conditions of the linear VARMA models.

The results depicted in the previous pages are only a first step to investigate the stationarity of the TVARMA process. It could be interesting to evaluate, among the others, if the stationarity of model (2) with $k > 2$ regimes can be based, as in the univariate domain and under well defined assumptions on p and q, on proper conditions on the parameters of only the first and the last regime. It will be object of future research.

References

Francq C., Zakoïan J.M. (2001), Stationarity of multivariate Markov-switching ARMA models, *Journal of Econometrics*, 102, 339-364.

Niglio M., Vitale C.D. (2012), Local unit roots and global stationarity of TARMA models, *Methodology and Computing in Applied Probability*, 14(1), 17-34.

Tong, H. (1983), *Threshold models in nonlinear time series analysis*, Springer-Verlag, London.

Tong H. (2011), Threshold models in time series analysis - 30 years on, *Statistics and Its Interface*, 4(2), 123-128.

Quaderni di Statistica
Vol. 14, 2012

The Factorial Asymmetric Multiplicative Error Model: preliminary results

Edoardo Otranto
Dipartimento di Scienze Cognitive e della Formazione and CRENoS,
University of Messina
E-mail: eotranto@unime.it

Summary: We consider the problem to modelize the volatility of a certain market including the transmission effects of other markets. We extend the Multiplicative Error Model, which is able to capture the dynamics of the volatility without resorting to logarithms of the time series, hypothesizing that the conditional mean can be decomposed into the sum of one factor, representing the *proper* volatility of the time series analyzed, and other factors, each one representing the volatility *transmitted* from another market. Each factor follows a proper dynamics with elements that can be usefully interpreted. This particular factorization provides the possibility to establish, for each time, the contribution of each market to the global volatility of the market object of the analysis.

Keywords: MEM, Realized volatility, Volatility transmission.

1. Introduction

The increasing degree of financial integration in terms of international investments and financial movements across borders has provided frequent cases of volatility transmission among markets (spillover effects, financial contagion, comovements, etc.). In our view, an important aspect to be considered in this framework is the fact that the volatility transmission is not constant along the time, but depends on the particular period considered; it is likely that the so-called dominant markets (e.g. the USA market) show very frequently their influence in the volatility of the other markets, whereas other markets show their effects only in particular periods and in correspondence of particular turmoils (a clear example is represented from the Greece crises in 2011-12). In modeling volatility, an important task is to capture the effect of the volatility of other markets in time-varying terms.

In this paper we propose an extension of the Multiplicative Error Model (MEM hereafter) of Engle (2002) to include this characteristic, but working in a univariate framework. The MEM approach is particularly interesting because, maintaining the simple structure of the GARCH model, is able to modelize non negative observations without the use of log transformation and to provide conditional expectations of the variables studied and not the expectations of the logarithms. The MEM structure is obtained from the product of two factors, one representing the mean level of the volatility and the other a positive disturbance. Our extension considers the possibility that the first factor can be decomposed into the sum of several sub-factors, the first one representing the volatility proper of the market and the others interpreted as the volatility transmitted from the other markets. Each sub-factor follows a sort of Threshold GARCH model (Zakoïan, 1994) and each part of the corresponding equation can be usefully interpreted, distinguishing among effects due to the recent information, persistence effect of the transmitted volatility, effects due to negative returns in the other markets. We call this model the Factorial Asymmetric MEM (FAMEM) and, as a final result, it is able to evaluate the presence and the weight of the volatility transmission from each market and in what measure it influences the studied market.

This reduced version of the paper is organized as follows: next section will describe the FAMEM, whereas Section 3 is devoted to some final remarks.

2. The Factorial Asymmetric MEM

Let $\mathbf{z}_t = (y_t, \mathbf{x}_t)'$ a $(n+1) \times 1$ vector of variables, each one representing the volatility relative to a certain financial market; in particular y_t represents the volatility at time t of the market to be analyzed, whereas $\mathbf{x}_t$ contains the other n variables. We hypothesize that the volatility y_t can be decomposed into the product of two factors: μ_t and a non negative disturbance ε_t with mean, conditional on the information at time $t-1$ (call it $\mathbf{\Psi}_{t-1}$), equal to 1. As a consequence, μ_t can be interpreted as the conditional mean of y_t. Similarly to Engle and Gallo (2006), we hypothesize that $\varepsilon_t | \mathbf{\Psi}_{t-1}$ follows a Gamma distribution with coefficients a and $(1/a)$. We propose to decompose the factor μ_t in $n+1$ sub-factors:

$$\mu_t = \zeta_t + \sum_{i=1}^{n} \xi_{i,t} \tag{1}$$

where ζ_t represents the *proper* volatility of the analyzed market, due to its proper dynamics and internal shocks, whereas $\xi_{i,t}$ represents the part of the volatility due to the volatility transmission from the $i-th$ market with volatility x_{it} included in $\mathbf{x_t}$. We suppose that both ζ_t and each $\xi_{i,t}$ follow a Threshold GARCH (Zakoïan, 1994) type dynamics.

Resuming, the model we propose, which we call Factorial Asymmetric MEM (FAMEM),

is characterized from the following set of equations:

$$\begin{array}{l} y_t = \mu_t \varepsilon_t \qquad \varepsilon_t | \mathbf{\Psi}_{t-1} \sim Gamma(a, 1/a) \text{ for each } t \\ \mu_t = \zeta_t + \sum_{i=1}^{n} \xi_{i,t} \\ \zeta_t = \omega + \sum_{h=1}^{p_0} \alpha_{0,h} y_{t-h} + \sum_{j=1}^{q_0} \beta_{0,j} \zeta_{t-j} + \delta_0 D_{0,t-1} y_{t-1} \\ \xi_{i,t} = \sum_{h=1}^{p_i} \alpha_{i,h} x_{i,t-h} + \sum_{j=1}^{q_i} \beta_{i,j} \xi_{i,t-j} + \delta_i D_{i,t-1} x_{i,t-1} \end{array} \tag{2}$$

where $D_{r,t}$ is a dummy variable assuming value 1 when the return of the corresponding market (with volatility y_t for $r = 0$, $x_{r,t}$ for $r = 1, \ldots, n$, respectively) is negative, 0 otherwise. Model (2) does not present particular estimation problems; it is possible to explicit the likelihood function and to maximize it. On the other side, this simple extension provides a lot of information. Firstly, it is possible to calculate the percentage of the explained volatility due to the transmission from other markets; it is given by:

$$tv_t = \frac{\sum_{i=1}^{n} \xi_{i,t}}{\mu_t} = 1 - \frac{\zeta_t}{\mu_t} \tag{3}$$

where $\frac{\zeta_t}{\mu_t}$ is the fraction of proper volatility. Moreover, it is possible to estimate the contribution of each market i (for each market i) to the volatility y_t by:

$$tv_{i,t} = \frac{\xi_{i,t}}{\mu_t} \tag{4}$$

Again, each element of the ζ_t and $\xi_{i,t}$ equations in (2) can have a proper interpretation:

- $\sum_{h=1}^{p_0} \alpha_{0,h} y_{t-h}$ is the part of the proper volatility due to the recent information about the volatility of the analyzed market;
- $\sum_{h=1}^{p_i} \alpha_{i,h} x_{i,t-h}$ is the part of the transmitted volatility due to the recent information about the volatility of market i;
- $\sum_{j=1}^{q_0} \beta_{0,j} \zeta_{t-j}$ is the inertial component of the proper volatility;
- $\sum_{j=1}^{q_i} \beta_{i,j} \xi_{i,t-j}$ is the inertial component of the transmitted volatility from market i;
- $\delta_0 D_{0,t-1} y_{t-1}$ is the effect due to negative returns in the analyzed market;
- $\delta_i D_{i,t-1} x_{i,t-1}$ is the effect due to negative returns in the market i;

3. Concluding Remarks

From a computational point of view, the FAMEM is very simple to be estimated and, using a general-to-specific specification, we are able to obtain feasible models. An advantage is the possibility to work in a univariate framework, reducing the number of

coefficients to be estimated, but, at the same time, obtaining very good results in terms of goodness-of-fitting and forecasting performance; we have experimented it on four volatility indices,[1] obtaining a better performance with respect to alternative models in three of the four cases studied, included a multivariate AMEM (Cipollini et al., 2006).

Formally, an extension of model (2) to the multivariate case is possible, but it would imply $(n+1)$ equations of the ζ_t type and $(n+1)n$ equations type ξ_t, and the coefficients of a multivariate Gamma (or other multivariate distributions with positive support) making the model unfeasible. The alternative would be to consider common factors for the $(n+1)$ variables, as in Engle et al. (1990), but we will lose the possible interpretation of the factors, described at the end of Section 2.

Acknowledgements: Financial support from Italian MIUR under Grant 20087Z4BMK_002 is gratefully acknowledged.

References

Cipollini F., Engle R.F., Gallo G. M. (2006), Vector multiplicative error models: Representation and inference, *Working Paper Series* n. 12690, National Bureau of Economic Research.

Engle R.F. (2002), New frontiers for ARCH models, *Journal of Applied Econometrics*, 17, 425–446.

Engle R.F., Gallo G.M. (2006), A multiple indicators model for volatility using intra-daily data, *Journal of Econometrics*, 131, 3–27.

Engle R.F., Ng V.K., Rothschild M. (1990), Asset pricing with a factor ARCH covariance structure: empirical estimates for treasury bills, *Journal of Econometrics*, 45, 213–238.

Zakoïan J.M. (1994), Threshold heteroskedastic models, *Journal of Economic Dynamic and Control*, 18, 931–955.

[1] To save space, we have not included the illustrative application, which is available on request.

Quaderni di Statistica
Vol. 14, 2012

Copula component analysis for dependence modelling

Paola Palmitesta
Department of Historical, Law, Political and Social Sciences
University of Siena
E-mail: paola.palmitesta@unisi.it

Corrado Provasi
Department of Statistical Sciences, University of Padua
E-mail: provasi@stat.unipd.it

Summary: A copula function can be employed to decompose the information content of a multivariate distribution into marginal and dependence components, with the latter quantified by the mutual information. From this statement, it is possible to state that a link between information and copula theories is valid. On the basis of these results, in the paper we show as it is possibile to use the independent component analysis to estimate the mutual information of a multivariate random sample and, then, to select the model of copula which better interprets the dependence in sample data.

Keywords: Copula functions, Independent component analysis, Mutual information.

1. Introduction

Casalsaverini and Vicente (2009) have shown that a copula function (Nelsen, 2006) can be employed to decompose the information content of a multivariate distribution into marginal and dependence components, with the latter quantified by the mutual information. As a matter of fact, in this contest the mutual information provides an upper bound for the asymptotic empirical log-likelihood of a copula. From this statement, we can state that a link between information and copula theories is valid.

On the basis of these results, in this paper we show as it is possible to use independent component analysis (Hyvärinen, Karhunen and Oja, 2001) to estimate the mutual information of a multivariate random sample and, then, to select the copula function

which better interprets the dependence in sample data. Therefore, we consider a procedure to investigate inferential aspects of the model proposed.

2. Copulas and mutual information

A d-dimensional copula C is a distribution function (df) on $[0,1]^d$ with standard uniform marginal distributions. Sklar's Theorem (see for example Nelsen (2006)) states that every df F with margins $F_1, \ldots, F_d$ can be written as

$$F(\boldsymbol{x}) = C(F_1(x_1), \ldots, F_d(x_d)), \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)'$, for some copula C, which is uniquely determined on $[0,1]^d$ for distributions F with absolutely continuous margins. Conversely any copula C may be used to join any collection of univariate dfs $F_1, \ldots, F_d$ using (1) to create a multivariate df F with margins $F_1, \ldots, F_d$. Here we concentrate exclusively on random vectors $\boldsymbol{X} = (X_1, \ldots, X_d)'$ whose marginal dfs are continuous and strictly increasing. In this case the so-called copula C of their joint df may be extracted from (1) by evaluating

$$C(\boldsymbol{u}) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)),$$

where $\boldsymbol{u} = (u_1, \ldots, u_d)'$ and the F^{-1} are the quantile functions of the margins. Introducing the copula density as $c(\boldsymbol{u}) = \frac{\partial^d C(\boldsymbol{u})}{\partial u_1, \ldots, \partial u_d}$, we can decompose the joint probability density as

$$f(\boldsymbol{x}) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{i=1}^{d} f_i(x_i), \tag{2}$$

where f are the density functions of the margins.

A convenient way to quantify statistical dependencies is by evaluating the mutual information of a copula. We recall here that the mutual information for a multivariable $\boldsymbol{X}$ with density function f is

$$I(\boldsymbol{x}) = \int f(\boldsymbol{x}) \log \frac{f(\boldsymbol{x})}{\prod_i f(x_i)} d\boldsymbol{x}.$$

Using (2) above, by performing the change of variables $u_i = F_i(x_i)$, $i = 1, \ldots, d$, this can be re-expressed as (Ma and Sun, 2011)

$$I(\boldsymbol{x}) = \int_{[0,1]^d} c(\boldsymbol{u}) \log c(\boldsymbol{u}) d\boldsymbol{u} = -H_c(x),$$

where H_c is the differential entropy associated with the c distribution. The mutual information can be regarded as a distance to statistical independence in the space of distributions measured by the relative entropy between the actual joint distribution and the product of marginals.

3. Copula component analysis

In independent component analysis (ICA), we observe a random vector $\boldsymbol{X} \in \mathbb{R}^d$ which is assumed to arise from a linear mixing of a latent random source vector $\boldsymbol{S} \in \mathbb{R}^d$,

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S},$$

where the underlying factors or sources, S_i, $i = 1, \ldots, d$, of $\boldsymbol{S}$ are statistically independent and $\boldsymbol{A}$ is the invertible $d \times d$ *mixing* matrix. ICA can be seen also as a linear transformation to a new set of variables $\boldsymbol{S} = \boldsymbol{W}\boldsymbol{X}$. The *demixing* matrix $\boldsymbol{W}$ is the inverse of the $\boldsymbol{A}$ matrix.

Statistical independence of sources means that the joint probability density of $\boldsymbol{X}$ and $\boldsymbol{S}$ can be factorized as

$$g(\boldsymbol{x}) = g(\boldsymbol{A}\boldsymbol{s}) = |\det(\boldsymbol{W})| g(\boldsymbol{s}), \quad g(\boldsymbol{s}) = \prod_{i=1}^{d} g_i(s_i).$$

Then, given a random vector $\boldsymbol{X}$ with df F determined with some copula C, the distance between g and f in a sense of Kullback-Liebler (K-L) divergence (cf. Cover and Thomas, 1991) can be represented as

$$\begin{aligned} K(g, f) &= \int g(\boldsymbol{x}) \log \frac{g(\boldsymbol{x})}{f(\boldsymbol{x})} d\boldsymbol{x} \\ &= \int g(\boldsymbol{x}) \log \frac{g(\boldsymbol{x})}{\prod_{i=1}^{d} g_i(x_i)} d\boldsymbol{x} - \int g(\boldsymbol{x}) \log c(\boldsymbol{u}) d\boldsymbol{x}, \end{aligned} \tag{3}$$

with the meaning of symbols seen before. The first term of (3) corresponds to the K-L divergence between g and ICA model and the second term corresponds to entropy of copula C.

4. Model Selection of Copula

The selection of copulas is an important aspect of dependence modelling. In many applications, we don't know the copula function which determines the df F of a multivariate distribution $\boldsymbol{X}$. Therefore it can be necessary to choose among the many copula classes the copula which better interprets sampling observations (see for example Manner, 2007). In this sense, a selection procedure can be obtained using a copula component analysis.

Given a data set $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ independently sampled from an unknown joint density f, the best approximation f_θ within a manifold $\mathcal{F}$, parameterized by θ, can be found by minimizing a sample estimate of the K-L divergence (3). Ma and Sun (2007) have

shown that, if a dependency structure is presented by copula function C, the objective function to be minimized can be written as

$$K(g, f; \boldsymbol{W}), \theta) = I(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{W}) + H_c(\boldsymbol{u}; \boldsymbol{W}, \theta),$$

whose optimization implies the minimization of the mutual information $I(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n; \boldsymbol{W})$ and the maximization of entropy $H_c(\boldsymbol{u}; \boldsymbol{W}, \theta)$. Therefore, the estimation procedure can be done in two steps: in the first step we solve for $\boldsymbol{W}$ through minimization of mutual information, while in the second step we determine θ in $C_\theta(\boldsymbol{u})$ so that entropy of copula is maximized. It is easy to prove that minimization of $K(g, f; \boldsymbol{W}, \theta)$ is equal to maximization of the likelihood function of the parametric model of copula subject to ICA.

In conclusion, the copula independent analysis can be a statistical approach to select a copula among many copula candidates; moreover, a bootstrap procedure can be used to verify the compliance of the copula respect to sample data.

Acknowledgements: This work is supported by a grant from MIUR (code 2008WKHJPK-002 PRIN2008).

References

Hyvärinen A., Karhunen J., Oja E. (2001), *Independent Component Analysis*, Wiley, New York.

Calsaverini R.S., Vicente, R. (2009), An information-theoretic approach to statistical dependence: Copula information, *Europhys. Lett.*, 88, 68003.

Cover T., Thomas J. (1991), *Elements of Information Theory*, Wiley, New York.

Ma J., Sun Z. (2007), Copula component analysis, in: Davies M., James C., Abdallah S., Plumbley M. (eds.), *Independent Component Analysis and Signal Separation, Lecture Notes in Computer Science*, 4666, 73–80, Springer, Berlin/Heidelberg.

Ma J., Sun Z. (2011), Mutual information is copula entropy, *Tsinghua Science & Technology*, 16, 51–54.

Manner H. (2007), Estimation and model selection of copulas with an application to exchange rates, *Research memorandum*, METEOR, Maastricht research school of Economics of TEchnology and ORganizations.

Nelsen R.B. (2006), *An Introduction to Copulas*, Springer, Berlin/Heidelberg.

Quaderni di Statistica
Vol. 14, 2012

Analysis of the covariance structure in manufactured parts

Giovanni Pistone
Collegio Carlo Alberto Moncalieri
E-mail: giovanni.pistone@carloalberto.org

Suela Ruffa
DIGEP Politecnico di Torino
E-mail: suela.ruffa@polito.it

Grazia Vicario
DISMA Politecnico di Torino
E-mail: grazia.vicario@polito.it

Summary: We discuss on a technological case study the use of estimators of covariance in a Kriging model to assess tolerances in a mechanical production process. It is a tricky issue to decide which class of covariance models is the best among the available options. The variogram is very informative about the spatial dependence and its estimation is essential in the spatial prediction, because it allows the computation of the weights assigned to the measured points in the Kriging prediction. In the paper we use various literature estimators for variograms for evaluating their estimation properties.

Keywords: Coordinate Measuring Machines, Kriging, Variogram, Spatial Correlation.

1. Introduction

The phenomenon considered in this paper is the surface error of manufactured parts, whose tolerances are assessed by Coordinate Measuring Machines (CMM), massively used in industry to check dimensions and shape of manufactured parts, because they allow to measure cartesian coordinates $(x; y; z)$ automatically and rapidly at a large enough number of points on the surface of interest. In our case study of variogram estimation in manufactured parts, we consider a planar grinded surface inspected in a dense regular grid with the aim of identifying a model of the covariance structure for the full surface.

In the paper we consider the Kriging model because of its recognized predictive

capability. The original Krige's idea of a positive correlation that decreases with increasing distance between the experimental points requires an accurate investigation of the covariance structure and its modelization. The variogram is very informative about the spatial dependence and it is favored by the researchers, mostly geo-statisticians, in the choice of the correlation function. The estimation of the variogram is essential in the spatial prediction, because it allows the computation of the weights assigned to the measured points in the Kriging prediction. In the paper we use different literature estimators for variograms for evaluating their estimation properties. Contrary to the common engineering belief, the variograms referring to three planar surfaces with different machine precisions give evidence of non-isotropy and of nugget effect. These features are interpreted both as technological signatures of the production process and systematic errors of the CMM measurement process. We refer to the recent monographs by Cressie (1993) for further details on Kriging.

2. Variogram and Nugget estimation

The Krige's idea of a positive correlation that decreases with increasing distance between the experimental points may be estimated by Matheron's variogram (G. Matheron, 1962 [2]). It is a statistical index describing the degree of spatial dependence of a stochastic process and it is defined as the variance of the difference between field values:

$$2\gamma(\boldsymbol{x}_1 - \boldsymbol{x}_2) = \text{var}[Z(\boldsymbol{x}_1) - Z(\boldsymbol{x}_2)] \text{ for all } \boldsymbol{x}_1, \boldsymbol{x}_2 \text{in } D \tag{1}$$

In Matheron's theory, the variogram is assumed to be null and continuous at 0; if it is not the case, i.e. $\lim_{h \to 0} \gamma(h) = c_0$, there is a *nugget*, mostly due to measurement errors (repeating the measurements a number of time, the measured values tend to fluctuate around the true value). Moreover, if is $C(\boldsymbol{h}) = \text{Cov}(Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2))$, it is $C(\boldsymbol{h}) = \gamma(\infty) - \gamma(\boldsymbol{h})$. Therefore, the covariance exists if, and only if,$\gamma(\infty)$, named *sill*, is finite.

There are various estimators for the variogram in literature; correctness, consistency, robustness and other desirable properties of the estimators are not satisfied by all the considered estimators. Matheron's original estimator is the empirical estimator. It has poor robustness properties because it is badly affected by possible outliers.

In literature there are more robust approaches to the variogram estimation. Cressie and Hawkins (1980) proposed two robust estimators: in this paper we consider the one based on the median $C_{med}(\boldsymbol{h})$:

$$C_{med}(\boldsymbol{h}) = \frac{1}{2B_h} = \left(med(|Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j)|^{\frac{1}{2}})\right)^4 \tag{2}$$

where $med(\cdot)$ is the median index, B_h is a correction factor for the bias when the field is Gaussian (asymptotically is B_h = 0.457).

The estimator $C_{diff}(\boldsymbol{h})$ proposed by Haslett (1997) is the half sample variance of the differences $Z(\boldsymbol{x}_i - Z(\boldsymbol{x}_j))$; $C_{diff}(\boldsymbol{h})$ proves to have good performances if the field is non-stationary.

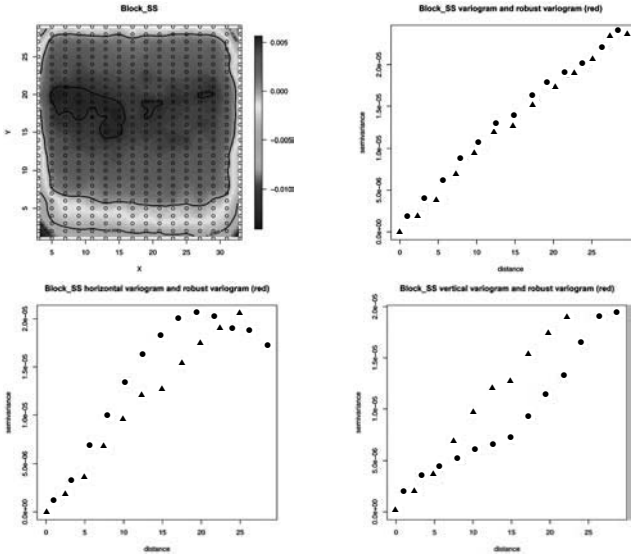


Figure 1. Image of the values, variogram and robust variogram, x-variogram and robust variogram, y-variogram and robust variogram in the case of the top grinded surface of a Johansson gauge

The last estimator $C_{Q_N(h)}(\boldsymbol{h})$ we considered is proposed by Genton (1998). We avoid to write its analytical expression (the computation is implemented in Software R), but its high robustness against the outliers is recognized.

A different option for discussing anisotropy is to consider a parametric model, typically exponential, for the covariance and compare the estimated values of the scale parameters in the two coordinate directions.

3. Tipical findings from the case study

Figure 1 shows the type of results obtained with the non parametric estimation of the variance in the case of the top grinded surface of a Johansson gauge. In this edition of the paper, we present results and outputs relating to one single planar surface because this is a short version. Analogous results are available for other two planar surfaces (a milled surfaces of a clamp for industrial air cushion guide and the lapped surface of a Johansson gauge), coming to different but peculiar conclusions.

In the present case study, the robust estimator does not give any information different from the Matheron's one, suggesting the absence or non relevance of outliers. Opposite, the directional Matheron estimator shows evidence of a considerable anisotropy of the production process. This finding is contrary to the common engineering belief, ac-

cording to it the isotropy is a common assumption. This features are interpreted both as technological signatures of the production process and systematic errors of the CMM measurement process, depending on the planar surface considered.

In a parametric study using the R package DiceKriging developed by Roustan (2009) we have estimated covariance of the form

$$g_{\mathrm{E}}(\boldsymbol{h}) = \exp\left(-\frac{|h_1|}{\theta_1} - \frac{|h_2|}{\theta_2}\right) \quad \text{Exponential}$$

$$g_{\mathrm{PE}}(\boldsymbol{h}) = \exp\left(-\left(\frac{|h_1|}{\theta_1}\right)^{p_1} - \left(\frac{|h_2|}{\theta_2}\right)^{p_2}\right) \quad \text{Power Exponential}$$

With the same data as used in Figure 1 we found

	θ_1	θ_2	p_1	p_2	Field Var	Nugget
exponential	55.9463	47.5896			$2.7e-05$	$9.6e-08$
power exponential	32.4741	36.0499	1.6745	1.625	$3.0e-4$	$1.2e-07$

The nugget effect can be neglected if compared with the estimated field variance.

The difference between the estimated parameters confirms the anisotropy of the production process already highlighted by directional variograms. Nugget can be caused by measurements errors.

References

Cressie N.A.C. (1993), *Statistics for spatial data*, Wiley, New York.

Cressie N., Hawkins, M. (1980), Robust estimation of the variogram I, *Mathematical Geology*, 12, 115–125.

Genton M. G. (1998), Highly robust variogram estimator, *Mathematical Geology*, 30, 213–221.

Haslett J. (1997), On sample variogram and the sample autocovariance for non-stationary time series, *The Statistician*, 46, 475–485.

Rousseew P.J., Croux. C. (1993), Alternatives of the median absolute deviation, *Journal of American Statistics Association*, 88, 1273–1283.

Roustant O., Ginsbourger D., Deville Y. (2009), The DiceKriging package: kriging-based metamodeling and optimization for computer experiments, The R User Conference 2009 July 8-10, Agrocampus-Ouest, Rennes, France.

Quaderni di Statistica
Vol. 14, 2012

A mixture model for predicting football teams' performance

Silvia Polettini
Dipartimento di Scienze e Biotecnologie Medico-chirurgiche
Sapienza Università di Roma
E-mail: silvia.polettini@uniroma1.it

Francesco De Icco
IMM-CNR, Napoli
E-mail: francesco.deicco@cnr.it

Summary: Predicting the outcome of a football match is of primary importance in fixed odds betting markets. Dixon and Coles (1997) have proposed a simple bivariate Poisson mixed effects model generalizing Maher's (1982), showing that the odds of specific match results can be better predicted by their model. Baio and Blangiardo (2010) propose a Bayesian hierarchical model with the additional feature that the teams are grouped into different classes according to their attack/defense performance, leading to a mixture model. We propose a generalization of the approach above, based on the introduction of a Dirichlet Process (DP) Prior on the team performance, allowing for a random number of groups; the application refers to data from the 2010-2011 Italian championship.

Keywords: Dirichlet process, Semiparametric mixed effects model, Football betting.

1. Data and model

The league teams amount to a total of 20, playing each other twice in a season (one home, one away). For the $g-$th match we denote by $Y_{g,1}$ and $Y_{g,2}$ the number of goals scored by the home and away team, respectively. We make the standard assumption that conditional to θ_1, θ_2, the observables have independent Poisson distributions with parameters $\theta_{1,g}\, \theta_{2,g}$, respectively, and parameterize the mean scores of the home and away team, respectively, through home, attack and defense parameters as follows:

$$log(\theta_{1,g}) = home + att_{h(g)} + def_{a(g)}; \quad log(\theta_{2,g}) = att_{a(g)} + def_{h(g)}.$$

We model the attack and defense parameters by introducing for each a Dirichlet Process (DP) prior with the same base measure F_0 and precision α; the Sethuraman (1994) representation of the DP implies that realizations from the DP are infinite mixtures of point masses:

$$f(\cdot) = \sum_{k=1}^{\infty} p_k I(\theta_k), \quad \theta_k \overset{i.i.d.}{\sim} F_0.$$

The base measure for each parameter can be chosen to be a Normal (indeed we assume F_0 to be $N_(\mu, \sigma^2)$), or a non-central Student's *t* distributions as in Baio and Blangiardo (2010). Choice of a DP prior allows to accommodate outliers (and therefore avoid the shrinkage effect) through a model generalization. A further generalization would be to use a mixture of Dirichlet processes (Antoniak, 1974), so that the above mixture representation can be written as

$$f(\cdot) = \sum_{k=1}^{\infty} p_k h(\cdot|\theta_k), \quad \theta_k \overset{i.i.d.}{\sim} F_0.$$

Baio and Blangiardo (2003) use independent priors for the teams' parameters. Since stronger teams may exhibit strong attack and defense parameters, whereas for other teams strong attack could be accompanied by weak defense, we also allow for possible dependence by specifying a mixture joint distribution for the pair (att, def). To this aim we adopt a single DP prior with bivariate base measure G_0. The parameters of the base measure are in turn modelled through normal and inverse gamma/Wishart "flat" distributions, while α is assigned a Gamma prior distribution.

2. Results

Data from the Italian 2010-2011 season league were used to assess the approach described; the teams' attack/defence performances are summarized in Figure 1. We also compare our framework with the one by Baio and Blangiardo (2010). The model fit seams reasonably good; for brevity we restrict attention to the out-of-sample predictive performance of the model in forecasting results for the 26th round of the league based on data from all the previous matches.

We predict correctly 7 out of the next 10 matches; for the model by Baio and Blangiardo this number is 5; comparing with bookmakers' odds, we see that, apart from the match Napoli-Catania, there is general agreement, and certain outcome probabilities are predicted more closely by our model.

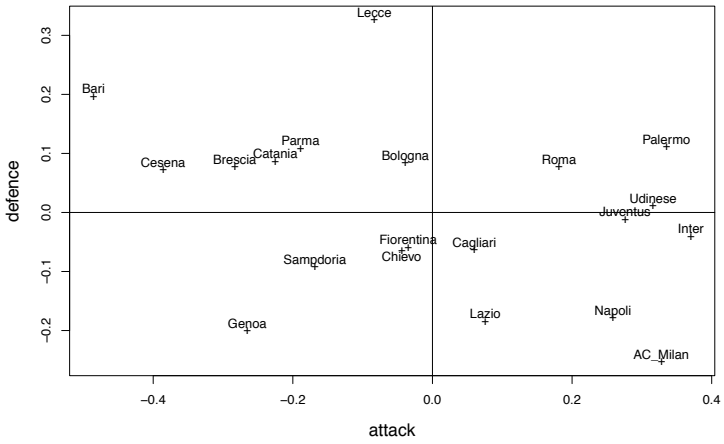


Figure 1. Plot of attack vs defense posterior means – independent mixtures of Dirichlet processes

References

Antoniak C. (1974), Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics*, 2, 1152–1174.

Baio G., Blangiardo M. (2010), Bayesian hierarchical model for the prediction of football results, *Journal of Applied Statistics*, 37, 253–264.

Dixon M.J., Coles, S.G. (1997), Modeling association football scores and inefficiencies in the football betting market, *Applied Statistics*, 46, 265–280.

Karlis D., Ntzoufras I. (2003), Analysis of sports data using bivariate Poisson models, *Journal of the Royal Statistical Association, D*, 52, 381–393.

Maher M.J. (1982), Modelling association football scores, *Statistica Neerlandica*, 36, 109–118.

Sethuraman J. (1994), A constructive definition of Dirichlet priors, *Statistica Sinica*, 4, 639–650.

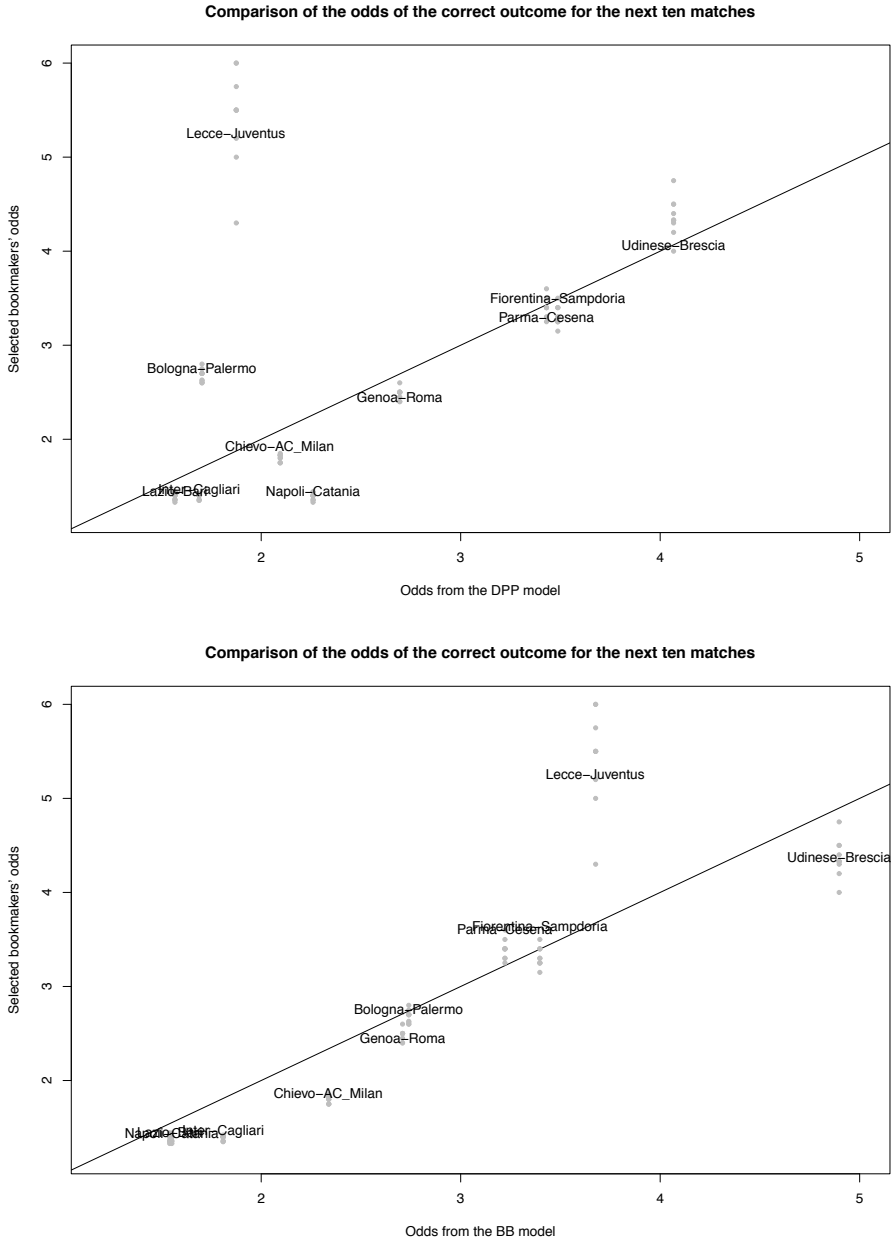


Figure 2. Comparison of model based odds vs bookmakers' odds – independent mixtures of Dirichlet processes (upper panel) vs Baio and Blangiardo model (lower panel)

Quaderni di Statistica
Vol. 14, 2012

Hierarchical Bayesian models for the estimation of correlated effects in multilevel data: a simulation study to assess model performance

Giulia Roli Paola Monari
Department of Statistical Sciences, University of Bologna
E-mail: g.roli@unibo.it, paola.monari@unibo.it

Summary: In this paper, we aim at assessing hierarchical Bayesian modeling for the analysis of multiple exposures and highly correlated effects in a multilevel setting. We exploit an artificial data set to apply our method and show the gains in the final estimates of the crucial parameters. As a motivating example to simulate data, we consider a real prospective cohort study designed to investigate the association of dietary exposures with the occurrence of colon-rectum cancer in a multilevel framework, where, e.g., individuals have been enrolled from different countries or cities. We rely on the presence of some additional information suitable to mediate the final effects of the exposures and to be arranged in a level-2 regression to model similarities among the parameters of interest (e.g., data on the nutrient compositions for each dietary item).

Keywords: Hierarchical Bayesian modelling, Correlated effects, Multilevel data.

1. Introduction

The estimation of multiple effects often faces problems of some sort of complications that need to be somehow controlled during the analysis. We consider two kinds of such complications. The first one concerns the structure of the data and occurs whenever units are nested into higher level units involving their own variability and a dependence among the related observations. This dependence is "neither accidental nor ignorable" (Goldstein, 1999) and the risks of drawing wrong conclusions are high if the clustering of the data is disregarded. The joint analysis of multiple exposures gives rise to the second complication. Indeed, many studies involve a set of potential effects to be compared, especially in epidemiologic field, and, as a result, face problems of multiple inference. When a conventional analysis is carried out, these are revealed by failures

in the convergence of the estimation process or by implausible large and unstable estimates, especially when the samples are small and sparse (Witte *et al*., 1994). The main reason is that these effects are often correlated. Therefore, we need to take into account for a covariance structure among them to reduce the random errors in the estimates.

Both these complications have been tackled separately in various applications and simulations by using hierarchical modeling (see for example, Diex-Roux, 2000; Witte *et al*., 1994). Although developed separately and for different purposes, hierarchical models for correlated effects and for multilevel data have important communalities, which can be strengthened especially when a Bayesian perspective is adopted (Gelman *et al*., 2003). In a previous work, we have extended the hierarchical Bayesian approach for the analysis of multiple exposures to a multilevel setting with an application to real data (Roli and Monari, 2011). Here, we aim at assessing the model on artificial data. As a motivating example to simulate data, we consider a real prospective cohort study designed to investigate the association of dietary exposures with the occurrence of colon-rectum cancer in multilevel data (e.g., individuals enrolled from different countries or cities). We rely on the presence of some additional information suitable to mediate the final effects of the exposures and to be arranged in a level-2 regression to model similarities among the parameters of interest (e.g., data on the nutrient compositions for each dietary item). Using this artificial data set, we apply our method to show and measure the gains in the final estimates of the crucial parameters with respect to the conventional analysis results.

2. Modelling framework

We consider J groups, commonly defined by geographical areas. For each group j, we have the total number of individuals N_j and the presence/absence of a disease is denoted by the individual indicator y_{ij} ($y_{ij} = 1$ for cases, $y_{ij} = 0$ for control units). We wish to model the number of diseases cases in terms of K explanatory variables, or exposures, denoted by X_k for each exposure k, further controlling for the effects of P potential confounders (such as age, sex and smoking status of individuals), denoted by W_p for each confounder p. We assume that the data are generated through the following underlying model. Individual i in group j, with exposures in $(x_{ij1}, \ldots, x_{ijK})$ and counfounders in $(w_{ij1}, \ldots, w_{ijQ})$, experiences the binary outcome y_{ij} with probability p_{ij}, where:

$$logit(p_{ij}) = \alpha_j + \sum_{k=1}^{K} \beta_{jk} x_{ijk} + \sum_{p=1}^{P} \gamma_p w_{ijp}. \tag{1}$$

α_j represents the group-specific logit-baseline risk, which is splitted into a common effect ψ_0 and a residual term u_j, whose values are assumed to be independent and normally distributed with null means and common variance σ_u^2. The effects of confounders are reasonably assumed to be the same in all the groups and, thus, simply denoted by γ_p.

As far as the exposures effects β_{jk} are concerned, they represent the key objective of the investigation and we assume they vary across the groups. In order to tackle the problem of interactions among the multiple exposures and the correlation among their effetcs, we assume that some kinds of Q group-specific prior or level-2 data are available (denoted by z_{jkq} for each group j exposure k and level-2 covariate q). These are arranged to form the exposures' coefficients through a level-2 regression model:

$$\beta_{jk} = \pi_{k0} + \sum_{q=1}^{Q} \pi_q z_{jkq} + \delta_{jk} \tag{2}$$

where π_q are the effects of such prior information on the exposures (and on the disease), which are assumed to be common to all the exposures; π_{k0} is the intercept reflecting our knowledge about any residual effects of the exposure k due to prior information not included in the second-stage model; δ_{jk} are the residuals, which are assumed to hold the simple hypothesis of independence and normal distribution with null means and constant variances, denoted by σ^2_δ, and to be further independent on u_j.

Under a Bayesian framework, the specification of the prior distributions for γ_p, ψ_0, σ^2_u, π_q, π_{k0} and σ^2_δ is needed and the choice can be problematic. As a first attempt, we assign flat and conjugate prior distributions and the model will be assessed with a simulation study described in the next section.

3. Simulation study

We generate an artificial dataset basing on a real prospective cohort study designed to investigate the association of dietary exposures with the occurrence of colon-rectum cancer with individuals enrolled from different centers and controlling for potential confounders, such as smoking status and age. The prior or level-2 information on the exposures are data on the nutrient compositions for each dietary item and center. We consider J=30 centers of enrollment of N_j=400 individuals, K=8 dietary exposures, P=2 confounders, one binary (such as smoking status) and one continuous (such as age), and Q=4 nutrients. We take γ_1=$log(1.15)$ as the effect of the binary confounder and γ_2=$log(1.01)$ as that of the continuous one. We vary the center-level proportion of individuals exposed to the binary confounder between groups, as $C \times l/3$, l=$1, 2, 3$, by fixing the basic case to be C=0.2. The continuous confounder variable is generated from a Normal distribution with mean 58.4 and standard deviation 6.3. The logit-baseline risks α_j of disease for the J groups are chosen as 10 equally spaced quantiles of a normal distribution, with mean $logit(0.1)$ and standard deviation 0.2 (Jackson *et al.*, 2006). For each exposure k, we generate values within group j from distributions with group-specific parameters and by fixing the ratios of the between-center standard deviation to the within-center standard deviation to be always ≥ 0.6. We fix the level-2 nutrients effects to be $log(1.06)$, $log(1.05)$, $log(0.9)$ and $log(1.015)$ and the k-specific level-2 intercepts to be -0.05, if we expect the residual effects for item k to be preventive, 0.05

if we expect to be causative and 0 if we expected little or no residual effects (Witte *et al*., 1994). The compositions of each dietary item are generated, separately by nutrient, from normal distributions defined by k-specific parameters and by fixing the ratios of the between-dietary standard deviation to the within-dietary standard deviation to be always ≥ 60. The residuals δ_{jk} are generated from a normal distribution with null means and common standard deviation 0.01. As a result, the $J \times K$=240 coefficients β_{jk} vary from a minimum of -0.085 (OR=0.918) to a maximum of 0.204 (OR=1.226). Finally, to generate the outcome, a Bernoulli distribution with probability p_{ij} is used. The overall prevalence of the disease is close to 15%.

We fit the hierarchical model described above to these simulated data by Markov chain Monte Carlo method implemented in the WinBUGS software. Here, we present results on the target parameters, i.e. Odds Ratios (OR) of dietary items ($exp(\beta_{jk})$), in terms of percentage bias and coverage of nominal 95 per cent credible intervals. Coverages of credible intervals are reasonable with 82.5% of cases including the true parameter. Over the 30 centers and the 8 dietary exposures, the percentage biases (in absolute values) vary from a minimum of 0.011 to a maximum of 11.284 (with mean 1.603 and standard deviation 1.509), which are much smaller than those obtained from several conventional logistic regressions separately by centers of enrollment (mean $\gg 100$).

References

Diez-Roux A.V. (2000), Multilevel analysis in public health research, *Annual Review of Public Health*, 21, 171–192.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (2003), *Bayesian Data Analysis* (2nd ed.), Chapman and Hall/CRC, New York.

Goldstein H. (1999), *Multilevel statistical models*, Institute of education, multilevel models project, London.

Jackson C., Best N., Richardson S. (2006), Improving ecological inference using individual-level data, *Statistics in Medicine*, 25, 2136–2159.

Roli G., Monari P. (2011), Improving the estimation of multiple correlated dietary effects on colon-rectum cancer in multicentric studies: a hierarchical Bayesuan approach, *Statistica*, 71, 437–452.

Witte J., Greenland S., Haile R., Bird C. (1994), Hierarchical regression analysis applied to a study of multiple dietray exposures and breast cancer, *Epidemiology*, 5(6), 612–621.

Quaderni di Statistica
Vol. 14, 2012

Estimating psychometric reliability with one observation per subject

Hoben Thomas
Department of Psychology, The Pennsylvania State University, USA
E-mail: hxt@psu.edu

Summary: Test a sample of individuals with the same test on two occasions. Correlate the pairs of scores. The resulting r is called the reliability of the test. More than 100 years ago Charles Spearman proposed a random effects model for interpreting r a model which remains extensively used today. However in many settings obtaining pairs of scores from the same individuals is often expensive or cannot be achieved. A model in the Spearman spirit for estimating reliability based on *one* observation per subject is proposed.

Keywords: Reliability, Mixtures, Spearman model.

1. A largely intuitive introduction

In 1904 Spearman proposed what was probably the first latent variables random effects model, and it was proposed for interpreting test-retest r reliability coefficients. The model was formalized by Yule in 1908 (Spearman, 1910). The population version is

$$X_o = T + E_o$$

where only X_o is observed with the test administered on occasions $o = 1, 2$. $\mathrm{var}(T) = \sigma_T^2, \mathrm{E}(E_o) = 0, \mathrm{var}(E_o) = \sigma_E^2$ and E_1, E_2 and T are independent of each other. (Spearman assumed uncorrelated variables; independence is needed here.) Spearman's model implies the following correlational structure

$$\rho_{XX} \equiv \rho(X_1, X_2) = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

which is the model used for interpreting sample reliability values, r.

However, in many settings obtaining two suitable assessments is difficult, costly, or impossible. Can reliability be estimated with *one* observation per subject? The following

fictitious story illustrates how this may be achieved; a slight modification of Spearman's the model underlies the example.

Consider a country in which all citizens live in one of two cities: In one city all residents have true score IQs of 80 , while in the other their true score IQs are 120. The cities are of equal size, with $\sigma_E = 5$ in each city. Suppose there is available a (large) $2n$-sized random sample of citizens which have been tested once. 1) Pair-up the scores randomly forming n pairs. 2) Within each pair, subtract one test score from the other. If the pair of citizens have the same true score, their difference should be around zero. If they have different true scores, the differences should be around ± 40. Discard those pairs with absolute differences greater than 15, suggesting the elements of the pair have different true scores. 4) Correlate the remaining pairs. The resulting r estimates ρ_{XX}.

The distribution theory for a test-retest version of this story leads to a bivariate mixture distribution with two equally weighted bivariate components, one centered at $(80, 80)$ a second centered at coordinates $(120, 120)$. While each component has independent marginals, the mixture structure induces the correlation, with $\rho_{XX} = .94$. (No distribution assumptions are needed at this point.) So .94 is the target value to be estimated with a univariate sample.

As an illustrative simulation: Take a random sample of size 100 from a univariate two component normal mixture with weights 1/2: $(1/2)[n(\mu = 80, \sigma_E = 5) + n(\mu = 120, \sigma_E = 5)]$. Randomly pair the observations. Obtain the difference scores within each pair discarding pairs with absolute difference 15 or greater. Correlate the remaining pairs. Repeating the simulation ten times yielded r values from .93 to .96.

Although the example is overly simplified, it does illustrate a general strategy: A bivariate correlational problem is transformed to a univariate setting; information culled from the univariate setting is transformed back to the bivariate reliability setting.

2. A brief more formal development

While the focus is on using one observation per subject it is convenient to develop the model from a test-retest perspective, i.e., with the same individual being assessed twice, then recognize the adjustment that can be made when a different individual replaces the second within-subject test score.

View Spearman's model from a conditional perspective,

$$X_{io} = t_{io} + E_{io}, i = 1, 2, \ldots n; o = 1, 2$$

where the assumptions are as above, except that t_{io} is the fixed true score for individual i on occasion o. Spearman assumes the true scores $t_{i1} = t_{i2}$ match for all i. This assumption can be evaluated by assuming that the pair (X_{i1}, X_{i2}) is bivariate normal with independent marginals (because of the assumed independent of E_{i1} and E_{i2}). Let $D_i = X_{i1} - X_{i2}$. Note that D_i is normal in distribution and under Spearman $\mathrm{E}(D_i) = \delta_i = 0, \mathrm{var}(D_i) = 2\sigma_E^2 = \sigma^2$ so under Spearman's model $D_i \sim n(0, \sigma^2)$.

Suppose $t_{i1} \neq t_{i2}$, then $\mathrm{E}(D_i) = \delta_i \neq 0$. Omitting some details (See Thomas et al, 2011) if i's true scores do not match, δ_i will be either negative or positive, and its departure from zero will vary depending on the magnitude of the mismatch of the true scores. Thus collectively, δ_i can be less than greater than or equal to zero. Consequently, unconditionally, D_i is distributed as a three component normal mixture:

$$D_i \sim \underbrace{\pi_1\, n(0, \sigma^2)}_{\text{Comp.1, } \delta=0} + \pi_2\, n(-\mu_\delta, \sigma^2 + \sigma_\delta^2) + \pi_3\, n(\mu_\delta, \sigma^2 + \sigma_\delta^2)$$

The π_j are component weights, $0 < \pi_j < 1 = \sum_j \pi_j$. The extra source of variance in the second and third components arises because of the assumed random effects associated with the varying non-zero δ_i which have means $\pm\mu_\delta$ and variance σ_δ^2.

Critical is component 1, called the Spearman Component, because under Spearman, difference scores D_i should have mean zero. If this assumption holds, then it would be expected π_1 would be at or near one. In fact, in real data estimates of $\hat{\pi}_1$ are rarely near one, suggesting that failure of the matching true score assumption of Spearman's model is common. The goal of Thomas et al (2011) was to use those i with $\delta_i = 0$ as the basis for estimating reliability. This was achieved using the R mixture package `mixtools`, specifically the `normalmixEM` call, then using $P(\text{Component } 1|D_i = d_i) = w_i$ weights obtained at convergence to form a weighted r denoted r_w which is taken as Spearman reliability and thus estimates ρ_{XX}.

Consider now two different individuals i and j: $X_i = t_i + E_i, X_j = t_j + E_j$ with the same structural assumptions as above. In particular assume (X_i, X_j) are circular bivariate normal. The above theory applies equally as well with a slight change in notation, so for example, $D_{ij} = X_i - X_j, \mathrm{E}(D_{ij}) = \delta_{ij}$ and $\mathrm{var}(D_{ij}) = \sigma^2$. The added difficulty of course is knowing how to suitably pair-up different individuals.

The following algorithm cycles through two distinct procedures, 1) combinatorial optimization, and 2) EM mixture decomposition with the goal of finding the appropriate pairings X_i and $X_j, i = 1, 2, \ldots, n; j = 1, 2, \ldots, n$ that maximize π_1, the Spearman mixture component weight.

3. Computational algorithm

(1) Let vector $\mathbf{V}$ be a random sample of test scores of length $2n$, ordered from smallest to largest (the order statistics).
(2) Put the odd numbered order statistics in one vector $V_{(1)}, V_{(3)}, V_{(5)}, \ldots, V_{(2n-1)}$, the even numbers in another $V_{(2)}, V_{(4)}, V_{(6)}, \ldots, V_{(2n)}$. Relabel these vectors respectively as $\mathbf{X}$ and $\mathbf{Y}$.
(3) $\hat{\pi}_1 =$ `normalmixEM`$(\mathbf{X} - \mathbf{Y})$ # Decompose the difference scores
(4) For $(s = 1$ to $s = S)\{$
$shuffle : \mathbf{Y} \rightarrow \mathbf{Y}_s$ # permutes elements of $\mathbf{Y}$ to $\mathbf{Y}_s$
`normalmixEM`$(\mathbf{X} - \mathbf{Y}_s) = \hat{\pi}_{1s}$ # Spearman component weight for permutation s

If($\hat{\pi}_{1s} > \hat{\pi}_1$) replace $\mathbf{Y}$ with $\mathbf{Y}_s$ and replace $\hat{\pi}_1$ with $\hat{\pi}_{1s}$; otherwise retain both $\mathbf{Y}$ and $\hat{\pi}_1$}
(5) For the $\mathbf{Y}$ associated with $\hat{\pi}_1$, after S iterations compute $r_w(\mathbf{X}, \mathbf{Y})$ which is reliability.

Steps (1) and (2) are regarded as producing a reasonable starting condition for a correlational problem. The loop in Step (4) is a localized random search algorithm (Givens & Hoeting, 2005) with the function $shuffle$ producing a rearrangement of the elements of $\mathbf{Y}$.

4. Conclusion

The model does assume a certain "clumpiness" of data: Formally it requires the marginal distributions in the test-retest setting be more appropriately modeled by mixture distributions with normal components with constant error σ_E, for all individuals; further specification is not required. For example, neither the number of components nor their weights need be known. Said simply, the data required need to be in the spirit of the above "two cities" example. The usefulness of the approach can be evaluated with both simulations and with real data. Suppose n test-retest pairs of data are available. Then concatenating the paired data into a single data vector of length $2n$ the model should be able to approximate the test-retest r. This is illustrated by considering Kelly's (1927, p. 159) ABC paragraph meaning test-retest data administered to 36 school children.

References

Givens, G. H., & Hoeting, J. A. (2005), *Computational statistics*. New York: Wiley.

Kelly, T. L. (1927), *Interpretation of educational measurements*. New York: World Book Company.

Spearman, C. (1904), General intelligence objectively defined and measured. *American Journal of Psychology*, 5, 201-293.

Spearman, C. (1910), Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271-295.

Thomas, H., Lohaus, A., & Domsch, H. (2011), Stable unstable reliability theory. *British Journal of Mathematical and Statistical Psychology*. Article first published online : 2 February 2011, DOI: 10.1111/j.2044-8317.2010.02011.x

Quaderni di Statistica
Vol. 14, 2012

Ordinal longitudinal data analysis using multilevel Rasch model in the context of chemotherapy side effects

Maria Chiara Zanarotti
Department of Statistics, Catholic University of Milan
E-mail: chiara.zanarotti@unicatt.it

Laura Pagani
Department of Economics and Statistics, University of Udine
E-mail: laura.pagani@uniud.it

Summary: In this paper we consider panel data obtained by administering a set of items to a group of persons during time. Responses to each item are collected trough an ordinal scale (like a Likert scale, for example) and the goals of the analysis are: (a) to obtain a person and item measure on an interval scale that located each individual and each item into a unique continuous scale; (b) to study this person and item measures behaviour during the time intervals considered. Rasch model (RM) looks as a standard method to perform goal (a) because of RM capacity to concurrently estimate a measure on an interval scale for both persons and items. Even if not so trivial, trough RM it is also possible to perform target (b). In fact, by considering the multilevel extension of RM, also longitudinal data analysis is executable.

Keywords: Longitudinal data, Multilevel Rasch models, Side effects.

1. Introduction

The use of multilevel model (MLM) also called "hierarchical linear model", or "variance component model" , or "random coefficient model" in longitudinal data analysis is a well know (Bryk and Raudenbush, 1987 and 1992; Goldstein, 1987; Snijders, 1996) alternative to other competing techniques. Longitudinal data can, in fact, be view as multilevel data, with repeated measurements nested within individuals. One, between others, of the advantage of using MLM for longitudinal data is that *"the development*

over time is modelled by a linear regression equation, with possibly different regression coefficients for different individuals. Thus, each individual gets their own growth curve, specified by individual regression coefficients that may depend on individual attributes" (Hox, 2009). As pointed out by several authors, RM can be formulated as a member of Hierarchical Generalized Linear Model ($HGLM$) considering items as first-level units and persons as second-level units, in a fully crossed design (see, for example, for dichotomous responses: Kamata, 2001, Dorans *et al*., 2007; for polytomous responses: Pastor and Beretvas, 2006). The main advantages of considering RM in a $HGLM$ framework are the possibility to:
- include both item-level and person-level covariates into the model;
- estimate the model using any software programs for $HGLM$;
- easily add higher levels to the model to take into account the dependency among observations when lower-level units are nested within over-setting.
Longitudinal data are a special case of dependency among observations, where time repeated measures are nested within subjects. The simplest Longitudinal Rasch Multilevel Model ($LRMM$) is a 3-levels model where: first level is represented by items, the second level is represented by the time variable (taking values $0, 1, 2, \ldots$) and the third level is represented by persons. The model proposed is applied to cancer data to investigate malaise severity and site effects during chemotherapy cycle.

2. The Model

Lets consider a Polytomous Rasch Model (PRM) where data are responses of n persons ($i = 1, \ldots, n$) to J item ($j = 1, \ldots, J$), each having K ($k = 1, \ldots, K$) ordinal response categories, and denote with Y_{ij} response of person i to item j. According to PRM, probability of Y_{ij} is function of three sets of parameters: a set referred to persons (α_i), a set referred to item (θ_j) and a set to thresholds (τ_k). Following Pastor and Beretvas (2006) model formulation and denoting by p_{ij} the probability that Y_{ij} take values higher than or equal to k, for $k = 2, \ldots, K$, ($p_{ij} = P(Y_{ij} \geq k)$, Rasch model is defined by choosing the logit function to relate p_{ij} to the three sets of parameters, obtaining the so called cumulative logit model, *i.e.*: $\ln[p_{ij}/(1-p_{ij})] = \alpha_i - \theta_j + \tau_k$. Using standard GLM notation, let η_{ij} denote the logit link function, *i.e.* $\eta_{ij} = \log[p_{ij}/(1-p_{ij})]$. Extending Kamata (2001) formulation of Dichotomous Rasch Model (DRM) to a PRM, it is possible to write the model as a two level Hirearchical Generalized Linear Model. Let X_{qij} denote a dummy variable for person i taking value 1 when $q = j$ and 0 if $q \neq j$. Then the PRM assume the form:

$$\eta_{ij} = \beta_{0i} + \delta_{ki} + \sum_{q=1}^{J-1} \beta_{qi} X_{qij} \quad \text{and} \quad \begin{cases} \beta_{0i} = \gamma_{00} + u_{0i} \\ \beta_{ij} = \gamma_{j0} \\ \delta_{ki} = \delta_k \end{cases} \tag{1}$$

Note that in model (1) there are $J-1$ indicator variables: the dummy variable for the last item is dropped to obtain a full rank matrix. Right side of model (1) alone is a structural

model, the so called "level-1" or "item-level model": at this level β_s are item parameters, were β_{0i} (the intercept term) is associated to dropped item (usually the last) and can be interpret as the expected item effect of this item for person i. The other β_{qi} parameters are specific effects of items $1, \ldots, J-1$, expressed in term of difference form β_{0i}. The δ_{ki} are the threshold parameters for each category k. Note that at item-level model, items parameters are not constant across persons. A level-2 model is obtained adding the left-side of the (1): this two-level logistic model is equivalent to PRM where u_{0i} are person ability and β_s are item parameters, fixed across persons. Person abilities vary across persons and are fixed across items. Note that while in RM both item and person are fixed effect, in $HGLM$ formulation (1) person effects u_{0i} are random components of β_{0i} and it is typically assumed that $u_{0i} \sim N[0, var(u_{0i})]$. Combining the two sides of the (1), the following model for item q is obtained:

$$ln\left[\frac{P(Y_{ij} \geq k)}{1 - P(Y_{ij} \geq k}\right] = (\gamma_{00} + \gamma_{j0}) + \delta_k + u_{0i} \tag{2}$$

where parameters (γ_0 + γ_{q0}), δ_k and u_0 are equivalent, respectively, to α_i , τ_k and θ_j of the PRM. In this $HGLM$ framework it is easy to add another level to consider longitudinal data. The three-level model is:

$$\eta_{jmi} = \beta_{0mi} + \delta_{kmi} + \sum_{q=1}^{J-1} \beta_{qmi} X_{qjmi}, \tag{3}$$

$$\begin{cases} \beta_{0mi} = \gamma_{00i} + \gamma_{01i} d_{mi} + u_{0i} \\ \beta_{jmi} = \gamma_{j0i} + \gamma_{j1i} d_{mi} \\ \delta_{kmi} = \delta_{ki} \end{cases} \quad \text{and} \quad \begin{cases} \gamma_{00i} = \pi_{000} + r_{00i} \\ \gamma_{01i} = \pi_{010} + r_{01i} \\ \gamma_{j0i} = \pi_{j00} \\ \gamma_{j1i} = \pi_{j10} \\ \delta_{ki} = \delta_k \end{cases}$$

where: d_{mi} ($m = 1, \ldots, M$) is time variable coded so that it takes value 0 for $m = 1$; subscript m indicates second-level units; $u_{0mi} \sim N[0, var(u_{0mi})]$, with variance of u_{0mi} constant for every third-level units; note that al "level-2" item location β_{jmi} are supposed nonrandom function of time. Model (2) allows the third-level mean effect to vary across units, so that it models variation in growth trajectories among persons. In fact, while item location's intercept and linear coefficient for time remain constant across both the second and the third-level units, the latent trait parameters can vary randomly across persons. The π_s parameters represent: π_{000} the average overall latent trait at time 0; π_{010} the overall linear time effect; π_{q00} overall item location at time 0; π_{q10} the overall change in item location over time. The random effects r_{00i} and r_{01i} are assumed $N[0, \Sigma]$, where Σ is the variance-covariance matrix. Combining the three sides of the (3), model for item q can be obtain in a similar way as in (2).

3. Application

The *LRMM* is applied to observational cancer data. The analysis regards responses to 22 side effects, in a four-point ordinal scale (1=not at all, 2=a little, 3=quite a bit, 4=very much) of 88 women with the same stage of breast cancer, nested in six consecutive chemotherapy cycles. Site effects refer to physical and psychological malaises. After a preliminary explorative analysis seven physical side effects were deleted because of absence of variability during cycles. The parameters of the longitudinal multilevel model model are estimated using *Stata*. Site effects related to great malaise are mostly psychological (Sleepiness, Anger, Anguish, Depression) than physical (Alopecia and Nausea). Thresholds are: $\delta_1 = 0.935, \delta_2 = 2.217, \delta_3 = 3.785$ and show that is it very hard to endorse category "quite a bit" and "very much". Variance of the second level (cycles), equal to 0.282 with S.E.=0.025, is significative, but variance of the third level (patients) is almost zero. This means that the variability of responses depends mainly on cycles and not on women, therefore it makes sense to study individual trajectories distinguishing, for example, psychological from physical site effects behaviour.

References

Bryk A.G., Raudenbush S.W. (1987), Application of hierarchical linear models to assessing change, *Psychological Bulletin*, 101, 147-158.

Bryk A.S., Raudenbush S.W. (1992), *Hierarchical Linear Models in Social and Behavioral Research: Applications and Data Analysis Methods (First Edition)*, Sage Publications, Newbury Park, CA.

Doran H., Bates D., Bliese P., Dowling M. (2007), Estimating the Multilevel Rasch Model: With the lme4 Package, *Journal of Statistical Software*, 20, 1-18.

Goldstein H. (1987), *Multilevel Models in Educational and Social Research*, Griffin, London.

Hox J.J. (2009), in: T.D. Little, K.U. Schnabel, & J. Baumert (Eds.) *Modeling longitudinal and multiple-group data: Practical issues, applied approaches and specific examples*, Erlbaum, Hillsdale, NJ.

Kamata A. (2001), Item Analysis by the Hierarchical Generalized Linear Model, *Journal of Educational Measurement* , 38, No.1, 79-93.

Pastor D.A., Beretvas S.A. (2006), Longitudinal Rasch Modeling in the Context of Psychotherapy Outcomes Assessment, *Applied Psychological Measurement*, 30, 100-120.

Snijders T. (1996), Analysis of longitudinal data using the hierarchical linear model, *Quality & Quantity*, 30, 405-426.

Author Index

Quaderni di STATISTICA

Volume 14 - 2012

ISSN 1594-3739