Studies in Theoretical and Applied Statistics
Selected Papers of the Statistical Societies

Maurizio Carpita
Eugenio Brentari
El Mostafa Qannari *Editors*

Advances in Latent Variables

Methods, Models and Applications

Springer

Studies in Theoretical and Applied Statistics

Selected Papers of the Statistical Societies

Series Editors

Societa Italiana di Statistica (SIS)

Spanish Society of Statistics and Operations Research (SEIO)

Société Française de Statistique (SFdS)

Sociedade Portuguesa de Estatfstica (SPE)

Federation of European National Statistical Societies (FENStatS)

More information about this series at
http://www.springer.com/series/10104

Maurizio Carpita • Eugenio Brentari
El Mostafa Qannari
Editors

Advances in Latent Variables

Methods, Models and Applications

Springer

Editors
Maurizio Carpita
University of Brescia
Dept. of Economics and Management
Brescia
Italy

El Mostafa Qannari
Oniris Nantes National College
Dept. of Chemometrics and Sensometrics
Nantes
France

Eugenio Brentari
University of Brescia
Dept. of Economics and Management
Brescia
Italy

ISSN 2194-7767
ISSN 2194-7775 (electronic)
ISBN 978-3-319-02966-5
ISBN 978-3-319-02967-2 (eBook)
DOI 10.1007/978-3-319-02967-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2015934840

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The Italian Statistical Society (Società Italiana di Statistica - SIS) promotes every 2 years an international specialized statistical conference. The meeting focuses on both methodological and applied statistical research.

The SIS 2013 Statistical Conference "Advances in Latent Variables. Methods, Models and Applications", held in June 19–21, 2013 at the Department of Economics and Management of the University of Brescia, focused on advances in statistical methods and models for analyses with unobservable variables. Recently, an increasing interest has been devoted to this topic, from both methodological and applied points of view. Indeed, the latent variable approach allows us to effectively model complex real-life phenomena in a wide range of research fields.

The SIS 2013 Statistical Conference brought together statisticians from different research fields who exchanged experiences related to the analysis of latent variables and to the investigation of the relationships among them. The meeting was attended by 317 Italian and foreign scholars, who proposed 205 papers, which were accepted after a review process and presented in different sessions of the conference (5 plenary, 1 invited, 22 specialized, 16 solicited, 12 spontaneous and 1 poster). During the 3 days of the meeting, several special events took place: a special track on "Space and Space–Time Models: Methods and Environmental Applications" organized by the GRASPA-SIS group and devoted to the environmental statistics, the "*Sensory Sessions*" organized by the "Centro Studi Assaggiatori di Brescia" and the "International Academy of Sensory Analysis", the invited session on "Latent Models" organized by the Federation of European National Statistical Societies (FENStatS), and "The BES Day" on the measure of equitable and sustainable well-being organized by the Italian National Institute of Statistics (Istituto Nazionale di Statistica - ISTAT).

The 25 papers included in this book were selected from 38 extended versions presented at the SIS 2013 Meeting. A careful double-blind review process was adopted. We are grateful to the members of the Scientific Committee and to the 76 referees for their very helpful assistance. For convenience, the volume is organized

in seven parts: these only serve to orient the reader since methods, models and applications presented in the 25 papers overlap in some cases.

Finally, we would like to thank Alice Blanck and Carmina Cayago from Springer for their valued assistance in preparing this volume.

Brescia, Italy — Maurizio Carpita
Brescia, Italy — Eugenio Brentari
Nantes, France — El Mostafa Qannari

Contents

Identification of Clusters of Variables and Underlying Latent Components in Sensory Analysis

Evelyne Vigneau

Abstract

The Clustering of Variables around Latent Variables (CLV) approach aims to identify groups of features in a data set and, at the same time, to identify the prototype, or the latent variable, of each group. The procedure makes it possible to search for local groups or directional groups. Moreover, constraints on the latent variables may be added in order to introduce, if available, additional information about the observations and/or the variables. This approach is illustrated in two different contexts encountered in sensory analysis: (1) the clustering of sensory descriptors by taking into account their redundancy; and (2) the segmentation of a panel of consumers according to their liking, by taking into account external information about the products and the consumers.

Keywords

Clustering of variables • Sensory analysis • Segmentation of consumers • L-shaped data

1 Introduction

The clustering of variables may be relevant for a broad set of issues in sensory analysis. For instance, in hedonic studies, how many segments of consumers are there in the population under investigation? In quantitative descriptive analysis,

E. Vigneau (✉)
Nantes-Atlantic College of Veterinary Medicine, Food Science and Engineering (Oniris), Sensometrics and Chemometrics Laboratory, Site de la Géraudière, CS 82225, FR-44322 Nantes Cedex 3, France
e-mail: evelyne.vigneau@oniris-nantes.fr

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_19, Published online: 12 November 2014

could the list of attributes be reduced by selecting a subset of terms and finding the redundant attributes?

The Clustering of Variables around Latent Variables (CLV) is an approach that may be used in such a context. However, there is already a wealth of work on cluster analyses for variables, most of them being hierarchical. They vary according to both the nature of the variables at hand and, the choice of the measure of similarity between the variables. Considering quantitative variables only, most of the algorithms use Pearson's correlation coefficient between the variables, and sometimes the square of this coefficient. With some exceptions, such as the likelihood linkage analysis [7], the most common approaches could be qualified as empirical descriptive methods. Beside the great majority of hierarchical techniques based on similarity (or dissimilarity) matrices, the SAS Varclus procedure [12], the diametrical clustering method [3] and the CLV approach [15] are constructed on linear factor analysis. Moreover, the Varclus procedure and the CLV approach can both be used when it is desirable to merge strong positive or negative associated variables or, alternatively, when the aim is to separate anti-correlated variables. However, the underlying clustering strategies differ: the Varclus procedure is based on a divisive hierarchical algorithm, whereas the CLV method uses both an ascendant hierarchical algorithm and a partitioning algorithm. More importantly, the CLV approach consists of maximizing well-defined criteria (see Sect. 2.2) while the Varclus procedure does not. This feature enables the CLV approach to be used in a wide range of situations by introducing various constraints in the optimization problem (see Sects. 2.3–2.5). Finally, the specificity of the CLV is not only the determination of a partition of the variables but also the explicit definition of a latent variable associated with each cluster. From this point of view, it can be considered a method suitable for the extraction of a simple structure from a dataset, and is an alternative to the principal components rotation techniques, like Varimax [6].

In principle, the CLV approach consists of determining K clusters of variables and, simultaneously, K latent variables, such that the variables in each cluster are related as much as possible to the corresponding latent component. Nevertheless, the CLV methodology includes a collection of specific situations that will be detailed in Sect. 2. Illustrative case studies, in the context of sensory studies, will be presented in Sect. 3.

2 Methodology

2.1 Data Structure and Notation

Let us consider the $n \times p$ data matrix, $\mathbf{X}$. The aim is to split the p variables into a finite number, K, of clusters. If $\mathbf{X}$ contains the scores of acceptability of p consumers regarding n products (as in Sect. 3.2), a segmentation of the panel will be achieved. If the p variables are the sensory attributes of n products (as in Sect. 3.1), the clusters can be used as the basis of a selection procedure. The X-variables are centered. Optionally, they can be standardized by their standard deviation.

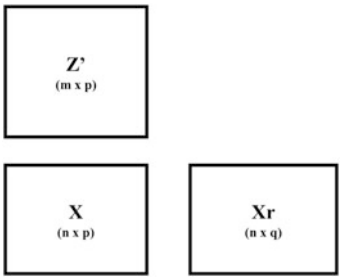

Fig. 1 Organization of the data: the variables in **X** are to be clustered, the **Xr** and **Z** blocks contain additional information (if this is available) which can be introduced within the CLV clustering procedure

Additional data may also be available on the n observations. These are arranged in a matrix **Xr** of size $n \times q$ (where "r" stands for right, as illustrated in Fig. 1). For instance, the **Xr** matrix can be associated with the experimental factors used for the formulation of the products or with their physico-chemical characterization (as in Sect. 3.2). Finally, if additional information is available on the variables to be clustered, another block of data, denoted **Z**, is considered (Fig. 1). The matrix **Z** is of size $p \times m$. In the case study presented in Sect. 3.2, **Z** allows socio-demographic information about the p consumers to be taken into account.

2.2 CLV for the Clustering of the X-Variables

We consider first the case where only the block of the X-variables is available. In this simple situation, two objectives for the clustering of the variables may be distinguished: if the aim is to separate variables that are highly, but negatively, correlated, each cluster must be defined locally around a latent variable that has the same orientation as the variables in the cluster; on the contrary, if the aim is to group together correlated variables in the same cluster, whatever the sign of the correlation coefficient, each cluster is to be defined directionally around a new axis.

Both cases are associated with a maximization problem [15]. The criterion to be maximized (1) involves the covariance between the variables and the latent variables for local groups:

$$\mathrm{S} = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, \mathrm{cov}(x_j, c_k) \qquad \text{with } c_k^t \, c_k = 1 \tag{1}$$

For directional groups, the criterion in (2) involves the squared covariance:

$$\mathrm{T} = \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} \, \mathrm{cov}^2(x_j, c_k) \quad \text{with } c_k^t \, c_k = 1 \tag{2}$$

In (1) and (2), K stands for the number of groups in the partition, $\boldsymbol{x}_j$ represents the j^{th} variable and $\boldsymbol{c}_k$ the latent variable associated with the k^{th} group. δ_{kj} is an indicator equal to 1 if the variable j belongs to the group k, and equal to 0 otherwise.

It is noteworthy that, in the case of the local groups, the latent variable $\boldsymbol{c}_k$ in a cluster G_k is proportional to the mean of the variables of this cluster. For

directional groups, the optimum value of T within each cluster is obtained when the latent variable $\boldsymbol{c}_k$ in G_k is the first normalized principal component of the variables belonging to this cluster.

2.3 CLV with External Information on the Observations

Let us consider the situation where additional information on the observations is available, for instance in external preference mapping when, as well as the acceptability scores in $\mathbf{X}$, the products have been described by a trained sensory panel using q attributes ($\mathbf{Xr}$). The CLV, with the criterion defined in (1) or (2), can be designed in such a way that each latent component in each segment of consumers, now denoted t_k, is a linear combination of the sensory attributes, as in (3):

$$t_k = \mathbf{Xr}\, a_k \qquad \text{with } a_k^t\, a_k = 1 \tag{3}$$

Thus, segmentation and external preference mapping can be achieved simultaneously. This procedure is much more straightforward than the usual method, which consists of separate steps of analysis: (1) a clustering of the consumers according to their acceptability scores and (2) a Principal Components Analysis of the sensory data. Finally, linear models are fitted in order to predict the mean acceptability scores in each cluster of consumers as a function of the first two sensory principal components. Using the criterion S in (1) with the constraint in (3), the hedonic segments are directly defined with reference to the sensory space. Case studies related to this situation have been presented in [14] and [16], for instance.

With the linear constraint in (3), it can be shown that the latent variable in each cluster is, in fact, the first PLS regression component of a PLS1 regression (in the case of local groups) or a PLS2 regression (in the case of directional groups).

2.4 CLV with External Information on the Variables

External information is often available on the variables to be clustered. This external information ($\mathbf{Z}$) may be the a priori knowledge of the characteristics of each variable.

For the clustering of the variables, in such a way that clusters of X-variables are related as much as possible to the external data, intermediate matrices must be defined [17]. In each group G_k, we consider the product matrix $\mathbf{P}_k = \mathbf{X}_k\, \mathbf{Z}_k$, where $\mathbf{X}_k$ and $\mathbf{Z}_k$ are formed by taking into account only the variables belonging to G_k. Thus, $\mathbf{P}_k$ may be regarded as the X-information weighted by something from the Z-information, or, if $\mathbf{Z}$ is centred, as an interaction matrix.

Finally, the group latent variable, denoted τ_k in this case, is expressed as a linear combination of the P-variables.

$$\tau_k = \mathbf{P}_k\, u_k \qquad \text{with } u_k^t\, u_k = 1 \tag{4}$$

2.5 CLV with External Information on Both the Observations and the Variables

This is a straightforward extension, mixing the strategies described in Sects. 2.3 and 2.4. Because of the L-shaped structure of the data matrices (Fig. 1), the clustering procedure is called the L-CLV. Two latent components are now determined in each cluster of variables: (1) the first, t_k, is defined in the space spanned by the co-variables in $\mathbf{Xr}$; and (2) the second, τ_{k}, is associated with the external information in $\mathbf{Z}$. The criterion to be maximized in each cluster involves the covariance between these two latent components. It is defined by:

$$S_{Xr}^{Z} = \sum_{k=1}^{K} \mathrm{cov}(t_k, \tau_k) \text{ with } t_k = \mathbf{Xr}\, a_k \text{ and } \tau_k = \mathbf{P}_{\mathrm{k}}\, u_k, \text{ where } \mathbf{P}_{\mathrm{k}} = \mathbf{X}_{\mathrm{k}}\, \mathbf{Z}_{\mathrm{k}} \tag{5}$$

Normalization constraints are set on the loadings vectors a_k and u_k, as in (3) and (4). The maximization of the criterion in (5) can also be presented as the maximization, in each cluster, of the quantity $u_k^t\, \mathbf{Z}_{\mathrm{k}}^{\mathrm{t}}\, \mathbf{X}_{\mathrm{k}}^{\mathrm{t}}\, \mathbf{Xr}\, \mathrm{a_k}$. In this way, it turns out that the L-CLV approach has many similarities with L-PLS regression [9]. Nevertheless, with the L-CLV, a specific triplet involving the three types of information is diagonalized and updated within each cluster G_k.

2.6 Algorithmic Point of View

The maximization of the CLV criteria, given in (1), (2) or (5), possibly subject to the constraints (3) or (4), may be achieved by an iterative alternating procedure whose monotonicity can be proven (except for the criterion in (5) for which it could only be observed based on experimental results). From an initial partition of the variables into K clusters, the two basic steps of the algorithm can be described as follows: (1) estimation of the group latent variables as indicated at the end of Sects. 2.2–2.5, (2) allocation of the variables to a cluster if its (squared) coefficient of covariance with the latent variable of this cluster is higher than with the other latent variables. For the criterion in (5) the variables considered in the allocation step are the P-variables instead of the X-variables.

In addition, within the CLV framework, we advocate the choice of the initial partition for the partitioning algorithm described above, on the basis of the results of a hierarchical procedure. In fact, the CLV criteria can also be involved in a hierarchical clustering process. Moreover, the advantage of performing a hierarchical algorithm before the partitioning algorithm is that the variation of the clustering criterion in the course of the hierarchy provides help in the choice of the number of clusters that can be retained.

All the CLV procedures have been implemented in an R-package, ClustVarLV, freely available on the CRAN Website [13].

3 Illustrative Examples

3.1 Clustering of Sensory Attributes in Sensory Profiling

In descriptive sensory profiling, the selection of attributes aims to provide a reduced list of terms by selecting the relevant and non-redundant attributes [11]. The initial list may be long, with, for instance, 30–50 attributes, organized a priori in various categories of sensory perception (odor, texture, flavor, off-flavor, etc).

Let us consider a case study [2] dealing with the sensory analysis of 12 varieties of apple from South Africa and New Zealand. The apples were peeled and quartered, and then assessed according to the 43 sensory descriptors listed in Table 1. The objective of the clustering of the 43 sensory attributes was to identify synthetic sensory latent traits using group latent variables. As directional groups of variables were sought, the CLV procedure was performed using the criterion T (2).

One of the outputs of the CLV is a graph showing the variation of the clustering criterion when passing from a partition into K clusters to a partition into (K-1) clusters. As shown in Fig. 2, the criterion clearly jumps when passing from four to three clusters. This means that there is a significant loss with only three clusters, and that four clusters should be retained. The four groups contained 12, 14, 12 and 5 attributes, respectively. The membership of the groups and the correlation coefficient of each sensory descriptor with its group latent variable are indicated in Table 1. Four main sensory latent traits were thus highlighted. The first latent variable (associated with the first group) was related to the internal odor and color of the apples, and had a gradient from green apple type to red apple type. The second latent variable also had a gradient from green to red apples but was related to their flavor. The third latent variable mainly gave information about the texture. The internal appearance and flavor attributes belonging to this group had a rather clear link with the texture of the apples. The last latent variable was related to bitterness, an undesirable taste for an apple.

Since each CLV latent variable is associated with a subset of the observed sensory descriptors, the underlying sensory traits are easy to interpret. More precisely, the CLV approach aims to identify a perfect simple structure, in the sense that each variable has exactly one non-zero loading for one latent variable. Bernaards and Jenrich [1] have shown that if this perfect simple structure exists, orthomax rotations (including Varimax) are able to retrieve it. This is a very interesting theoretical result. However, from an empirical point of view, we considered the results of the Varimax (orthogonal rotation) and Promax (oblique rotation) techniques [6] for the apple sensory data. In practice, the number of components of the loading matrix before rotation must be fixed. This is quite a complex problem. We considered solutions with four rotated components based on eleven selected sensory attributes: three from the group G_1 (“iogreen”, “ioredap”, “iagreen”), three from G_2 (“flgreen”, “flredap”, “flsweet”), three from G_3 (“txcrisp”, “txjuicy”, “iajuicy”) and two from G_4 (“flbitte”, “asbitte”). It turned out that neither the Varimax rotation nor the Promax rotation led to the expected partition. Using the Varimax rotation, the

Table 1 List of the sensory attributes organized according to the sensory category of the perception, with the results of the CLV procedure, i.e. the group to which each attribute was allocated, and the correlation between the attribute and the latent component of its group

Category	Attribute		CLV		Category	Attribute		CLV	
			G_k	$r(x_j, c_k)$				G_k	$r(x_j, c_k)$
Internal odor	iosweet	Sweet	1	0.86	Flavor during chewing	flgreen	Green apple	2	−0.93
	ioacids	Acidic/sour	1	−0.96		flredap	Red apple	2	0.93
	iogreen	Green apple	1	−0.92		flsweet	Sweet	2	0.86
	ioredap	Red apple	1	0.97		flaccid	Acidic/sour	2	−0.86
	iograss	Grassy	1	−0.89		flbitte	Bitter	4	−0.90
	iounrip	Unripe apple	1	−0.96		flgrass	Grassy	2	−0.85
	iocooka	Cooking apple	1	−0.96		flfresh	Fresh	3	−0.90
	iodampt	Damp twigs	2	−0.33		fldrop	Peardrops	2	0.76
	ioplums	Plums/cherries	1	−0.90		flwater	Watery	4	−0.75
Internal appearance	iawhite	White color	1	−0.76		floffl	Off flavor	4	−0.81
	iagreen	Green color	1	−0.92		flplumc	Plum/cherry	2	0.83
	iayelow	Yellow color	1	0.89		flunrip	Unripe apple	2	−0.93
	iagreli	Green lines	1	−0.89		flcoxli	Cox like	4	0.84
	iajuicy	How juicy	3	−0.90		flpearl	Pear like	3	0.81
	iatrans	Translucent	3	−0.79		flsoapy	Soapy	3	0.64
	iapulpy	Pulpy	3	0.83	After-swallow	assweet	Sweet	2	0.90
First bite	fbjuicy	Juicy	3	−0.79		asacid	Acidic/sour	2	−0.83
	fbhardn	Hardness	3	−0.93		asbitte	Bitter	4	−0.95
Texture during chewing	txcrisp	Crispness	3	−0.97		asgreen	Green apple	2	−0.94
	txjuicy	Juiciness	3	−0.94		asredap	Red apple	2	0.93
	txslowb	Slow break	3	−0.78		asastri	Astringent	2	−0.92
	txspong	Spongy	3	0.94					

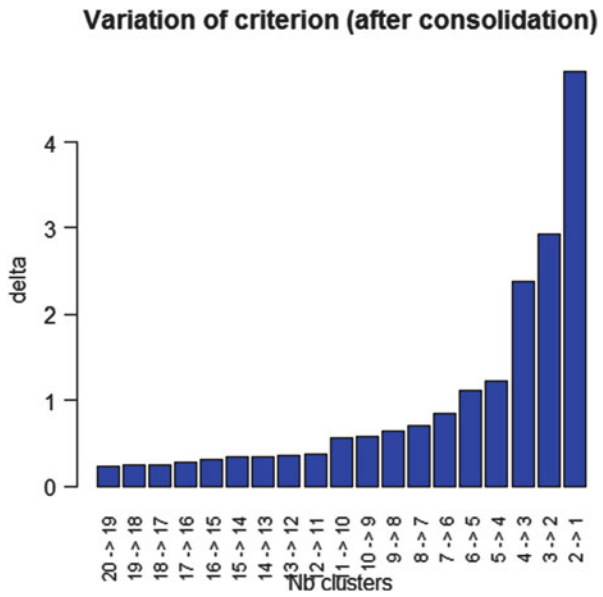

Fig. 2 Variation of the clustering criterion (delta) during the CLV procedure applied to the sensory attributes of apples

four groups obtained were {"iogreen", "ioredap", "iagreen", "flgreen", "flredap"}, {"iajuicy", "txcrisp", "txjuicy"}, {"flbitte", "asbitte"} and {"flsweet"}. In this solution, the internal odor and the flavor characteristics of the green/red apples were mixed, but the sweet flavor was set alone. With the Promax rotation, the four expected subgroups were better retrieved, except that the flavour "flgreen" was associated with the texture group instead of with "flredap" and "flsweet".

As illustrated in this case study, the CLV method is a simple strategy, which seems to provide results that make sense. Besides the identification of clusters of variables, the latent variables associated with the clusters give rise to a reduction in the dimensionality of the problem.

3.2 L-CLV for the Segmentation of Consumers

In order to identify the main "drivers of liking" of consumers regarding the sensory (or physico-chemical) properties of products of interest, different strategies for External Preference Mapping [10] may be adopted. The CLV procedure, as described in Sect. 2.3, is a simple approach to answer the question. Moreover, socio-demographic, usage and attitudinal information about the consumers is often collected by means of a questionnaire in an attempt to gain a greater understanding of the segmentation in the light of consumer characteristics. For a better and

straightforward integration of all the available information for the purpose of segmentation, the CLV has been extended to the L-CLV approach (Sect. 2.5).

Data collected in 2009 [4] are used to illustrate the L-CLV method. During the project mentioned, 25 juice prototypes were obtained by mixing one of five juices (apple, blood orange, orange, pineapple and pomegranate) with one of five freshly-squeezed berry fruit juices (strawberry, raspberry, blackberry, redcurrant and blueberry). The mixing proportions were 80 and 20 %.

Three blocks of information were collected and organized in an L-shaped structure, as illustrated in Fig. 1. The core data matrix, **X**, contained the hedonic scores, on a nine-point scale, of $p = 69$ consumers regarding the $n = 25$ juice mixes. The external data matrix, **Xr**, which gave additional information on the products, contains two parts: the amounts of five chemical compounds (sugar, malic acid, citric acid, ascorbic acid and total polyphenols) in the juices, and the factors of the mixture design. The responses of the consumers to a usage and attitude questionnaire were coded in the matrix **Z**. This included the usual demographic characteristics, information on the consumption of fruit and fruit juice (when, how, etc.), liking assessments for 24 different fruits and for nine types of fruit juice, criteria sought for fruit juices, opinions on new foods, and opinions on berry fruit. In total, 97 questions were taken into account, which is much greater than the number of items considered in a previous work [17]. Twenty-one of these questions were qualitative items coded by means of dummy variables, so that **Z** finally contained $m = 126$ columns.

Regarding the evolution of the L-CLV criterion, two main segments of consumers emerged, but a finer partition into four segments (Seg1–Seg4) was retained. These segments contained 15, 20, 23 and 11 consumers, respectively. Comparing these with the partition into two segments, it turns out that Seg2, Seg3 and Seg4 were aggregated into a mega-segment whereas Seg1 remained separate.

Among the outputs of the L-CLV procedure, the loadings associated with the external information on the products (a_k) and with the consumer background descriptors (u_k) were especially interesting. Figure 3 gives the (standardized) loadings associated with the five chemical compounds or related to the type of fruit mixed. Clearly, the type of berry fruit used (S, R, BLK, RC or BLU) did not segment the panel. In contrast, the nature of the main juice (A, BO, O, P or PG) led to different profiles of liking or disliking. The first segment was formed by those consumers who did not reject the juices made with pomegranate. On the contrary, the consumers in Seg2 and Seg3 clearly rejected the juices with pomegranate. A difference between these latter two segments was related to juices with pineapple. Regarding the fourth segment, it seems that these consumers did not like juices made from apple or pomegranate.

Concerning the consumer attributes, the relatively high number of items makes it rather difficult to scrutinize the loadings vectors of each segment. We have chosen to illustrate the results regarding the age and gender of the consumers as well as their criteria for juices. For instance, from Fig. 4 it can be seen that the consumers in Seg1, who liked juices based on pomegranate (contrary to those in the other segments) but did not much like juices with orange, were not the youngest students.

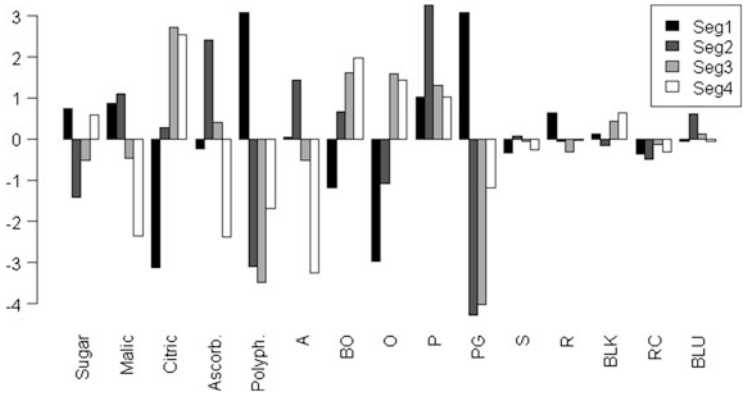

Fig. 3 Loadings associated with the chemical compounds (sugar, malic acid, citric acid, ascorbic acid, polyphenols) and the design factors (*A*: apple, *BO*: blood orange, *O*: orange, *P*: pineapple, *PG*: pomegranate for the main juice, and *S*: strawberry, *R*: raspberry, *BLK*: blackberry, *RC*: redcurrant and *BLU* blueberry for the co-fruit)

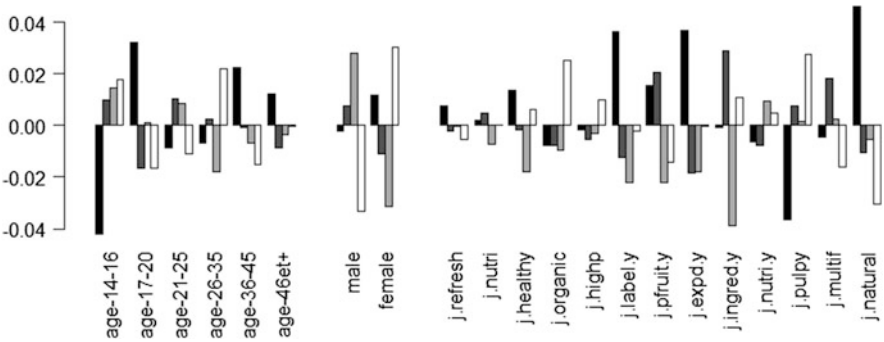

Fig. 4 Loadings associated with chosen attributes for the consumers, i.e. age category, gender and criteria regarding fruit juice (13 items: 5 "must be" items (must be refreshing_j.refresh, nutritious_j.nutri, healthy_j.healthy, organic_j.organic, high % of fruit_j.highp); 5 "I read" items (I read the label_j.label.y, the percentage of fruit_j.pfruit.y, the expiration date_j.expd.y, the ingredients_j.ingred.y, the nutritional facts_j.nutri.y); and 3 items of choice which are pulpy vs liquid_j.pulpy, multi-fruit vs single fruit_j.multi-fruit, enriched vs natural_j.enriched)

They stated that they read the label and the expiration date, thought that juices had to be healthy, and preferred liquid and natural juices to pulpy and enriched juices. The consumers in Seg2, who liked juices with pineapple, said that they paid attention to the percentage of fruit in a juice. The number of men in Seg3 was rather high. These consumers did not seem to pay much attention to fruit juices. Finally, in Seg4 there was a higher proportion of women than in the other groups. These consumers said that juices had to be organic, that, unlike people in Seg1, they preferred enriched and pulpy juices, and, not surprisingly, that they did not much appreciate juices with apple.

The segmentation of the panel obtained when neither the external information about the products nor the external information collected about the consumers was taken into account did not distinguish between Seg3 and Seg4. However, even though people belonging to the fourth segment represented only 16 % of the panel, their attitude to fruit juice was less neutral than that of people in Seg3, and was, more or less, opposite to that of people in Seg1.

Conclusion

The CLV approach has been demonstrated as attractive for explorative data analysis, and complementary to Principal Components Analysis or Factor Analysis, for instance. In the context of sensory analysis, the use of the CLV approach can provide answers to several common issues, from the reduction of the number of attributes necessary for the sensory description of the products to the analysis of the preference of the consumers. However, its field of application is not restricted to the sensory domain (see, for instance, in vibrational spectroscopy [18] or in the health domain [5, 8]).

One of the specificities of the CLV approach is that each cluster is associated with a latent variable. Thus, the CLV is not only a clustering method but also a useful way of reducing the dimensionality and identifying the perfect simple structures in a data set. Moreover, the interpretability of these latent variables can be made easier by the use of external information on the observations and/or the variables.

In particular, the L-CLV procedure has been developed specifically to address the question of the segmentation of consumers, with a better capacity for interpretation in terms of sociological and behavioral parameters, and in relation to the sensory or compositional key-drivers.

References

1. Bernaards, J., Jennrich, R.: Orthomax rotation and perfect simple structure. Psychometrika **69**(4), 585–588 (2003)
2. Daillant-Spinnler, B., MacFie, H.J.H., Beyts, P.K., Hedderley, D.: Relationships between perceived sensory properties and major preference directions of 12 varieties of apples from the Southern Hemisphere. Food Qual. Prefer. **7**, 113–126 (1996)
3. Dhillon, I.S., Marcotte, E.M., Roshan, U.: Diametrical clustering for identifying anti-correlated gene clusters. Bioinformatics **19**(13), 1612–1619 (2003)
4. Endrizzi, I., Pirretti, G., Calò, D.G., Gasperi, F.: A consumer study of fresh juices containing berry fruits. J. Sci. Food Agric. **89**, 1227–1235 (2009)
5. Golden, J., Conroy, R.M., O'Dwyer, A.M., Golden, D., Hardouin, J.-B.: Illness-related stigma, mood and adjustment to illness in persons with hepatitis C. Soc. Sci. Med. **63**, 3188–3198 (2006)
6. Jolliffe, I.: Principal Component Analysis, 2nd edn. Springer, New York (2002)
7. Kojadinovic, I.: Hierarchical clustering of continuous variables based on the empirical copula process and permuation linkages. Comput. Stat. Data Anal. **54**, 90–108 (2010)
8. Lovaglio, P.G.: Model building and estimation strategies for implementing the balanced scorecard in health sector. Qual. Quant. **45**, 199–212 (2001)

9. Martens, H., Anderssen, E., Flatberg, A., Gidskehaug, L.H., Hoy, M., Westad, F.: Regression of a matrix on descriptors of both its rows and its columns via latent variables: L-PLSR. Comput. Stat. Data Anal. **48**, 103–123 (2005)
10. Næs, T., Brockhoff, P.B., Tomic, O.: Statistics for Sensory and Consumer Science. Wiley, Chichester (2010)
11. Sahmer, K., Qannari, E.M.: Procedures for the selection of a subset of attributes in sensory profiling. Food Qual. Prefer. **19**, 141–145 (2008)
12. Sarle, W.S.: SAS/STAT User's Guide: The VARCLUS Procedure, 4th edn. SAS Institute, Cary (1990)
13. Vigneau, E., Chen, M.: ClustVarLV : Clustering of Variables Around Latent Variables, R Package Version 1.3.1 (2014). http://cran.r-project.org/web/packages/ClustVarLV
14. Vigneau, E., Qannari, E.M.: Segmentation of consumers taking account of external data: a clustering of variables approach. Food Qual. Prefer. **13**, 515–521 (2002)
15. Vigneau, E., Qannari, E.M.: Clustering of variables around latent component. Comm. Stat. Simul. Comput. **32**, 1131–1150 (2003)
16. Vigneau, E., Charles, M., Chen, M.: External preference segmentation with additional information on consumers. Food Qual. Prefer. **32**, 83–92 (2014)
17. Vigneau, E., Endrizzi, I., Qannari, E.M.: Finding and explaining clusters of consumers using CLV approach. Food Qual. Prefer. **22**, 705–713 (2011)
18. Vigneau, E., Sahmer, K., Qannari, E.M., Bertrand, D.: Clustering of variables to analyze spectral data. J. Chemom. **19**, 122–128 (2005)

Clustering the Corpus of Seneca: A Lexical-Based Approach

Gabriele Cantaluppi and Marco Passarotti

Abstract

We present a lexical-based investigation into the corpus of the *opera omnia* of Seneca. By applying a number of statistical techniques to textual data we aim to automatically collect similar texts into closely related groups. We demonstrate that our objective and unsupervised method is able to distinguish the texts by work and genre.

Keywords

Hierarchical Clustering Analysis • Contribution Biplots • Principal Component Analysis • Latin • Seneca

1 Introduction

We present a lexical-based investigation into the corpus of the *opera omnia* of Seneca. By applying a number of statistical techniques to textual data, we aim to automatically organize the texts in such a way that those works that share a relevant amount of lexical items are considered to be very similar to each other and get automatically collected into closely related groups.

G. Cantaluppi (✉)
Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italy
e-mail: gabriele.cantaluppi@unicatt.it

M. Passarotti
Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione, Università Cattolica del Sacro Cuore, Milano, Italy
e-mail: marco.passarotti@unicatt.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_6, Published online: 12 November 2014

In order to detail the lexical similarities and differences between the texts, we apply a technique that is able to highlight the words that mostly characterize one or more texts in comparison to the others.

The paper is organized as follows. Section 2 presents the data that we used. Section 3 details our method for clustering the data and performing principal component analysis. Section 4 shows and evaluates the results. Section 5 reports a number of conclusions and introduces our future work.

2 Data

Lucius Anneus Seneca (4 BC–65 AD) was a Roman Stoic philosopher, statesman and gramatis. He is considered to be among the most important authors of the Classical era of Latin literature. His tragedies (most of which are based on Greek original texts) are the only complete Latin tragedies extant. The corpus of the opera omnia of Seneca is quite diverse in terms of both literary genres featured and topics addressed. This motivates its clustering analysis, aimed to collect together those texts that feature a similar lexicon, checking if the results are consistent with differences in literary genre and topics.

The corpus featuring the *opera omnia* of Seneca is taken from the lexicon of the Stoics provided by [15]. The corpus comprises 23 works, among which are eight tragedies, ten dialogues and the full text of *Apocolocyntosis*, *Epistulae morales*, *Naturales quaestiones*, *De clementia* and *De beneficiis* (divided into seven books). Two tragedies of disputed attribution (*Hercules Oetaeus* and *Octavia*) are provided as well. The size of the corpus is approximately 364,000 words. All texts come from authoritative editions. For more details, see [15], XV–XVI. The corpus is fully lemmatized.

3 Method

We applied two statistical techniques to textual data, namely clustering and principal component analysis.

All the experiments were performed with the R statistical software [14]. In particular, we used the "tm" package to build and analyze the document-term matrices that are employed for clustering [4, 5]. Distance and similarity measures provided by the package "proxy" were used as well [12].

3.1 Clustering

Clustering methods can be applied to several different kinds of data, among which are textual data, whose "objects" are occurrences of words in texts. As far as word sense disambiguation is concerned, clustering lies on the theoretical assumption stated by Harris' Distributional Hypothesis, according to which words that are used in similar contexts tend to have the same or related meanings [9]. This basic assumption is well summarised by the famous quotation of Firth [6]:

You shall know a word by the company it keeps.

In this work, we apply hierarchical agglomerative clustering in order to compute and graphically present similarity/dissimilarity between texts. As we deal with texts instead of occurrences of words, this led us to slightly modify the two basic theoretical assumptions mentioned above. Thus, here we assume that

1. texts that feature a similar (distribution of) lexicon tend to address the same or related topics (Harris-revised);
2. you shall know a text by the words it keeps (Firth-revised).

These two assumptions are reflected in our clustering method, which compares the texts by computing their distance in terms of similarity as follows:

Data Cleaning. We remove punctuations and function words from input data. All characters were translated to lower case. In particular, we remove all (both coordinative and subordinative) conjunctions, prepositions, pronouns and those adverbs that cannot be reduced to another lemma (like *diu*, *nimis* and *semper*);

Hierarchical Agglomerative Clustering Analysis: Distance. Clustering analysis is run on document-term matrices by using the cosine distance $\mathrm{d}(i, i') = 1 - \cos\{(x_{i1}, x_{i2}, \ldots, x_{ik}), (x_{i'1}, x_{i'2}, \ldots, x_{i'k})\}$. The arguments of the cosine function in the preceding relationship are two rows, i and i', in a document-term matrix; x_{ij} and $x_{i'j}$ provide the number of occurrences of word j $(j = 1, \ldots, k)$ in the two texts corresponding to rows i and i' (profiles).

Zero distance between two documents holds when two documents with the same profile are concerned (i.e. they have the same relative conditional distributions of terms). In the opposite case, if two texts do not share any word, the corresponding profiles have distance 1;

Hierarchical Agglomerative Clustering Analysis: Clustering. We run a complete linkage agglomeration method. While building clusters by agglomeration, at each stage the distance (similarity) between clusters is determined by the distance (similarity) between the two elements, one from each cluster, that are most distant. Thus, complete linkage ensures that all items in a cluster are within some maximum distance (or minimum similarity) of each other.

Roughly speaking, according to our clustering method, works that share a high number of lemmas with similar distribution are considered to have a high degree of similarity and, thus, fall into the same or related clusters.

3.2 Principal Component Analysis

While clustering computes and represents the degree of similarity/dissimilarity between texts by clusters, it does not inform about which features distinguish one text from the other. These features are those properties that make two texts similar or dissimilar to each other.

As our method is highly lexical-based, the features that we consider are words (either lemmas or forms). In order to know which words distinguish one or more texts from the others, we apply the principal component analysis technique.

Principal component analysis is a method used to retrieve a structure built according to one or more latent dimensions. This structure can be defined by using different features: in our case, the features are words, which are used as bag-of-words representations of texts. Such representations of texts get mapped into a vector space that is assumed to reflect the latent dimension structure.

We follow the Principal Component Analysis (PCA) presentation (described e.g. by [11]) and produce contribution biplots that graphically represent a vector space [8, p. 67]. Starting from an $I \times J$ term-document matrix $\mathbf{Y}$ (whose values were previously standardized by column, in order to overcome the size differences between texts), a reduction of the column (document) space can be achieved by using principal component analysis and considering dimensions which relate texts that show high correlation in their term distributions.

A singular value decomposition (SVD) of $\mathbf{Y}/(IJ)^{1/2}$ is then performed

$$\mathbf{S} = \mathbf{Y}/(IJ)^{1/2} = \mathbf{U}\mathbf{D}_\beta\mathbf{V}'$$

where $\mathbf{U}$ and $\mathbf{V}$ are matrices containing respectively the left and the right singular vectors and $\mathbf{D}_\beta$ is a diagonal matrix containing the singular values in decreasing order.

The SVD allows the calculation of coordinates $\mathbf{U}$ for terms and $\mathbf{G} = J^{1/2}\mathbf{V}\mathbf{D}_\beta$ for documents. By considering the first two columns of $\mathbf{U}$ and $\mathbf{G}$, we have the coordinates with respect to the first two principal components.

The squares of the elements in $\mathbf{D}_\beta$ divided by their total inform about the amount of variance explained by the principal components. By considering the squared values of the coordinates of terms we obtain their contribution to principal axes.

4 Results and Evaluation

The results on Seneca's works are reported by a genre-based order: first the dialogues, then the tragedies and, finally, the *opera omnia*.

4.1 Dialogues

Figure 1 presents the clustering plot for the ten dialogues of Seneca.

According to agglomerative hierarchical clustering, each text starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy by an always lower degree of similarity. Clustering ends once all the texts are collected into one common cluster, in this case showing that the dialogues of Seneca are dissimilar at the height of 0.20 (i.e. similar at 0.80).

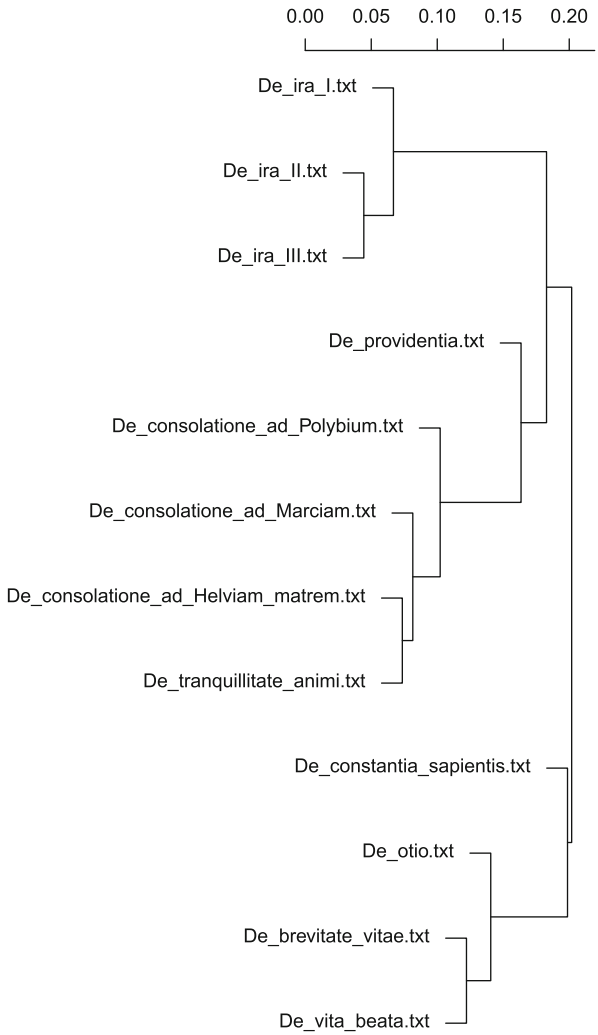

Fig. 1 Clustering the dialogues

For instance, the three books of *De ira* (which are clustered together) are dissimilar from *De providentia*, from the three *Consolationes* and from *De tranquillitate animi* at the height of 0.18 (i.e. similar at 0.82), while they are dissimilar from each other at the height of 0.07 (i.e. similar at 0.93). Among the three books of *De ira*, the second and the third are closer to each other than to the first one.

In Fig. 1 we can see that the three books of *De ira* are clustered apart from the other dialogues. Principal component analysis is able to answer the question about what makes *De ira* different from the other dialogues. As our method is lexical-based, this question concerns the words (in this case, the lemmas) that distinguish *De ira* from the other dialogues.

Figure 2 is a contribution biplot that presents the results of the principal component analysis performed on the term-document matrix of the dialogues of Seneca. In particular, the biplot represents the rows and the columns of the term-document matrix through a graph whose axes are the two first principal components, as we observed that these are able to explain over the 90 % of the total variance among texts.[1]

The first principal component gets graphically represented on the horizontal axis of the contribution biplot and it is able to explain alone most of the variance (0.876). As all the dialogues polarize in the same direction (the rightside of the biplot), the first principal component describes a dimension that is common to all the texts involved.

The second principal component is reported on the vertical axis of the biplot and it explains the 0.026 of the variance among texts. This component describes a dimension that is able to detail what mostly characterize one or more texts in comparison to the others.

For instance, the verb *sum* is placed right in the center of the vector (approximately at height 0.0 on the vertical axis). This means that *sum* is a kind of a "neuter" lemma, which is common to all the texts and does not characterize any of them in

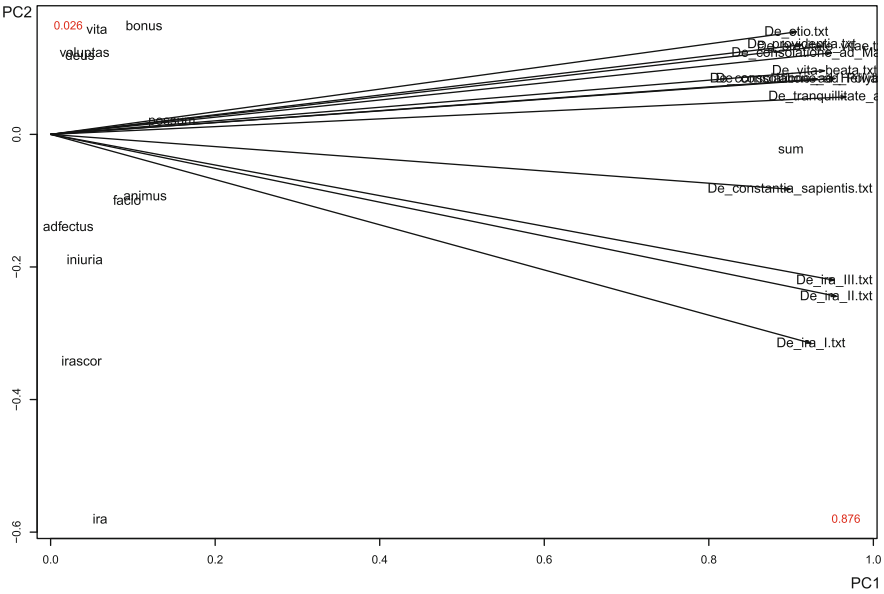

Fig. 2 Principal component analysis of the dialogues

[1] In more detail, the first two principal components explain the 0.902 of the variance, this proportion resulting from the sum of the explaining power of each of the two components (respectively, 0.876 and 0.026).

comparison to the others. Instead, the lemmas *iniuria*, *ira* and *irascor* are moved from the center and characterize the three books of *De ira*, which are all set apart from the other dialogues in the biplot.

Although the second principal component explains just the 0.026 of the total variance among texts, it is still able to report meaningful differences, which allow to recognize the specific lexical features that distinguish *De ira* from the other dialogues.

4.2 Tragedies

Figure 3 reports the clustering plot for the eight tragedies of Seneca plus the two ones of disputed attribution.

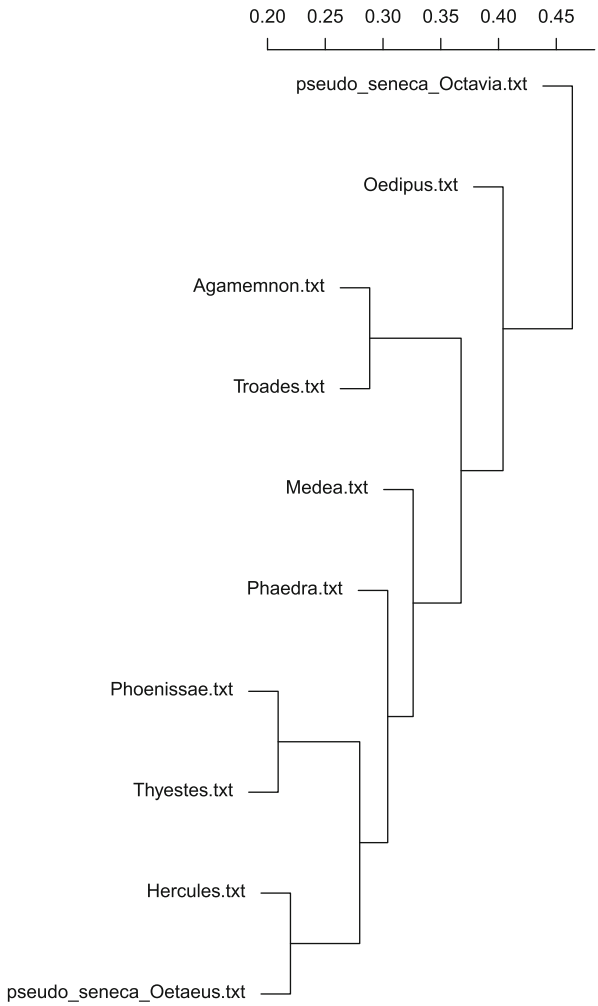

Fig. 3 Clustering the tragedies

The long lasting debate about the attribution to Seneca of the *Hercules Oetaeus* and the *Octavia* has led to the generally assumed conclusion that the *Hercules Oetaeus* is much probably original, while the *Octavia* is an imitation of the tragic style of Seneca.[2] This is reflected by our results too. Indeed, the *Hercules Oetaeus* is collected into the cluster of the original tragedies and, in particular, it is clustered together with the *Hercules* (which is not surprising, because these two tragedies cover a much similar topic). Conversely, the *Octavia* is clustered apart from the other tragedies (like the *Oedipus*).

Figure 4 shows the results of the principal component analysis performed on the tragedies. The lemmas that characterize the *Octavia* in comparison to the other tragedies are *coniunx*, *nero*, *nutrix*, *octavia*, and *seneca*. These words summarize well the contents of this *fabula praetexta* that tells the story of Octavia, who was the first wife (*coniunx*) of the emperor Nero. Further, one of the main arguments in favour of considering the *Octavia* a not original tragedy of Seneca is that one of the characters of the story is named Seneca, which is again reflected by our principal component analysis.

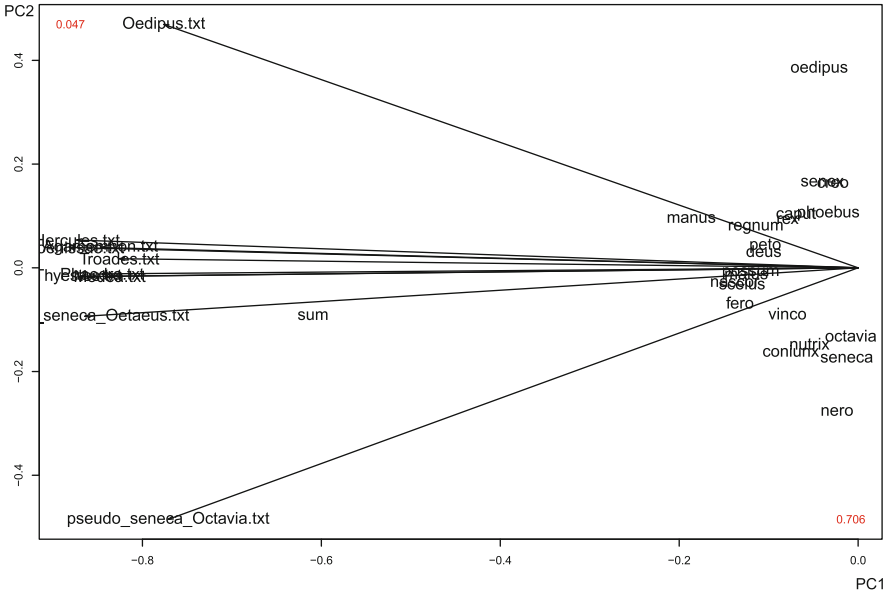

Fig. 4 Principal component analysis of the tragedies

[2] About the attribution of the *Hercules Oetaeus* see [3, 10, 13]. On the *Octavia* see [1, 2, 7, 17]. Among the contributions in favour of the authenticity of both the tragedies see [16].

4.3 Opera Omnia

Figure 5 presents the results of clustering the *opera omnia* of Seneca.

Texts get organized into two main clusters, one including the tragedies and the other featuring the dialogues and the other writings. Within the latter, the *Apocolocyntosis* is clustered apart from the other works and all the seven books of *De beneficiis* get clustered together.

The *Apocolocyntosis* is a menippean satyre (a kind of mixture of prose and poetry) and it is indeed a text quite different from the others of Seneca: it is, thus, not surprising that it belongs to a separate cluster.

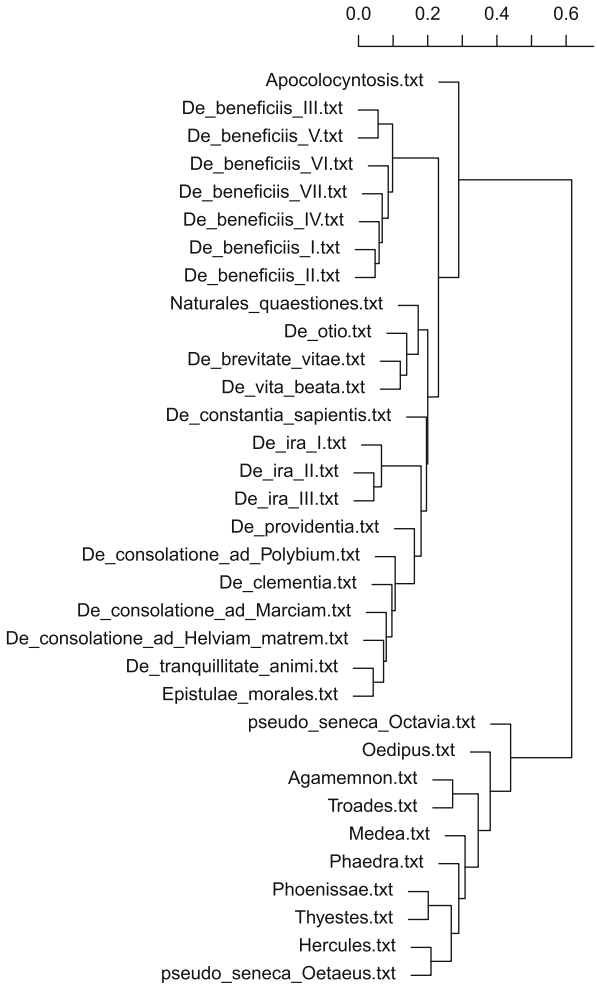

Fig. 5 Clustering the *opera omnia*

The fact that the tragedies are set apart from the other works and that all the books of *De beneficiis* and all those of *De ira* are clustered together shows that our method is able to distinguish texts not only by genre but also by single work.

Figure 6 presents the results of the principal component analysis performed on the *opera omnia* of Seneca. The seven books of *De beneficiis* deviate from the other works and are characterized by the following lemmas: *beneficium*, *gratia*, *gratus*, *ingratus* and *reddo*.

Along all our experiments we observed that *De consolatione ad Polybium* was always clustered separately from the other two *consolationes*. Figure 7 reports the contribution biplot that highlights the lexical features of these three texts, showing that *De consolatione ad Polybium* is characterized by the lemmas *bonus*, *caesar*, *dolor*, *fortuna* and *frater*, while *De consolatione ad Marciam* and *De consolatione ad Helviam matrem* are distinguished by *filius*, *locus*, *mater*, *vir* and *vivo*. The three texts share an high average relative frequency of lemmas like *animus*, *homo* and *natura*.

The biplot reported in Fig. 7 looks different from those presented so far, as it features a massive central black area formed by those lemmas that are shared by the three *consolationes*. Although such an area was present also in all the biplots reported above, it was always removed for presentation purposes. In this case, we left the black area in on purpose, in order to show how big is the number of lemmas with similar relative frequency that are shared by these three texts which, indeed, appear as clustered very close to each other in Fig. 5.

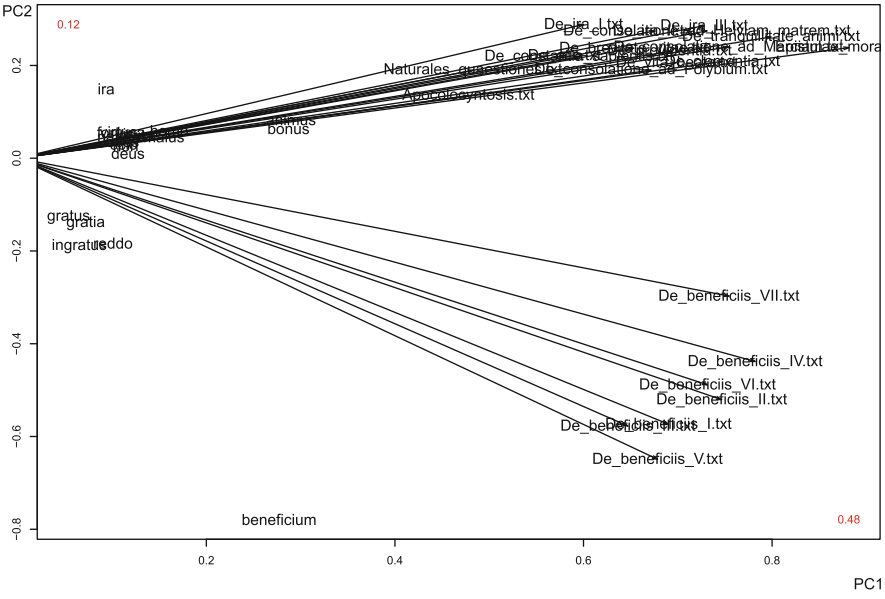

Fig. 6 Principal component analysis of the *opera omnia*

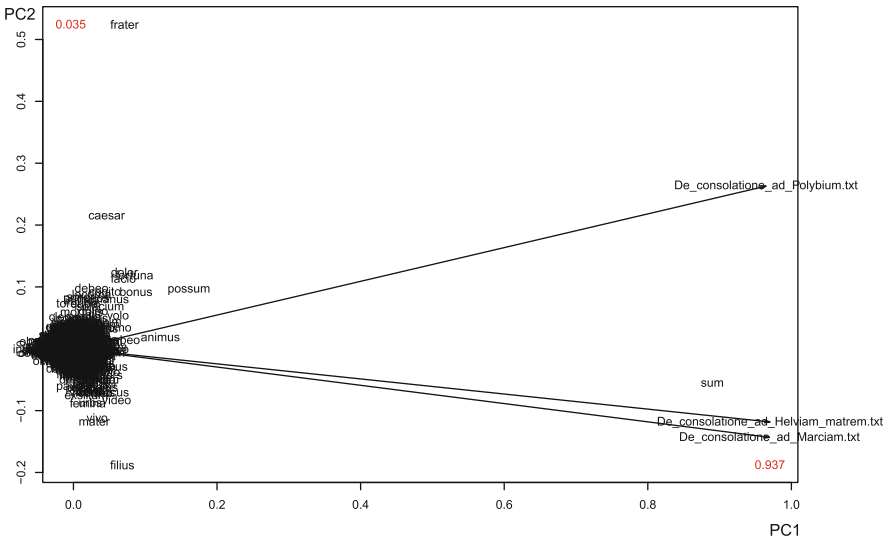

Fig. 7 Principal component analysis of the three *Consolationes*

5 Discussion and Future Work

The main feature of our method is that it provides an objective and unsupervised analysis of textual data. Our results can be replicated and the method is open to be refined in order to achieve better results.

The R software allows an efficient managing of big amounts of data. This gave us the opportunity to perform always full-text analyses, instead of grounding our experiments on manually selected excerpts built in a subjective fashion.

At first, our method wants to quantify and verify (or not) previous intuition-based assumptions on the evidence provided by textual data. For instance, our results do not just show that the tragedies of Seneca are different from his dialogues (which is indeed neither a new nor a really interesting fact), but they report objectively how much different the tragedies are from the dialogues and how much different they are from each other according to their lexical features.

Then, our fully data-driven approach does not only add empirical evidence to subjective intuitions about texts, but it allows also to bring to light previously overlooked relations between texts, like in the case of the relation between *De consolatione ad Polybium* and the other two *consolationes*.

Further, such a method can also be used for authorship attribution purposes, like in the case of the *Octavia*. However, this may lead to promising results just in those cases where the works of disputed attribution differ from the original ones by lexical features. If differences concern other linguistic properties of the texts (ranging from syntax to semantics and, more generally, to literary style), a lexical-based approach is not the best fitting one.

As mentioned above, all our results were driven by the assumption that "you shall know a text by the words it keeps". This entails that the texts involved in our experiments get clustered according to their lexical properties. In this context, thus, saying that one work is close to another means that they share a relevant amount of (non-function) words showing a similar distribution. In light of the results achieved, such a basic assumption seems to be working, as the organization of texts that automatically results from applying our method corresponds to the different works and genres involved in the several experiments performed.

In the near future, we want to refine our method both by providing a more fine-grained subdivision of data and by exploiting higher layers of linguistic annotation of texts.

As the former is concerned, we shall organize the data according to the sub-parts of the texts (books, chapters etc.): for instance, we should provide one separate file for each letter of the *Epistulae morales*.

As for the latter, we first want to compare the texts by distribution of Parts of Speech (PoS) and colligations (i.e. co-occurrences of PoS). At the higher level, we have to exploit syntactically annotated data (produced by parsers, or made available in treebanks) in order to compare the texts by phrases and/or chuncks instead of single words. And finally, we can use second-order features as well (like semantic descriptions of lexical items provided by Latin WordNet) to enhance the information provided by words.

References

1. Beck, J.W: «Octavia» Anonymi: zeitnahe «praetexta» oder zeitlose «tragoedia»?: mit einem Anhang zur Struktur des Dramas, Duehrkohp und Radicke, Göttingen (Göttinger Forum für Altertumswissenschaft. Beihefte 15) (2004)
2. Bruckner, F.: Interpretationen zur Pseudo-Seneca-Tragödie Octavia, Offsetdruckerei Hogl, Erlangen (1976)
3. del Río Sanz, E.: Problemas de autenticidad del Hercules Oetaeus. Estado de la cuestión, Cuadernos de Investigación Filológica, XII-XIII, 147–153 (1987)
4. Feinerer, I., Hornik, K.: tm: Text Mining Package. R package version 0.5-9.1. http://CRAN.R-project.org/package=tm (2013)
5. Feinerer, I., Hornik, K., Meyer, D.: Text mining infrastructure in R. J. Stat. Softw. **25**(5), 1–54. http://www.jstatsoft.org/v25/i05/ (2008)
6. Firth, J.R.: Papers in Linguistics 1934–1951. London University Press, London (1957)
7. Giancotti, F.: L'Octavia attribuita a Seneca. Loescher-Chiantore, Torino (1954)
8. Greenacre, M.: Biplots in Practice. Fundación BBVA, Madrid (2010)
9. Harris, Z.S.: Distributional structure. Word **10**, 146–162 (1954)
10. Iorio, V.: L'autenticità della tragedia Hercules Oetaeus di Seneca. Rivista Indo-Greca-Italica di filologia, lingua, antichità, 20, 1–59 (1936)
11. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Prentice Hall, Upper Saddle River (2002)
12. Meyer, D., Buchta, C.: proxy: Distance and Similarity Measures. R package version 0.4-10. http://cran.r-project.org/web/packages/proxy/index.html (2013)
13. Paratore, E.: Lo Hercules Oetaeus è di Seneca ed è anteriore al Furens, in Acta classica: verhandelinge van die Klassieke Vereniging van Suid-Afrika = Proceedings of the Classical Association of South Africa, I, 72–79 (1958)

14. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/ (2012)
15. Radice, R., Bombacigno, R.: Lexicon IV: Stoics. Biblia, Milano (2007)
16. Ruiz de Elvira, A.: La «Octauia» y el «Hercules Oetaeus»: tragedias auténticas de Séneca, in Urbs aeterna: actas y colaboraciones del Coloquio Internacional Roma entre la Literatura y la Historia: homenaje a la profesora Carmen Castillo, Ediciones Universidad de Navarra, Pamplona, pp. 909–919 (2003)
17. Runchina, G.: Sulla pretesta Octavia e le tragedie di Seneca. In: Rivista di cultura classica e medioevale, vol. 6, pp. 47–63 (1964)

Modelling Correlated Consumer Preferences

Marcella Corduas

Abstract

The CUB model is a mixture distribution recently proposed in literature for modelling ordinal data. The CUB parameters may be related to explanatory variables describing the raters or the object of evaluation. Although various methodological aspects of this class of models have been investigated, the problem of multivariate ordinal data representation is still open. In this article the Plackett distribution is used in order to construct a bivariate distribution from CUB margins. Furthermore, the model is extended so that the effect of rater characteristics on their stated preferences is included.

Keywords

CUB models • Food quality • Ordinal data • Plackett distribution

1 Introduction

The level of consumer satisfaction with products and services is often investigated by complex surveys. In such a context, respondents may be requested to rate several attributes and to express their preferences by means of a Likert scale. In addition, interviewees may belong to different categories since they are usually stratified according to relevant features such as geographic location or gender.

The statistical analysis of this type of data traditionally relies on Generalized Linear Models which offer a fundamental framework for methodological developments and empirical applications [21, 22].

M. Corduas (✉)
Department of Political Sciences, University of Naples Federico II, Napoli, Italy
e-mail: corduas@unina.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_9, Published online: 12 November 2014

In this article we consider an alternative class of models, denoted CUB, which has been recently introduced in literature [6, 28] in order to represent ordinal data. In particular, the attention will be focussed on the problems related to the joint modelling of correlated preferences. This is a very common situation arising in consumer surveys since interviewees are generally unable to describe objectively their perceptions and then they tend to give hedonically based judgements rather than specific and thoughtful evaluations. For this reason, the ratings concerning connected items are often correlated and a multivariate approach is needed in order to understand the authentic map of consumer preferences.

In this respect, despite the CUB model has proved to be a successful statistical tool for describing the judgements on a single item in food quality studies [2,15,19], the problem of multivariate ordinal data representation within this class of models is still open.

This article moves a first step in such a direction. In particular, we introduce the Plackett's bivariate distribution with CUB margins and we discuss how the distribution parameters can be related to the subject's covariates. Finally, we illustrate an application of the proposed method to data from the EuroSalmon project [7].

2 The Plackett Distribution with CUB Margins

A bivariate Plackett random variable (X, Y) is characterised by the following joint cumulative distribution function:

$$H(x, y; \psi) = \frac{M(x, y) - [M^2(x, y) - 4\psi(\psi - 1)F(x)G(y)]^{1/2}}{2(\psi - 1)}, \tag{1}$$

where $\psi \in (0, \infty)$. Here, $F(x)$, $x \in S_x$, and $G(y)$, $y \in S_y$, are the pre-defined marginal distributions. Finally $M(x, y) = 1 + (F(x) + G(y))(\psi - 1)$ [20,31]. The parameter ψ is a measure of association between X and Y; in particular, $\psi = 1$ implies that X and Y are independent, whereas $\psi < 1$ and $\psi > 1$ refer to negative and positive association, respectively.

The distribution $H(x, y; \psi)$ satisfies the Fréchet bounds: $max\{F(x) + G(y) - 1, 0\} \leq H(x, y; \psi) \leq min\{F(x), G(y)\}$ where the lower and upper bounds are attained when $\psi \to 0$ and $\psi \to \infty$, respectively.

The original derivation of the Plackett distribution moves from considering the case of continuous margins and observing that one can always construct a joint cumulative distribution $H(x, y; \psi)$ having the property that when it is cut by lines parallel to the x and y axes, anywhere, the probabilities in the four quadrants viewed as a contingency table have a cross-product ratio which remains constant for any choice of the cutting points (x, y).

The problem goes back to the earlier contribution of Yule [33], Pearson [26], Pearson and Heron [27] who lively debated about the probability model with constant association coefficient, its capability to reproduce the bivariate Normal and, therefore, to model frequency surfaces in actual practice. The genesis of

the distribution that Plackett introduced later, in 1965, is strictly related to that debate. The differences with the Normal distribution are mainly due to the fact that Plackett's model with Normal margins is characterized by the skewness of the conditional distributions, the nonlinearity of its regressions and, by definition, by the fact that the invariance property of the association coefficient is not verified (as follows by earlier Pearson's results [26], Mosteller [25], and Goodman [10]).

Despite the constraints established over the possible shapes of $H(x, y; \psi)$, the Plackett's distribution family has found numerous applications with reference to various type of models for continuous and discrete data. In the latter case, the distribution describes the latent random variable from which a contingency table is derived by a discretization process. In this regards, overcoming the initial dimensional limit to the bivariate or trivariate random variables, Molenberghs [23] successfully extended it to the multivariate case. Furthermore, Molensberghs and Lesaffre [24] exploited that result and proposed a modelling approach to take into account the dependence of the global cross ratios from explanatory variables using the Dale model [5].

Although marginal distributions are usually supposed to be continuous, the derivation of the Plackett distribution holds with convenient premises in the discrete case. In this regards, Genest and Neslehova [9] extensively discussed the consequences of using copula distributions when the marginal distributions are not continuous. One of the main result, remarked in that article, concerns the fact that common association measures for a discrete copula distribution, such as Kendall τ and Spearman correlation, may depend on the margins, so that the ψ parameter can not be estimated or interpreted by means of those measures.

In the rest of this article, we assume that (X, Y) is a discrete bivariate random variable with support $S_{xy} = \{(x, y) : x = 1, 2, \ldots, m; y = 1, \ldots, m\}$ and the margins are described by CUB models. Moreover, we will refer to the probability mass distribution implied by (1) as:

$$h(x, y; \psi) = \sum_{r=0}^{1} \sum_{s=0}^{1} (-1)^{r+s} H(x - r, y - s; \psi).$$

2.1 The Marginal Distributions

We briefly introduce the definition of the CUB distribution of the random variable X (similarly, the following results apply to Y). In particular, $X \sim F(x; \boldsymbol{\theta}_x)$ with $\boldsymbol{\theta}_x = (\pi_x, \xi_x)'$ (and $Y \sim G(y; \boldsymbol{\theta}_y)$) is characterised by the distribution function:

$$F(x; \boldsymbol{\theta}_x) = \pi_x \sum_{j=1}^{x} \binom{m-1}{j-1} (1-\xi_x)^{j-1} \xi_x^{m-j} + (1-\pi_x) \frac{x}{m}, \qquad x = 1, 2, \ldots, m, \tag{2}$$

where $\pi_x \in (0, 1]$, $\xi_x \in [0, 1]$ and $m > 3$ [11]. Statistical properties and extensions of CUB models have been widely investigated in literature, as reviewed

by Corduas et al. [3]. In particular, CUB models provide a useful parametrization for a model-based clustering approach by means of the Kullback–Leibler divergence [1]. In addition, special situations due to shelter choices and the presence of over-dispersion in the data can be taken into account [12, 14]. The framework of CUB models can also be generalized in order to exploit intra-class correlation of respondents originated from a hierarchical structure of data [13]. Finally, modelling is feasible in practice since an efficient algorithm for the maximum likelihood estimation has been implemented in R [16, 29].

The formulation of the CUB probability mass distribution highlights the role of the two characterising parameters:

$$p(x;\boldsymbol{\theta}_x) = \pi_x \binom{m-1}{x-1} (1-\xi_x)^{x-1} \xi_x^{m-x} + (1-\pi_x)\frac{1}{m}, \qquad x = 1, 2, \ldots, m. \tag{3}$$

The weight π_x determines the contribution of the Uniform distribution in the mixture, therefore, $(1-\pi_x)$ is interpreted as a measure of the *uncertainty* which is intrinsic to any judgment. Besides, the parameter ξ_x, characterises the shifted Binomial distribution and $(1-\xi_x)$ denotes the degree of liking (*feeling*) expressed by raters with respect to the item.

The mixture distribution (3) is rather flexible since it is capable of describing distributions having very different shapes in terms of asymmetry and kurtosis [28]. For a given $\pi \in (0, 1]$, the peakedness increases as ξ approaches the borders of the parameter space whereas the distribution is symmetric for $\xi = 0.5$, negatively skewed when $\xi < 0.5$ and positively skewed when $\xi > 0.5$.

The graphical representation of the estimated CUB parameters in the unit square provides a useful tool in order to interpret the results from empirical analyses in terms of the above mentioned unobserved components: the uncertainty and the feeling.

Moreover, the influence of external factors, $\boldsymbol{w}$, in the final judgement can be taken into consideration by relating the model parameters π_x and/or ξ_x to significant *covariates* describing the raters, by means of a logistic link function [30]. For this aim, two further relations are added:

$$(\pi_x|\boldsymbol{w}_i) = [1+\exp(-\boldsymbol{w}_i\boldsymbol{\beta}_x)]^{-1}, \quad (\xi_x|\boldsymbol{w}_i) = [1+\exp(-\boldsymbol{w}_i\boldsymbol{\gamma}_x)]^{-1}$$

where, with an obvious notation, $\boldsymbol{\beta}_x$ and $\boldsymbol{\gamma}_x$ are the parameter vectors, and $\boldsymbol{w}_i$ is the i-th row of the regressor matrices associated to the i-th rater. In general, in the univariate case, the covariates affecting the π and ξ parameters may not be the same.

Denoting with $\boldsymbol{\delta}_x = (\boldsymbol{\beta}_x', \boldsymbol{\gamma}_x')'$ the parameters of the CUB margins, the presence of covariates leads to $F(x;\boldsymbol{\delta}_x)$ and $G(y;\boldsymbol{\delta}_y)$ so that the Plackett cumulative distribution (1) becomes $H(x, y; \psi, \boldsymbol{\delta}_x, \boldsymbol{\delta}_y)$ where $log(\psi) = \boldsymbol{w}_i\boldsymbol{\eta}$.

In the following section, we will denote the association parameter as $\psi(\boldsymbol{\eta})$ in order to highlight the dependency from the $\boldsymbol{\eta}$ parameters.

2.2 The Estimation

Given an observed sample of ordinal data, (x_i, y_i), for $i = 1, 2, \ldots n$, the estimation is performed by means of the IFM (Inference For the Margins) method [17, 18]. This is a method which is particularly useful for models where the dependence structure can be separated from univariate margins. As described above, this is the case of the Plackett distribution family where the marginal distributions originate from univariate CUB models and the parameter ψ can be estimated separately from margins. Specifically, the estimation is performed by means of a two-stage estimation procedure. The first stage involves maximum likelihood from univariate margins, and the second stage involves the maximum likelihood estimation of $\psi(\boldsymbol{\eta})$ with the univariate parameters held fixed from the first stage.

Thus, the model (1) with CUB margins is estimated as follows. Firstly, the log-likelihoods:

$$l_1(\boldsymbol{\delta}_x) = \sum_{i=1}^{n} \log\left(p(x_i; \boldsymbol{\delta}_x)\right); \quad l_2(\boldsymbol{\delta}_y) = \sum_{i=1}^{n} \log\left(p(y_i; \boldsymbol{\delta}_y)\right), \tag{4}$$

are separately maximised to get estimates: $\hat{\boldsymbol{\delta}}_x$ and $\hat{\boldsymbol{\delta}}_y$. Secondly, the log-likelihood:

$$l(\boldsymbol{\eta}; \hat{\boldsymbol{\delta}}_x, \hat{\boldsymbol{\delta}}_y) = \sum_{i=1}^{n} \log\left(h(x_i, y_i; \psi(\boldsymbol{\eta}), \hat{\boldsymbol{\delta}}_x, \hat{\boldsymbol{\delta}}_y)\right), \tag{5}$$

is maximised to get $\hat{\boldsymbol{\eta}}$ and then $\hat{\psi}$ by means of the previous mentioned relationship.

This two step procedure can be associated to the use of the jackknife method for estimation of the standard errors of the parameters and functions of the parameters. In such a way, the analytic derivatives to obtain the asymptotic covariance matrix of the vector of parameter estimates are not needed. Furthermore, since each inference function derives from some log-likelihood of a marginal distribution, the inference can be obtained by the existing EM algorithm implemented for the CUB models.

3 An Empirical Application

In order to exemplify the use of the proposed model for a real case study we consider the distribution of ratings that consumers expressed about two types of commercial smoked salmon. The data originates from a large survey which was carried out during the Euro-salmon project and which involved consumers' panels from various European countries [4, 7]. This data set was analysed by Piccolo and D'Elia [30] who estimated the univariate CUB models for the 30 products and introduced both product and subject features as covariates affecting the CUB parameters.

In the present study, we consider the ratings that 230 Italian consumers and 618 consumers from France, Belgium and Germany gave about two types of smoked

salmon, PROD19 and PROD24, using a 9 points Likert scale (1 = dislike extremely, 9 = like extremely). We expect consumers from the last three mentioned countries to be generally more trained to taste and recognise the quality of smoked salmon with respect to Italian consumers because of the higher frequency of use of such a product. Moreover, as for other food products, national cooking traditions and foodpairing might affect the development of consumer abilities to appreciate certain flavours and tastes.

According to the FAO Globefish report [8], Italy is one of the larger importing countries in Europe with about 6,000 tonnes of imported product in 2005. The national production, about 1,200 tonnes, is completely absorbed by local market. The per capita consumption is rather low (0.13 kg) compared with that of France (0.42 kg), Belgium (0.47 kg) and Germany (0.28 kg). Although UK was included in the consumer panel of the Eurosalmon project, the data concerning the ratings of UK interviewees were not considered in the following analysis. The present case study has been in fact focussed on contrasting Italian consumer ratings with those of respondents living in countries with high per capita consumption of smoked salmon. In this regard, we notice that per capita consumption of such a product in UK is only 0.12 kg, in spite of the fact that UK is the third producer of smoked salmon in Europe. In addition, we recall that Sémenou et al. [32] studied consumer behaviour across various European countries and highlighted that Italian consumers place great importance on product appearance (the orange colour and the translucent appearance) rather than on spoilage indicators confirming that Italian consumers are generally less trained to taste smoked salmon.

The first type of smoked salmon, PROD19, originated from Scotland, has the following sensory features: orange and homogeneous colour, medium intensity of odour, firm texture, medium level of salt. The second one, PROD24, originated from Norway, is less salty than the other product. Moreover, it has an intense orange and homogeneous colour, translucent appearance, low odour, firm and crunchy texture. The sensometric study proved that PRO19 is generally less appreciated than PROD24 [32].

Firstly, we considered the two products separately and we estimated the CUB models for each margin introducing a dummy variable representing the consumer country of origin in order to explain both the feeling and the uncertainty parameter (Table 1). Specifically, the dummy variable has value 1 when the respondent is French/Belgian/German, and is 0 when he/she is Italian.

The estimated CUB distributions clearly show how the shape of the marginal distribution varies depending on the respondents' country of origin (Fig. 1).

Specifically, Italian consumers express their judgments with lower uncertainty with respect to the other group. Moreover, considering the feeling parameter estimates, it is evident that PROD24 confirms its attractiveness to consumers with respect to PROD19. The value $(1 - \hat{\xi})$ of PROD24 is always higher than that estimated for PROD19.

However, French, Belgian and German consumers tend to distinguish better between the two products. Consumers from those countries show a higher degree of appreciation of PROD24 than that expressed by Italian consumers. Yet, the

Table 1 Results (country: ITA = Italy BFG = Belgium, France, Germany)

X=PROD19	$\hat{\beta}_{x0} = 0.394$	$\hat{\beta}_{x1} = -1.162$	$\hat{\pi}_{x.ITA} = 0.597$	$\hat{\pi}_{x.BFG} = 0.317$
	$(SE = 0.284)$	$(SE = 0.360)$		
	$\hat{\gamma}_{x0} = -0.293$	$\hat{\gamma}_{x1} = 0.513$	$\hat{\xi}_{x.ITA} = 0.427$	$\hat{\xi}_{x.O} = 0.555$
	$(SE = 0.086)$	$(SE = 0.140)$		
Y=PROD24	$\hat{\beta}_{y0} = 1.023$	$\hat{\beta}_{y1} = -1.072$	$\hat{\pi}_{y.ITA} = 0.7356$	$\hat{\pi}_{y.BFG} = 0.488$
	$(SE = 0.313)$	$(SE = 0.353)$		
	$\hat{\gamma}_{y0} = -0.644$	$\hat{\gamma}_{y1} = -0.239$	$\hat{\xi}_{y.ITA} = 0.344$	$\hat{\xi}_{x.BFG} = 0.293$
	$(SE = 0.095)$	$(SE = 0.102)$		
Conditional joint Distribution			$\hat{\psi}_{ITA} = 1.761$	$\hat{\psi}_{BFG} = 2.189$
			$(SE_{jack} = 0.343)$	$(SE_{jack} = 0.314)$

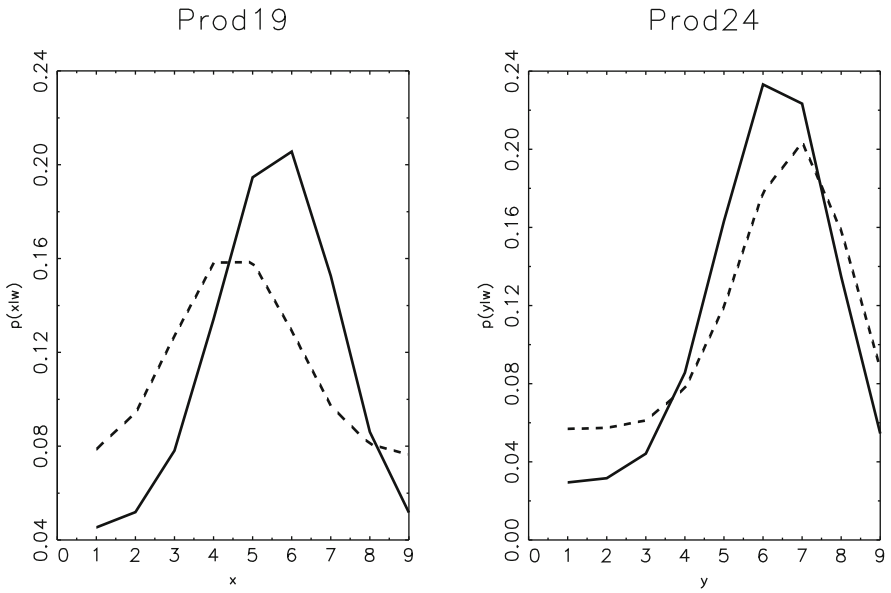

Fig. 1 CUB models (Italian consumers = *solid line*, Others = *dashed line*)

reverse statement applies when PROD19 is considered. Notwithstanding the positive judgement, French, Belgian and German consumers are less attracted by that product than Italian consumers.

Thus, we jointly model the ratings concerning the appreciation of the two products and again we consider the consumer country of origin as a meaningful covariate to describe the preferences.

In Fig. 2 the conditional joint probability distributions of the ratings about the two considered items is illustrated given the country of origin. The values of the estimated ψ parameters are both positive and rather close, although the consumers

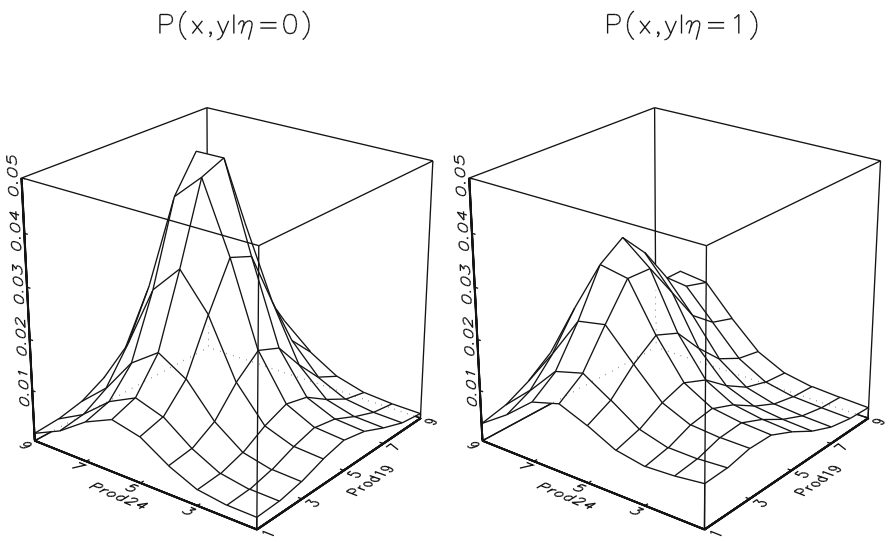

Fig. 2 Joint probability distribution of PROD19 and PROD24 (*left panel*: Italian consumers; *right panel*: Others

from Italy seems to show a lower level of association between the ratings that they express.

Furthermore, we evaluated the probability that a consumer will assign a rate over 5 to both products. The probability that Italian consumers are satisfied with both types of smoked salmon is 0.35 whereas the analogous probability for the other group is only 0.28. This is the effect of the appreciation that Italian consumers generally show towards both the products whereas consumers from the other countries seem more critical.

Concluding Remarks

The results described in the previous section encourage further studies on the proposed model for correlated ordinal data. On the one hand, CUB models represent an effective statistical tool which helps to identify the role of two latent components: the *uncertainty* of respondents in rating product attributes and the strength of *attraction* each attribute arouses. On the other hand, the joint modelling of ratings allows the study of the bonds that connect consumer preferences about alternative products providing further insights into consumer behaviour.

Further research is needed in order to implement the approach to the k-variate case. The Plackett distribution has in fact been generalised by Molenberghs [23]. However, the definition of such a distribution becomes computationally cumbersome for high-dimensional applications and because of the discrete nature of the involved random variables.

Further attention may also be paid to alternative modelling approaches, such as random effect modelling, which exploits the possible hierarchical structure of the data in order to take the intra-correlation at cluster level into account.

References

1. Corduas, M.: Assessing similarity of rating distributions by Kullback-Leibler divergence. In: Fichet, B., et al. (eds.) Classification and Multivariate Analysis for Complex Data Structures, pp. 221–228. Springer, Heidelberg (2011)
2. Corduas, M., Cinquanta, L., Ievoli, C.: The importance of wine attributes for purchase decisions: a study of Italian consumers perception. Food Qual. Prefer. **28**, 407–418 (2013)
3. Corduas, M., Iannario, M., Piccolo, D.: A class of statistical models for evaluating services and performances. In: Bini, M., et al. (eds.) Statistical Methods for the Evaluation of Educational Services and Quality of Products, pp. 99–117. Physica, Heidelberg (2009)
4. Courcoux, P., Qannari, E.M., Schlich, P.: Sensometric workshop: segmentation of consumers and characterization of cross-cultural differences. Food Qual. Prefer. **17**, 658–668 (2006)
5. Dale, J.R.: Global cross-ratio models for bivariate, discrete, ordered responses. Biometrics **42**, 909–917 (1986)
6. D'Elia, A., Piccolo, D.: A mixture model for preferences data analysis, Comp. Stat. & Data Anal. **49**, 917–934 (2005)
7. Eurosalmon Final Report: Improved quality of smoked salmon for the European consumer. Final Report for the EC "Quality of Life and Management of Living Resources" Programme (2004). www.mmedia.is/matra/eurosalmon
8. FAO: Commodity update: salmon. Globefish Report, Fisheries Department (2007). www.globefish.org
9. Genest, C., Neslehova, J.: A primer on copulas for count data. Astin Bull. **37**, 475–515 (2007)
10. Goodman, L.A.: Models and the bivariate normal for contingency tables with ordered categories. Biometrika **68**, 347–355 (1981)
11. Iannario, M.: On the identifiability of a mixture model for ordinal data. METRON **LXVIII**, 87–94 (2010)
12. Iannario, M.: Modelling shelter choices in a class of mixture models for ordinal responses. Stat. Methods Appl. **21**, 1–22 (2012a)
13. Iannario, M.: Hierarchical CUB models for ordinal variables. Commun. Stat. Theory Methods **41**, 3110–3125 (2012b)
14. Iannario, M.: Modelling uncertainty and overdispersion in ordinal data. Commun. Stat. Theory Methods **43**, 771–786 (2014)
15. Iannario, M., Piccolo, D.: CUB models: statistical methods and empirical evidence. In: Kenett, R.S., Salini, S. (eds.) Modern Analysis of Customer Surveys: With Applications Using R, pp. 231–258. Wiley, Chichester (2012)
16. Iannario, M., Piccolo, D.: A Short Guide to CUB 3.0 Program (2013). www.researchgate.net
17. Joe, H.: Asymptotic efficiency of the two-stage estimation method for copula-based models. J. Multivar. Anal. **94**, 401–419 (2005)
18. Joe, H., Xu, J.J.: The estimation method of inference functions for margins for multivariate models. Report No. 166, Department of Statistics, University of British Columbia (1996)
19. Manisera M., Piccolo D., Zuccolotto P.: Analyzing and modelling rating data for sensory data in food industry. Quaderni di Statistica **13**, 69–82 (2011)
20. Mardia, K.V.: Families of Bivariate Distributions. Griffin, London (1970)
21. McCullagh, P.: Regression models for ordinal data (with discussion). J. R. Stat. Soc. Ser. B **42**, 109–142 (1980)
22. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman & Hall, London (1989)

23. Molenberghs, G.: A multivariate Plackett distribution with given marginal distributions. Universitaire Instelling Antwerpen, No. 92/33 (1992)
24. Molenberghs, G., Lesaffre, E.: Marginal modelling of correlated ordinal data using multivariate Plackett distribution. J. Am. Stat. Assoc. **89**, 633–644 (1994)
25. Mosteller, F.: Association and estimation in contingency tables. J. Am. Stat. Assoc. **63**, 1–28 (1968)
26. Pearson, K.: On the theory of association. Biometrika **9**, 159–315 (1913)
27. Pearson, K., Heron, D.: Note on the surface of constant association. Biometrika **9**, 534–537 (1913)
28. Piccolo, D.: On the moments of a mixture of uniform and shifted binomial random variables. Quaderni di Statistica **5**, 85–104 (2003)
29. Piccolo, D.: Observed information matrix for MUB models. Quaderni di Statistica **8**, 33–78 (2006)
30. Piccolo, D., D'Elia, A.: A new approach for modelling consumers' preferences. Food Qual. Prefer. **19**, 247–259 (2008)
31. Plackett, R.L.: A class of bivariate distributions. J. Am. Stat. Assoc. **60**, 516–522 (1965)
32. Semenou, M., Courcoux, P., Cardinal, M., Nicod, H., Ouisse, A.: Preference study using a latent class approach. Analysis of European preferences for smoked salmon. Food Qual. Prefer. **18**, 720–728 (2007)
33. Yule, G.U.: On the methods of measuring association between two attributes. J. R. Stat. Soc. **75**, 579–642 (1912)

Modelling Job Satisfaction of Italian Graduates

Stefania Capecchi and Silvia Ghiselli

Abstract

Different models have been implemented to observe worker conditions, abilities, leadership, decision-making attitudes and other related concerns. This paper aims to investigate the job satisfaction of a large sample of Italian graduates with a model-based approach derived by a mixture distribution. Sample data have been collected in the 2010 AlmaLaurea survey on graduates employment conditions, 5 years after their degree. We highlight several issues which are effective in assessing the performance of the academic system and detecting graduates' responses towards labour market using CUB models approach. A specific contribution of this paper consists in emphasizing the possibility to achieve immediate interpretation and visualization of the main relationships between responses concerning job satisfaction and characteristics of the interviewees.

Keywords

Job satisfaction • Ordinal data • CUB models

1 Introduction

The relationship between job satisfaction and worker characteristics has been heavily researched over the years in various domains such as Sociology, Economics and Management Sciences [28], mostly in the field of industrial-organizational

S. Capecchi (✉)
Department of Political Sciences, University of Naples Federico II, Via L. Rodinò, 22, 80138 Naples, Italy
e-mail: stefania.capecchi@unina.it

S. Ghiselli
Almalaurea Inter-University Consortium, Viale Masini, 36, 40126 Bologna, Italy
e-mail: silvia.ghiselli@almalaurea.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_7, Published online: 12 November 2014

Psychology and in the goal-setting theory [27]. The positive link between workers' satisfaction and their productivity levels initially motivated these analyses. Since the early 1970s many definitions of job satisfaction have been proposed and different models have been introduced [30]. A frequently quoted statement explicates job satisfaction through a behavioural variable: "(...) a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences" [19].

In labour market dynamics job satisfaction has become a leading determinant of productivity, mobility, unionism, etc., and can be considered both an explicative variable of job performance and a dependent one, based on individual as well objective conditions [8]. Though job satisfaction could be analyzed as an element of job performance it cannot be related only to incentives, especially to economic ones. Indeed, the incentives sometimes act as counterproductive [26], so job satisfaction has to be investigated as a relevant issue for individuals' general life well-being [3, 18]. To study such a phenomenon, a collection of adequate data is needed so to analyze individual attitudes and satisfaction through an ordinal scale, according to the level of agreement of the respondent. Questions used to measure job satisfaction are usually related to an overall dimension and also to several specific items (such as income level, education/job mismatches, overqualification, and so on).

The main objective of the paper is to introduce a model-based approach to job satisfaction to investigate and check the significance of important relationships of the responses with subjects' covariates. In this respect, graphical devices are introduced to effectively visualize different facets of the estimated models.

The paper is organized as follows: in the next section, the notation and main properties of CUB models are introduced. After a brief mention about data collection, Sect. 3 examines global job satisfaction and its components with a special reference to their relationships. Then, in Sect. 4 we consider the effect of some subjects' covariates on the expressed satisfaction, with respect to final grades and typology/sector of work. Some concluding remarks end the work.

2 CUB Models

In recent years remarkable advances have been made in the analysis of categorical ordinal data [1, 29] and most of them rely on Generalized Linear Models [21, 22]. A different paradigm has been proposed with the introduction of CUB models [24]: a comparison between these approaches, with reference to job satisfaction, has been conducted by [9]. Further extensions have been exploited to take into account real data problems as in case of hierarchical, *shelter* and overdispersion effects [13–15]. These models have been used with efficacy in presence of ordinal data expressing preference and evaluation, and collected in different circumstances, as discussed by [5] and [16], among others. Hereafter, we apply CUB models to explain the responses about job satisfaction and measure their possible relationships with selected covariates.

CUB models are generated by a class of discrete probability distributions which takes into account two latent components pertaining to the response, denoted as

feeling and uncertainty. More specifically, these latent variables are modelled as a shifted Binomial and a discrete Uniform random variable, respectively.

Let the response variable Y take values in $\{1, \ldots, m\}$, where $m > 3$ for identifiability constraints [11]. Then, we collect a sample $(y_1, y_2, \ldots, y_n)'$ where y_i is the rating expressed by the ith subject, for $i = 1, 2, \ldots, n$, on an m point scale. Moreover, let $\mathbf{x}_i$ and $\mathbf{w}_i$, $i = 1, \ldots, n$, be subjects' covariates selected for explaining feeling and uncertainty, respectively.

The general formulation of a CUB model is:

$$Pr(Y = y|\mathbf{x_i}, \mathbf{w}_i) = \pi_i \binom{m-1}{y-1} (1-\xi_i)^{y-1} \xi_i^{m-y} + (1-\pi_i)\frac{1}{m}, \qquad (1)$$

with $y = 1, 2, \ldots, m$, and two logistic links for the systematic components:

$$logit(\pi_i) = \mathbf{x_i}\boldsymbol{\beta}; \qquad logit(\xi_i) = \mathbf{w_i}\boldsymbol{\gamma}; \qquad i = 1, \ldots, n; \qquad (2)$$

where $logit^{-1}(z) = \left[1 + e^{-z}\right]^{-1}$, for any real z.

We support the interpretation of the parameters (π, ξ) in terms of uncertainty and feeling components, respectively. Each respondent answers thoughtfully or with a completely uncertain behaviour, with propensities measured by π and $1 - \pi$, respectively. As a consequence, $(1-\pi)$ is a measure of uncertainty. In a rating survey $(1 - \xi)$ may be considered as a measure of agreement to the item. The meaning of ξ changes according to the empirical framework since that parameter is determined by the frequency of responses with low degree of agreement (liking, approval). In general, the ξ parameter has been related to degree of perception, measure of closeness, assessment of proficiency, rating of concern, index of selectiveness, pain threshold, personal confidence, subjective probability, and so on. In our context, $(1 - \xi)$ is the level of satisfaction of the respondent.

Although ξ and π may be related to the mean level and variability of the ordinal response, respectively, such an interpretation may be considered as a partial one and, in some circumstances, also misleading. Indeed, the expectation of a CUB random variable is a non-linear function of both parameters and thus infinitely many pairs (π, ξ) would give the same mean value. In addition, if we compute the variance or the mean difference of the distribution (1) we get a non-monotone relationship with respect to π. As a consequence, the parameter π cannot be strictly related to a dispersion aspect of the random variable. Instead, if we measure the heterogeneity of the CUB distribution with respect to $(1 - \pi)$ we get a regular increase over the whole parameter space and for any fixed ξ. In fact, it is possible to prove that the parameter π is formally related to the heterogeneity Gini index [10, 12].

For a given m, there is one-to-one correspondence among CUB probability distributions and the parameters (π, ξ). Then, we may represent each CUB model as a point in the unit square with coordinates $(1-\pi,\ 1-\xi)$ corresponding to uncertainty and satisfaction, respectively, in our study. We can summarize several estimated models as a collection of points in the parameter space and check for possible

effects of covariates when time and circumstances are different. In addition to this representation, other graphical devices are convenient to emphasize the relationships between uncertainty or feeling and the relevant covariates (as we experiment in Sects. 3–4). Finally, if prediction issues have to be pursued, selected profiles of distribution functions for given covariates of respondents may also be plotted.

From an inferential point of view, the estimation procedure is obtained by the Maximum Likelihood method exploiting the EM algorithm [23], as specified for CUB models in [25]. Detailed information on computational aspects can be found in [17]; the R program used in the present study is freely available.

The validation of CUB models with covariates (for uncertainty and/or feeling) is achieved by considering the parameter significance and the increase of the log-likelihood function when covariates are inserted in the standard model. In that regard, Wald and likelihood ratio tests may be exploited; often, *AIC* and *BIC* criteria have been used.

From a fitting point of view, if the sample size (n) is moderate or large, as it happens in our data set, traditional X^2 measures are not effective since they tend to reject adequate fitting even in case of empirical and estimated distributions nearly overlapping. Thus, we prefer relying on dissimilarity indexes or using an observed/predicted table where prediction of ordinal data have been obtained by some location index (as modal values, median, expectation).

3 Global Satisfaction and Its Components

The analyzed data set is a subset of the archive of AlmaLaurea Inter-University Consortium, which now covers 64 Italian universities and accounts for about 78 % of Italian graduates well distributed from a geographical point of view. In a broad sense, it is representative of the whole population of graduates. Our research concerns the survey carried out in 2010 and refers to 59 % of all graduates in the period May–August 2005 [2, 6]. In addition to a global satisfaction rating, data set includes responses to several facets of job satisfaction as listed in Table 1 and also covariates concerning personal, socio-demographic and economic variables.

The survey refers to the students who obtained their degree according to the former Italian university education system and are employed after 5 years from

Table 1 Job satisfaction items

1. Security of the job	8. Involvement in the decisional processes
2. Coherence with studies	9. Flexibility of time
3. Acquisition of professionalism	10. Availability of free time
4. Prestige	11. Workplace
5. Connection with cultural interests	12. Relationships with co-workers
6. Social utility	13. Expectation of future gains
7. Independence or autonomy in the job	14. Perspectives of career

Table 2 Average of job satisfaction and items, with corresponding missing values

Components	*Missing*	*Average*		*Components*	*Missing*	*Average*
Global satisfaction	5	7.57				
Security	5	6.74		Involvement	15	7.54
Coherence	3	6.92		Flexibility	14	7.06
Professionalism	14	7.66		Free time	10	6.19
Prestige	20	7.10		Workplace	63	7.46
Cultural interests	4	7.27		Co-workers	581	8.02
Social utility	37	7.37		Future gains	117	6.53
Autonomy	12	7.83		Career perspectives	139	6.57

their graduation. All statistical analyses have been performed on 17,387 validated questionnaires where the job satisfaction items are based on a modified 9-point response scale (1=very dissatisfied, 9=very satisfied).

In Table 2, for each item, the number of missing data and the average of job satisfaction responses are reported. Notice that the large number of missing data for the item "Relationships with co-workers" is mainly due to single workers. Anyway, the percentage of missing values is not so relevant. In this respect, univariate CUB models for each item on the complete subset have been estimated.
Since, for a given m, ML estimates are invariant with respect to sample size, it is possible to visualize the estimated CUB models on the same parameter space, as in Fig. 1. This analysis confirms that the estimated CUB models are not coincident (that is the differences among estimated parameters are always significant); in fact, even if we add confidence regions on the parametric representation in Fig. 1 they never overlap. All items present a high level of satisfaction since, for all estimated models, it turns out that $(1 - \hat{\xi}) > 0.744$. Instead, uncertainty is more dispersed since $(1 - \hat{\pi}) \in (0, 0.6]$.

In this representation, "Availability of free time", "Expectation of future gains" and "Perspectives of career" get the lowest satisfaction. "Security of the job", "Relationships with co-workers" and "Coherence with studies" are considered very satisfying, comparatively. With regard to the indecision in the answers, we observe that respondents are more uncertain about "Security of the job", "Coherence with studies" and "Availability of free time". Finally, we observe that the global satisfaction is not a mere average of expressed satisfaction for the different facets.

Hitherto, analyses have been pursued by comparing the univariate distributions of ratings expressed for both global and facets of job satisfaction. The closeness of points in the parameter space (as depicted in Fig. 1) implies a similar shape of distribution but not necessarily a strong relationship among the corresponding facets. Indeed, a multivariate analysis based on CUB models is not widespread yet [7] whereas a multivariate analysis adapted to ordinal data should be applied, as pursued by [4, 20], for instance.

However, in the context of CUB models, it is possible to interpret the response to global satisfaction as the multifacet result of the different features, and to introduce

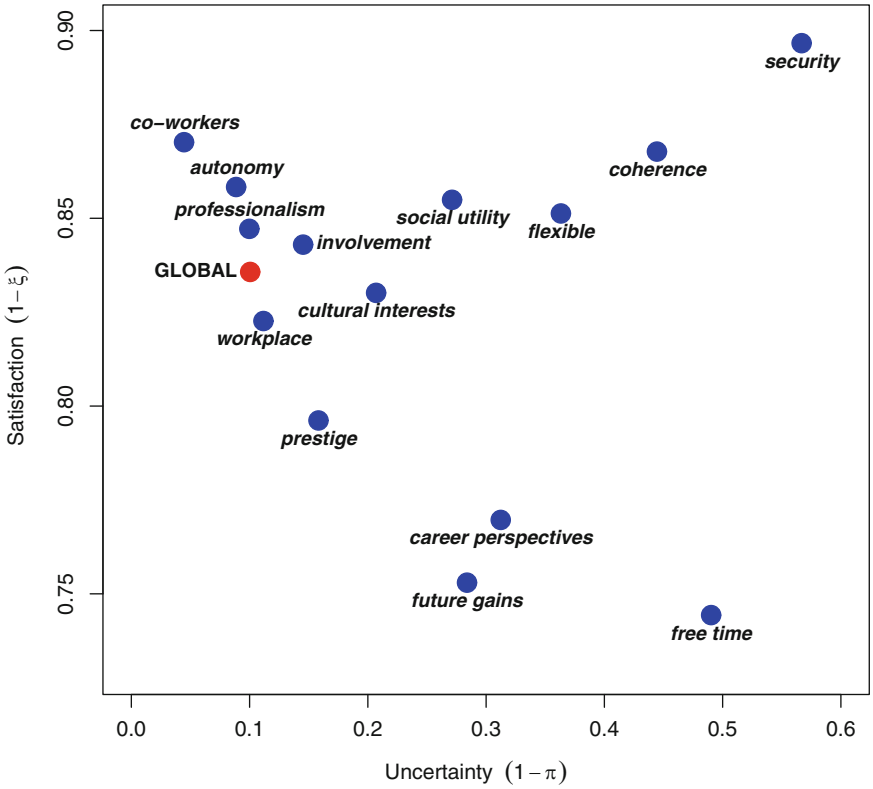

Fig. 1 Estimated CUB models of global job satisfaction and 14 items

these aspects as covariates in the model. Assuming a constant uncertainty in this model (this approximation may be fairly accepted with our data set), if we denote the global satisfaction and the items with Y and $Y^{(j)}$, $j = 1, 2, \ldots, 14$, respectively, we fit a model (1) where the specification (2) is given by:

$$\pi_i = \pi; \qquad logit(\xi_i) = \gamma_0 + \sum_{j=1}^{14} \gamma_j \, Y_i^{(j)}; \qquad i = 1, \ldots, n\,. \tag{3}$$

Adopting a stepwise strategy, based on both significance of parameters and increase in log-likelihood functions, we sequentially upgrade a standard CUB model (without covariates) for job satisfaction which presents a log-likelihood function estimated at maximum as $-27{,}006$, with $n = 16{,}547$ complete observations.

It turns out that all items are largely significant to explain the global satisfaction but "Flexibility of time" and "Perspectives of career": such a model increases log-likelihood function up to $-21{,}300$. In addition, since all covariates have the same range, the estimated γ_j parameters express the importance of a single component to contribute to the global job satisfaction. In Table 3 we report the facets which

Table 3 A CUB model of global job satisfaction as a function of significant items

Facets	Estimates $\hat{\gamma}_j$	Facets	Estimates $\hat{\gamma}_j$	Facets	Estimates $\hat{\gamma}_j$
Professionalism	−0.100 *(0.007)*	Involvement	−0.040 *(0.006)*	Social utility	−0.034 *(0.004)*
Prestige	−0.127 *(0.007)*	Co-workers	−0.054 *(0.007)*	Coherence	−0.023 *(0.004)*
Autonomy	−0.145 *(0.006)*	Future gains	−0.065 *(0.005)*	Security	−0.047 *(0.003)*
Cultural interests	−0.067 *(0.006)*	Workplace	−0.033 *(0.005)*	Free time	−0.011 *(0.004)*

Asymptotic standard errors in parentheses

are significant to explain the global job satisfaction. In this model, we get $\hat{\pi} = 0.993\,(0.001)$ and $\hat{\gamma}_0 = 3.735\,(0.060)$. The facets are ordered according to their contribution to improve the log-likelihood function.

Thus, "Acquisition of professionalism", "Prestige", "Independence or autonomy in the job", "Connection with cultural interests", "Involvement in the decisional processes", "Relationships with co-workers" are the most relevant features whereas "Coherence with studies", "Security of the job", "Availability of free time" exert a minor (although significant) impact. Notice that "Expectation of future gains" is scored as the 7th.

Model (3) confirms the theory that job satisfaction is a very complex latent variable, not strictly related to monetary compensation. Its expression is the final outcome of several causes (subjective and objective, related to workplace and colleagues), each of them contributing with separate but significant impacts.

4 Covariates Effects on Job Satisfaction

To better understand the expressed job satisfaction, we present a selection of results focused on the effects of gender, grades and typology of work.

First, we check if Gender exerts a significant effect on the responses by means of model (1) where:

$$logit(\pi_i) = \beta_0 + \beta_1\, Gender_i\,; \quad logit(\xi_i) = \gamma_0 + \gamma_1\, Gender_i\,; \quad i = 1, 2, \ldots, n$$

and $Gender_i = 0$ or 1 if the ith subject is men or woman, respectively.

In Table 4 we list the sign (if significant) of the estimated $\hat{\beta}_1$ and $\hat{\gamma}_1$ parameters to measure the impact of Gender on uncertainty and satisfaction, respectively. It turns out that the significant effect of uncertainty is greater for women (except for "Social utility" and "Autonomy") and their satisfaction is significantly greater for the facets "Security", "Coherence", "Cultural interests", "Social utility", "Free time", and "Workplace".

Let $SC = log(Grade)$ be a score variable for the final grade; for convenience, we denote *summa cum laude* as 112. Then, the best estimated CUB model with score as a covariate implies the following relationship for the level of satisfaction:

$$1 - \xi_i = logit^{-1}\big(-95.547 + 40.996\, SC_i - 4.472\, SC_i^2\big), \qquad i = 1, 2, \ldots, n.$$

Table 4 Sign of estimated parameters for Gender covariate in CUB models for all items

Components	*Uncertainty*	*Satisfaction*	*Components*	*Uncertainty*	*Satisfaction*
Security	−	−	Involvement	−	+
Coherence	∘	−	Flexibility	−	+
Professionalism	−	∘	Free time	+	−
Prestige	−	+	Workplace	−	−
Cultural interests	∘	−	Co-workers	−	∘
Social utility	+	−	Future gains	−	+
Autonomy	−	+	Career perspectives	−	+

− negative effect, + positive effect, ∘ non statistically significant

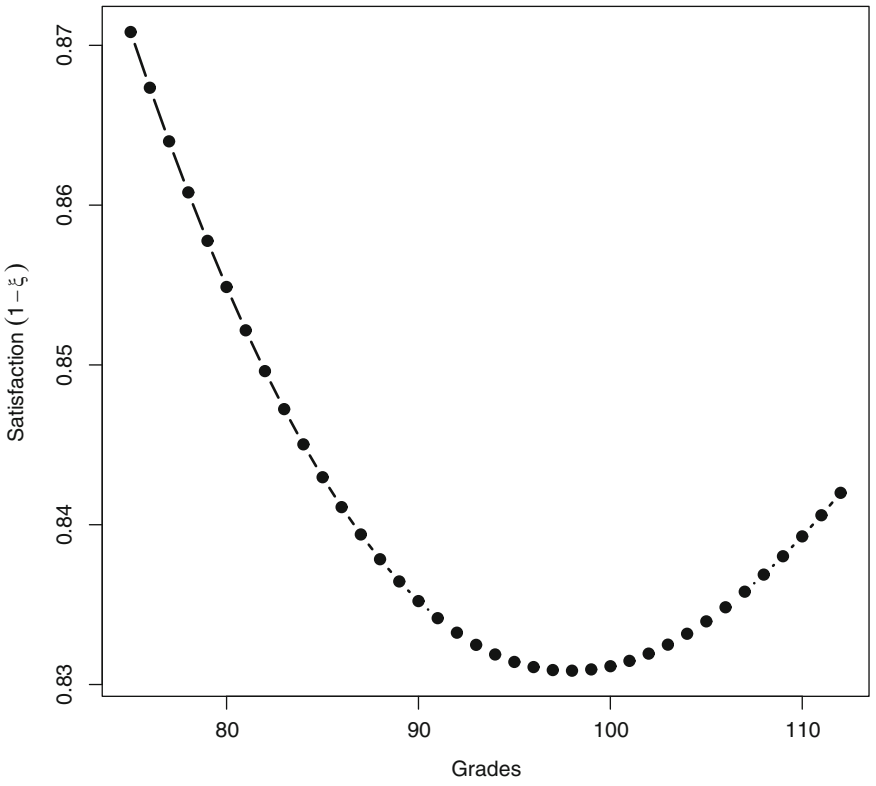

Fig. 2 Global satisfaction as a function of grades, estimated by a CUB model

Gender is also a significant covariate for determining a shift in the uncertainty of the responses (women are more uncertain); however, the shape of the relationships between satisfaction and final grade is the same for both genders.

Figure 2 illustrates the level of the job satisfaction as function of final grades. As expected, the function is asymmetric since students receiving low scores are generally older and conclude their University training with some difficulties: when

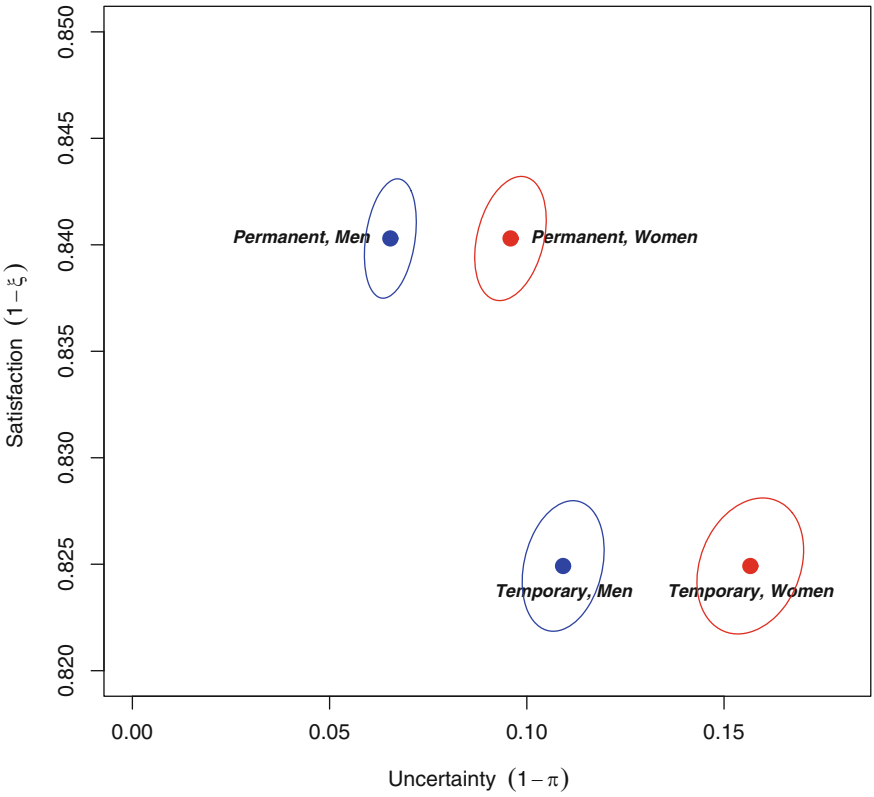

Fig. 3 CUB models for job satisfaction as functions of contract typology and gender

they get a degree, they are often already employed and such a result may improve their career. As a consequence, they manifest a greater satisfaction in the job despite the achieved low scores. On the contrary, students receiving very high scores are often clever in their professional abilities; some of them obtain jobs adequate to their skills and express a greater job satisfaction. Otherwise, the overqualification can be a possible explanation of a lower satisfaction expressed by respondents with high degree scores. Ceteris paribus, graduates with scores around 97 show the minimum job satisfaction. These general considerations have been empirically checked in the selected data set; specifically, the correlation between age at degree and grade is -0.21 .

In Fig. 3, we describe how contract typology affects the expressed satisfaction. In particular a permanent position, that allows for a sense of security for long-term individual and family planning (as declared by 69 % of the sample), generates a higher level of satisfaction for both men and women. The significance of this covariate is verified by defining a dummy covariate: “permanent” versus all other

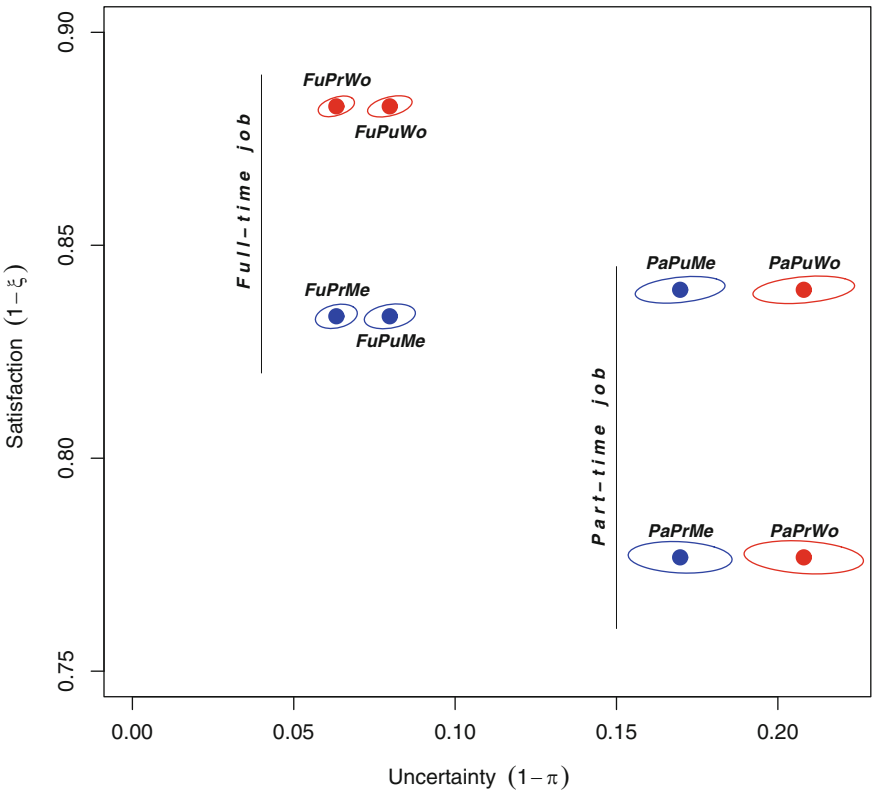

Fig. 4 CUB models for job satisfaction as functions of contract typology, sector and gender

typologies of job. We found that job security affects both uncertainty and feeling parameters (people with permanent job are less uncertain and more satisfied as clearly shown in Fig. 3) whereas Gender is significant only for uncertainty in the expressed satisfaction (women are more uncertain in any case). The well separated 95 % confidence regions plotted around estimated models suggest significant differences among genders, sector and typology of contract.

Figure 4 summarizes the estimated CUB models of job satisfaction when covariates are Full-time/Part-time (denoted on the plot as Fu/Pa), Public/Private (Pu/Pr) and Gender (Wo/Me). It is evident that the main discrimination is Full-time versus Part-time, with the second one characterized by a greater uncertainty. Regarding Full-time jobs, the satisfaction expressed by women is higher, whereas in Private sector the effect of Gender is evident: women are a bit more satisfied. Finally, in all cases, to work in the Public sector generates more satisfaction than in the Private one.

Concluding Remarks

The results so far discussed motivate the flexibility and versatility of CUB models as a different paradigm for the analysis of job satisfaction data. More specifically, a remarkable added value of the approach is the possibility to represent the estimated models in a proper parameter space and to see how they are modified with respect to subgroups and/or covariates by using several different graphical displays (for instance, the study of feeling as a function of selected covariates, the location of estimated CUB models related to different characteristics of respondents, and so on).

In any case, the consideration that all models contain an uncertainty component is a relevant one since this presence may alter the interpretation of the observed data if we summarize all information by some average or other location indexes.

Finally, if we adhere to the logic of CUB models, we are implicitly accepting that all data related to job satisfaction may be effectively summarized by just few parameters within a specific class of discrete mixture distributions, so to improve interpretation, prediction and classification.

Acknowledgements This research has been partly supported by an agreement between Department of Political Sciences, University of Naples Federico II and AlmaLaurea Inter-University Consortium, Bologna, and by a STAR Project grant (CUP E68C13000020003).

References

1. Agresti, A.: Categorical Data Analysis, 3rd edn. Wiley, New York (2013)
2. AlmaLaurea: Condizione Occupazionale dei Laureati, XIII Indagine 2010, www.almalaurea.it (2011)
3. Blanchflower, D.G., Oswald, A.J.: Well-being over time in Britain and the USA. J. Public Econ. **88**, 1359–1386 (2004)
4. Brentari, E., Golia, S., Manisera, M.: Models for categorical data: A comparison between the Rasch model and nonlinear principal component analysis. Statistica Applicazioni **5**, 53–77 (2007)
5. Capecchi, S., Piccolo, D.: Modelling approaches for ordinal data: the case of orientation service evaluation. Quaderni di Statistica **12**, 99–124 (2010)
6. Capecchi, S., Iannario, M., Piccolo, D.: Modelling Job Satisfaction in AlmaLaurea Surveys, AlmaLaurea Working Papers n. 56 (2012)
7. Corduas, M.: Modelling correlated bivariate ordinal data with CUB marginals. Quaderni di Statistica **13**, 109–119 (2011)
8. Freeman, R.B.: Job satisfaction as an economic variable. Am. Econ. Rev. **68**, 135–141 (1978)
9. Gambacorta, R., Iannario, M.: Measuring job satisfaction with CUB models. LABOUR **27**, 198–224 (2013)
10. Iannario, M.: Selecting feeling covariates in rating surveys. Ital. J. Appl. Stat. **20**, 121–134 (2009)
11. Iannario, M.: On the identifiability of a mixture model for ordinal data. METRON **68**, 87–94 (2010)
12. Iannario, M.: Preliminary estimators for a mixture model of ordinal data. Adv. Data Anal. Classif. **6**, 163–184 (2012a)

13. Iannario, M.: Modelling *shelter* choices in a class of mixture models for ordinal responses. Stat. Methods Appl. **21**, 1–22 (2012b)
14. Iannario, M.: Hierarchical CUB models for ordinal variables. Commun. Stat. Theory Methods **41**, 3110–3125 (2012c)
15. Iannario, M.: Modelling uncertainty and overdispersion in ordinal data. Commun. Stat. Theory Methods **43**, 771–786 (2013)
16. Iannario, M., Piccolo, D.: CUB models: Statistical methods and empirical evidence. In: Kenett, R.S., Salini, S. (eds.) Modern Analysis of Customer Surveys: With Applications Using R, pp. 231–258. Wiley, Chichester (2012)
17. Iannario, M., Piccolo, D.: A Short Guide to CUB 3.0 Program, Technical report available from Authors (2013)
18. Judge, T.A., Watanabe, S.: Another Look at the Job Satisfaction - Life Satisfaction Relationship. J. Appl. Psychol. **78**, 939–948 (1993)
19. Locke, E.: The nature and causes of job satisfaction. In: Dunnette, M.D. (eds.) Handbook of Industrial and Organizational Psychology, pp.1297–1349. Rand McNally, Chicago (1976)
20. Manisera, M., Van der Kooij, A.J., Dusseldorp, E.M.L.: Identifying the component structure of satisfaction scales by nonlinear principal components analysis. Qual. Technol. Quant. Manag. **7**, 97–115 (2010)
21. McCullagh, P.: Regression models for ordinal data (with discussion). J. R. Stat. Soc. Ser. B **42**, 109–142 (1980)
22. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall, London (1989)
23. McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, New York (2008)
24. Piccolo, D., On the moments of a mixture of uniform and shifted binomial random variables. Quaderni di Statistica **5**, 85–104 (2003)
25. Piccolo, D.: Observed information matrix for MUB models. Quaderni di Statistica **8**, 33–78 (2006)
26. Pugno, M., Depedri, S.: Job performance and job satisfaction an integrated survey. Economia Politica **27**, 175–210 (2010)
27. Spector, P.E.: Measurement of human service staff satisfaction: Development of the job satisfaction survey. Am. J. Commun. Psychol. **13**, 693–713 (1985)
28. Spector, P.E.: Job Satisfaction: Application, Assessment, Causes, and Consequences. SAGE, Los Angeles (1997)
29. Tutz, G.: Regression for Categorical Data. Cambridge University Press, Cambridge (2012)
30. Vroom, V.H., Deci, E.L.: The stability of post-decision dissonance: a follow-up study of the job attitudes of business school graduates. Organ. Behav. Hum. Perform. **6**, 36–49 (1971)

Identification of Principal Causal Effects Using Secondary Outcomes

Fabrizia Mealli, Barbara Pacini, and Elena Stanghellini

Abstract

Unless strong assumptions are made, identification of principal causal effects in causal studies can only be partial and bounds (or sets) for the causal effects are established. In the presence of a secondary outcome, recent results exist to sharpen the bounds that exploit conditional independence assumptions (Mealli and Pacini, J. Am. Stat. Assoc. 108:1120–1131, 2013). More general results, though not embedded in a causal framework, can be found on concentration graphs with a latent variable (Stanghellini and Vantaggi, Bernoulli 19:1920–1937, 2013). The aim of this paper is to establish a link between the two settings and to show that adapting results contained in the latter paper can help achieving identification of principal casual effects in studies with more than one secondary outcome. An empirical illustrative example is also provided, using data from a real social job training experiment.

Keywords

Binary latent variable models • Causal estimands • Graphical models • Identification • Latent class • Principal stratification

F. Mealli
Dipartimento di Statistica, Informatica, Applicazioni "Giuseppe Parenti", Università di Firenze, Firenze, Italy
e-mail: mealli@disia.unifi.it

B. Pacini (✉)
Dipartimento di Scienze Politiche, Università di Pisa, Pisa, Italy
e-mail: barbara.pacini@sp.unipi.it

E. Stanghellini
Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Perugia, Italy
e-mail: elena.stanghellini@stat.unipg.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_15, Published online: 28 October 2014

1 Introduction

Causal inference studies, including randomized clinical trials, are often subject to possible post-assignment complications. Those include noncompliance to assigned treatment, censoring of primary outcomes due to death, missing outcome data, and can be thought of as selection processes not under experimental control.

If the causal problem is formalized using the potential outcome framework, also known as the Rubin Causal Model [12], the presence of intermediate variables can be represented by means of a stratification of units into latent classes (also named principal strata, [5]) defined by the joint value of the intermediate variables under each possible treatment level.

The goal of the causal analysis is usually that of identifying and estimating the so called principal causal effects, that is, contrasts of features of the distribution of potential outcomes under different treatment levels, conditional on a latent stratum. The identification of the outcome distributions conditional on the latent classes is thus crucial. Without strong assumptions, such as exclusion restriction assumptions (ER), these distributions are typically only partially identified. However, identification results exist [11], that exploit conditional independence assumptions between two outcomes (only one of which may be of primary interest). More general results, though not embedded in a causal framework, can be found in studies on concentration graphs with a latent variable [14].

The main contribution of the paper is to formally bridge the two settings and show how these results on concentration graphs can be adopted and adjusted to solve identification problems in causal studies in the presence of intermediate variables.

The paper is organized as follows: the potential outcome framework with one intermediate binary variable and multiple outcomes is presented in Sect. 2, and it is related to a general structure with a single binary latent variable and multiple discrete variables. Section 3 contains the main new results of identifiability of causal quantities of interest and includes some examples to explain the potential applicability of our results. Section 4 discusses the relevant problem of label swapping in our causal framework. An illustrative empirical example is reported in Sect. 5, where data from of a real social job training experiment are analyzed. Concluding remarks and directions for future research are provided in section "Concluding Remarks".

2 Framework and Notation

Let us introduce the potential outcome framework. Throughout the paper we will make the stability assumption (SUTVA; [13]) that there is neither interference between units nor different versions of the treatment. Under SUTVA, let Z_i be a binary treatment assignment for unit i ($Z_i = 0$ if unit i is assigned to the control group, $Z_i = 1$ if unit i is assigned to the treatment group).

We denote by $D_i(z)$ a potential intermediate binary variable for unit i when assigned to treatment z, which is, without loss of generality, assumed to be an

indicator equal to 1 if a specific post-treatment event happens and 0 otherwise. The units under study can be stratified into the following four subpopulations, according to the value of the two potential indicators $D_i(0)$ and $D_i(1)$:

$$\begin{aligned} 11 &= \{i : D_i(1) = D_i(0) = 1\}, \\ 10 &= \{i : D_i(1) = 1, D_i(0) = 0\}, \\ 01 &= \{i : D_i(1) = 0, D_i(0) = 1\}, \\ 00 &= \{i : D_i(1) = D_i(0) = 0\}. \end{aligned}$$

Because only one of the two potential indicators is observed, these four subpopulations are latent, in the sense that in general it is not possible to identify the specific subpopulation a unit i belongs to. Let U_i represent the latent group to which subject i belongs, $U_i = \{11, 10, 01, 00\}$.

Depending on the type of post-treatment event that variable D represents, the four groups may have different interpretations; we give a couple of examples. When the intermediate variable represents the treatment receipts in the presence of non-compliance (also known as instrumental variable setting), the four subpopulations are denoted as compliers, for whom $D_i(z) = z$ for $z \in \{0, 1\}$; never-takers, for whom $D_i(z) = 0$ for $z \in \{0, 1\}$; always-takers, for whom $D_i(z) = 1$ for $z \in \{0, 1\}$; and defiers, for whom $D_i(z) = 1 - z$ for $z \in \{0, 1\}$ [2]. In the presence of censoring due to death, the intermediate variable is an indicator of survival, so that the four subpopulations are usually denoted as always-survivors, survivors only under treatment, survivors only under control, and never-survivors.

We define four potential outcomes for a k-variate binary outcome,

$$\mathbf{Y}_i(z, d) = [Y_{i1}(z, d), Y_{i2}(z, d), \cdots, Y_{ik}(z, d)]',$$

for all possible combinations of treatment assignment and intermediate binary variable, $z \in \{0, 1\}$ and $d \in \{0, 1\}$. However, for every subject i, only two of the four potential outcomes are potentially observed, namely, $\mathbf{Y}_i(z, D_i(z))$, $z \in \{0, 1\}$, the other two potential outcomes being *a priori counterfactuals* [5]. In order to avoid the use of such counterfactuals, we let the k-variate binary outcome variable depend only on treatment assignment: $\mathbf{Y}_i(z)$.

In what follows we will maintain the following assumptions:

Assumption 1 *Random assignment: Z_i is randomly assigned, implying that*

$$Z_i \perp\!\!\!\perp D_i(1), D_i(0), \mathbf{Y}_i(1), \mathbf{Y}_i(0),$$

that is,

$$Z_i \perp\!\!\!\perp U_i, \mathbf{Y}_i(1), \mathbf{Y}_i(0), \quad \forall i.$$

Random assignment of Z_i usually holds by design in randomized experiments. Sometimes it is assumed to hold conditional on pre-treatment covariates. However, without loss of generality, we avoid the use of covariates in this paper.

Assumption 2 *Nonzero effect of Z on D:* $0 < |E(D_i(1) - D_i(0))| < 1$.

This assumption essentially implies that the latent variable U_i may take on at least two values and can be verified from the data [2].

In a causal inference problem, causal estimands of interest are typically causal effects conditional on the values of U (contrasts of summaries of $\mathbf{Y}(1)$ and $\mathbf{Y}(0)$ within a latent group). For example, the effect of the instrument on compliers in the instrumental variable setting (usually interpreted as the causal effect of the receipt of the treatment), or the effect of the treatment on the quality of life in the subpopulation of always-survivors, in the case of censoring by death, are common quantities of interest. To this extent, we usually need to identify the distribution of one or more outcome variables under $Z = 1$ and $Z = 0$. We introduce the joint distribution of potential outcomes:

$$P[\mathbf{Y}_i(z) = (y_1, y_2, \cdots, y_k)|U_i = u] \tag{1}$$

for $z = \{0, 1\}$, $u = \{11, 10, 01, 00\}$.

In the following, we assume that the focus of the analysis are intention-to-treat (ITT) effects on a single outcome, let this be the first outcome, Y_1, for each latent subgroup, which are defined as:

$$E[Y_{i1}(1) - Y_{i1}(0)|U_i = u] \quad u \in \{11, 10, 01, 00\}. \tag{2}$$

The data we can observe are Z_i, $D_i^{obs} = D_i(Z_i)$ and $\mathbf{Y}_i^{obs} = \mathbf{Y}_i(Z_i)$, so that the distributions that are asymptotically revealed by the sampling process are the following:

$$P[\mathbf{Y}_i^{obs} = (y_1, y_2, \cdots, y_k)|Z_i = z, D_i^{obs} = d], P[D_i^{obs} = d|Z_i = z].$$

We can observe four different groups of units, defined by the observed values of Z and D^{obs}, $O(Z, D^{obs})$; each group results from a mixture of two latent strata, as shown in Table 1.

Table 1 Correspondence between observed and latent groups

Observed subgroups $O(Z, D^{obs})$	Latent strata
$O(1,1) = \{i : Z_i = 1, D_i^{obs} = 1\}$	11 or 10
$O(1,0) = \{i : Z_i = 1, D_i^{obs} = 0\}$	00 or 01
$O(0,1) = \{i : Z_i = 0, D_i^{obs} = 1\}$	11 or 01
$O(0,0) = \{i : Z_i = 0, D_i^{obs} = 0\}$	10 or 00

In general, a concentration graph can be used to represent the conditional independencies in the joint distribution of $\mathbf{Y}_i(z)$ and U_i, after conditioning on D_i^{obs} and Z_i, i.e. of $P[\mathbf{Y}_i(z) = (y_1, y_2, \cdots, y_k), U_i = u | Z_i = z, D_i^{obs} = d]$, for $z = \{0, 1\}$, and $d = \{0, 1\}$.

In case we do not exploit restrictions across arms (i.e., across values of Z, such as exclusion restrictions, [11]), the distributions of variables to be identified (within each of the four observed groups) can be generically denoted by $P[\mathbf{Y}_i = (y_1, y_2, \cdots, y_k), W_i = w]$, with W a latent binary variable.

To avoid complex notation, we omit the suffix i in the sequel. We therefore denote with $\mathbf{Y} = (Y_1, \ldots Y_k)$ the k-variate outcome, and with W the binary latent variable. We then consider the distribution of $(\mathbf{Y}, W)$ for observed groups ($Z = z$, $D^{obs} = d$), with W a latent binary variable taking values on a defined subset of U. For example, when $z = 1$ and $d = 1$, W is a latent binary variable taking on values in $\{11, 10\}$.

Under suitable conditional independence restrictions, it is possible to identify the joint distribution of $(\mathbf{Y}, W)$ within each observed group.

Once the outcome distributions are identified, solving a causal problem requires to estimate distributions involved in the causal estimands and contrast them across arms, conditional on the *same* value of U, so that it is fundamental to be able to identify the labels of different latent groups, i.e., the strata membership, as will be shown in Sect. 4.

Recent results on identification of models with one latent binary variable allow for conditional associations between the observable variables (see [1, 14]). Next section contains a review of these results together with a list of concentration graphs corresponding to identified models that can be of interest in our causal context.

3 Some Identification Results

Let $G^K = (K, E)$ be an undirected graph with node set $K = \{0, 1, \ldots, k\}$ and edge set $E = \{(s, j)\}$ whenever vertices s and j are adjacent in G^K, $0 \leq s < j \leq k$. A discrete random variable is associated to each node as follows: to node 0 the latent binary random variable W is associated, while to nodes $1, \ldots, k$ the observable outcome random variables $Y_1, \ldots, Y_k$ are associated. A concentration graphical model is a family of joint distributions of the variables $(\mathbf{Y}, W)$ satisfying the Markov property with respect to G^K, namely that the joint distribution of the random variables factorizes according to G^K; see [8] for definitions and concepts.

Let $\mathcal{M}(\Theta) = \{P_\theta : \theta \in \Theta\}$ be a family of probability distributions over the observable variables with parameter space Θ and $\psi : \theta \to P_\theta$ the parametrization map. A model is generically identified if ψ is finite-to-one almost everywhere in the parameter space (see for details [1]). When the mapping is one-to-one everywhere in the parameter space, then the model is *strictly identifiable*. Strict identification is also known as *global identification*. For generically identified models, a precise characterization of the values leading to the same P_θ, and of the null measure subset where identifiability breaks down, is essential to perform correct analysis [4].

Conditions for (generic/strict) identification of graphical models with one latent variable have been given in [1, 14]. Different from the former, the latter paper focuses on binary latent structures only, but provides explicit expression of the subspace where identifiability breaks down in generically identified models. The conditions of the two papers overlap only partly, but we here refrain from comparing them.

A common problem that arises when dealing with discrete latent variable models is known as "label swapping". This implies that two models with a relabelling of the latent classes generate the same marginal distribution over the observable variables. Therefore, in a binary latent variable model, the parametrization map is at most two-to-one. In the following we refer to strictly identified models as to models that are strictly identified up to label swapping. Awareness of this ambiguity is necessary when interpreting the results. However, in this causal setting, randomization allows us to identify labels across different observed groups, as we will discuss in next section.

It is well-known that the binary latent class model, see Fig. 1a, is strictly identified if and only if $k \geq 3$, see [7, 10]. Strictly identified models that allow for conditional associations among the observable variables require $k \geq 4$. In Fig. 1b the concentration graph corresponding to the most complex strictly identified model with $k = 4$ is presented (see [14], for details). More complex models are useful because they imply less restrictive conditional independence assumptions.

To give some examples, consider an open-label encouragement randomized study, aimed at assessing the effects of a new drug on an outcome. The new drug is known to have some frequent side effects, such as headaches and muscle aches. Noncompliance is present, but, to simplify the example, we hypothesize that the subjects in the control group cannot have access to the new drug. Because treatment is not blinded, we cannot rule out effect of encouragement on the outcomes, that is exclusion restrictions do not hold. Suppose, however, that it is plausible to assume that the three outcomes (primary outcome and the two indicators of side effects) are conditionally independent, given compliance status (U) and treatment assignment (Z). That is, after conditioning on Z and U, the implied model has conditional independence graph as in Fig. 1a.

Consider now the same setting as in the previous example, where there are three binary indicators of side effects. Suppose that, in a specific setting, side effects can be plausibly associated as shown in the graph (with Y_2 independent of Y_4 given

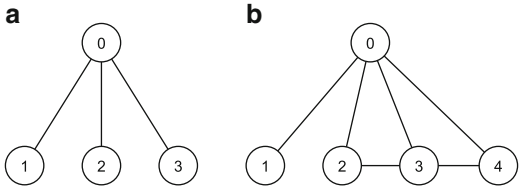

Fig. 1 Concentration graphs corresponding to (**a**) a latent class model for $k = 3$ and (**b**) the most complex strictly identified model for $k = 4$

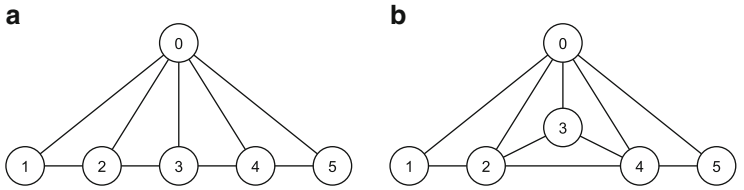

Fig. 2 Concentration graphs with $k = 5$ observables corresponding to (**a**) a strictly identified model and (**b**) a generically identified model

Y_3 and W), but are independent of the primary outcome, conditional on compliance status and treatment assignment. That is, after conditioning on Z and U, the implied model has conditional independence graph as in Fig. 1b.

In Fig. 2a the most complex strictly identified model with $k = 5$ is presented. The model can represent a social randomized experiment with noncompliance and where the exclusion restriction of the random assignment is questioned. A random subset of subjects are offered participation in a training program, while the other subset is denied participation. Both groups are tracked at baseline, and at five subsequent waves to gather information on their employment status. It is plausible that, for the dynamics of the labor market, employment statuses at non adjacent waves are independent, conditional on all other nodes. The implied model, after conditioning on Z and U, has conditional independence structure as in Fig. 2a.

Even more complex association structures allow identification and could be plausible description of empirical settings. An example of a generically identified model is in Fig. 2b. The models fails to be identified in a subspace of null measure, the expression of which can be derived using the results in [14] and is here omitted for brevity.

4 Identification of the Levels of the Latent Variable

The problem of label swapping reflects into the problem of exactly identifying the levels of U across observed groups. In this context, this is particularly necessary as principal effects are defined as contrasts of (features) of the outcome distributions across arms, conditional on the *same* value of U. Here, identification of the levels of the latent variable may be achieved due to randomization and assuming that $P(U = u)$ varies with u, i.e. $P(U = i) \neq P(U = j)$ for all $i, j \in (11, 01, 10, 00)$. In fact, randomization guarantees that U has the same distribution in both treatment arms ($Z = 0$ and $Z = 1$) and, from the observed groups (see Table 2), the distribution of U can be identified in both treatment arms. Therefore, given the two distributions we can identify the labels, again if the four values of U have different probabilities.

If the usually invoked monotonicity assumption is maintained ($D(1) \geq D(0)$), identification of the distribution of U is easier. Monotonicity rules out the presence of the $\{01\}$ latent stratum, so that the subgroup proportions can be identified as

Table 2 Sample sizes, observed and latent groups in Job Corps data

	D^{obs}			D^{obs}			D^{obs}	
Z	0	1	Z	0	1	Z	0	1
0	3275	0	0	{00,10}	–	0	NT, C	–
1	1471	3545	1	{00}	{10}	1	NT	C

$P[U = 11] = P[D^{obs} = 1|Z = 0]$, $P[U = 00] = P[D^{obs} = 0|Z = 1]$, and $P[U = 10] = 1 - P[U = 11] - P[U = 00]$.

When monotonicity is not plausible, however, we can still identify the labels, as the distribution of U is the same for $Z = 0$ and $Z = 1$. For example, thank to randomization, $P[U = 11] = P[U = 11|Z = 1] = P[U = 11|Z = 0]$. $P[U = 11|Z = 1]$ is identifiable because it is equal to $P[U = 11|Z = 1, D^{obs} = 1]P[D^{obs} = 1|Z = 1] + P[U = 11|Z = 1, D^{obs} = 0]P[D^{obs} = 0|Z = 1] = P[U = 11|Z = 1, D^{obs} = 1]P[D^{obs} = 1|Z = 1]$ and these are identifiable quantities. Analogously, $P[U = 11|Z = 0]$ is identifiable because it is equal to $P[U = 11|Z = 0, D^{obs} = 1]P[D^{obs} = 1|Z = 0] + P[U = 11|Z = 0, D^{obs} = 0]P[D^{obs} = 0|Z = 0] = P[U = 11|Z = 0, D^{obs} = 1]P[D^{obs} = 1|Z = 0]$. Given the identification of outcome distribution within strata and treatment arm and the identification of labels, we can identify the contrasts in (2) for each principal stratum (each level of the latent variable).

5 An Illustrative Empirical Example

For illustrative purposes, a randomized study with noncompliance, where the exclusion restriction of the random assignment has been questioned, is analyzed. The study is the National Job Corps (JC) Study, a randomized experiment performed in the mid-1990s to evaluate the effects of participation in JC (D), a large job training program for economically disadvantaged youths aged 16–24 years [3]. A random sample of eligible applicants was randomly assigned into treatment and control groups (Z), with the second group being denied access to JC for 3 years. Both groups were tracked at baseline, soon and at 12, 30 and 48 months after randomization. Previous works have concentrated on global ITT effects, i.e., effects of being assigned to enroll in Job Corps (e.g., [9,16]). However, noncompliance was present, as only 68 % of those assigned to the treatment group actually enrolled in JC within 6 months from assignment. When estimating the effect on compliers, the ER for never-takers was always maintained (e.g., [6]). The ER for never-takers rules out any effect of assignment on the outcomes for those who do not take the treatment. However being denied enrollment in JC, as opposed to deciding not to accept the offer to enroll, may, in principle, affect the labor market behavior of never-takers, especially in the short-term. For example, the denial may encourage applicants to temporarily look for alternative forms of training, possibly reducing their job search intensity.

Here, we concentrate on the following binary variables: smoking habits at 12th month ($CIG12$), employment indicators at 12th ($W12$), 30th ($W30$) and

48th month ($W48$); we use only observations where all the outcomes and the treatment indicator are not missing (N=8291). The three employment indicators are plausibly associated, with the employment indicator at 12th month independent of the indicator at 48th month, given the indicator at 30th month, compliance status and treatment assignment. The employment indicators are plausibly assumed to be independent of smoking habits at 12th month conditional on compliance status and treatment assignment. Thus, conditioning on Z and D, the implied model corresponds to the concentration graph in Fig. 1b and is a strictly identified model, as in [14].

By design, in the study there are only two latent groups, namely compliers (C) and never-takers (NT), and we can observe three different subgroups of units, with only the subgroup $O(0,0)$ resulting from a mixture of two latent strata, as shown in Table 2. Once identified, the outcome distributions involved in the causal estimands (ITT effects in our example) have been estimated by maximum likelihood and contrasted across arms, conditional on the same value of U (i.e., within a latent group). The proposed model is well fitting the data, as the LRT against the saturated model is 6.411 with 3 degrees of freedom (p-value 0.0932). Notice that, by the identification results of [14], the asymptotic distribution of the test is well approximated by a chi-squared distribution.

Results are summarized in Table 3, where the estimated distributions of each outcome variable, under treatment and under control, and the estimated ITT effects are reported, for compliers and never-takers respectively.

Table 3 Estimated outcome distributions and estimated ITT effects for compliers and never-takers. Maximum likelihood estimates and standard errors

Compliers		
Z=0	Z=1	ITT_C
$P(W12 = 1) = 0.44$	$P(W12 = 1) = 0.36$	-0.08
(*0.01*)	(*0.01*)	(*0.01*)
$P(W30 = 1) = 0.53$	$P(W30 = 1) = 0.58$	0.04
(*0.01*)	(*0.01*)	(*0.01*)
$P(W48 = 1) = 0.60$	$P(W48 = 1) = 0.62$	0.02
(*0.01*)	(*0.01*)	(*0.01*)
$P(CIG12 = 1) = 0.49$	$P(CIG12 = 1) = 0.50$	0.01
(*0.01*)	(*0.01*)	(*0.01*)
Never-takers		
Z=0	Z=1	ITT_{NT}
$P(W12 = 1) = 0.49$	$P(W12 = 1) = 0.44$	-0.05
(*0.03*)	(*0.01*)	(*0.03*)
$P(W30 = 1) = 0.31$	$P(W30 = 1) = 0.54$	0.23
(*0.03*)	(*0.01*)	(*0.03*)
$P(W48 = 1) = 0.38$	$P(W48 = 1) = 0.61$	0.22
(*0.03*)	(*0.01*)	(*0.03*)
$P(CIG12 = 1) = 0.60$	$P(CIG12 = 1) = 0.48$	-0.13
(*0.03*)	(*0.01*)	(*0.03*)

Focussing on the outcome variables of main interest, i.e. employment indicators, results point to a negative effect of assignment on employment for compliers in the short-run, confirming lock-in effects of those participating in the program (see, e.g., [6, 15]); while the effect becomes positive at month 30th. For never-takers we found a negligible negative effect of assignment on employment in the short-run, possibly due to a reduced search intensity of the never-takers denying participation and therefore looking for alternative training. The positive effect at both subsequent waves for never-takers can be partly attributed to these different forms of training, maybe better targeted to single individuals. This is also in line with former results (see [6]), which showed that never-takers are better educated and with longer labour market experience.

Concluding Remarks

Conditional independence assumptions are naturally embedded in studies on causality and graphical models are natural tools to entail those assumptions. We have proposed a way to take into account, within causal inference, results on identification pertaining to graphical models with one binary latent variable. Graphical models with one binary latent variable have been shown to describe principal stratification settings, where principal strata within observed groups take the form of a latent binary variable. Therefore, identification results for concentration graphs can be adapted and extended for identification of principal strata effects.

Importantly, our proposal allows one to use conditional independence structures that vary across observed groups, thereby providing a powerful tool that can flexibly adapt to many empirical settings. Directions for future research include the study of procedures to optimize the choice of secondary outcomes and for model specification and checking.

References

1. Allman, E.S., Matias, C., Rhodes, J.A.: Identifiability parameters in latent structure models with many observed variables. Ann. Stat. **37**, 3099–3132 (2009)
2. Angrist, J.D., Imbens, G.W., Rubin, D.B.: Identification of causal effects using instrumental variables. J. Am. Stati. Assoc. **91**, 444–455 (1996)
3. Burghardt, J., McConnell, S., Schochet, P., Johnson, T., Gritz, M., Glazerman, S., Homrighausen, J.: Job Corps Work? Summary of the National Job Corps Study. Document No. PR01-50. Mathematica Policy Research, Princeton (2001)
4. Drton, M.: Likelihood ratio tests and singularities. Ann. Stat. **27**, 979–1012 (2009)
5. Frangakis, C.E., Rubin, D.B.: Principal stratification in causal inference. Biometrics **58**, 191–199 (2002)
6. Frumento, M., Mealli, F., Pacini, B., Rubin, D.B.: Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. J. Am. Stat. Assoc. **107**, 450–466 (2012)
7. Goodman, L.A.: Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika **61**, 215–231 (1974)
8. Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)

9. Lee, D.: Training, wages, and sample selection: estimating sharp bounds on treatment effects. Rev. Econ. Stud. **76**, 1071–102 (2009)
10. McHugh, R.B.: Efficient estimation and local identification in latent class analysis. Psychometrika **21**, 331–347 (1956)
11. Mealli, F., Pacini, B.: Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. J. Am. Stat. Assoc. **108**, 1120–1131 (2013)
12. Rubin, D.B.: Estimating causal effects of treatments in randomized and non randomized studies. J. Edu. Psychol. **66**, 688–701 (1974)
13. Rubin, D.B.: Bayesian inference for causal effects: the role of randomization. Ann. Stat. **6**, 34–58 (1978)
14. Stanghellini, E., Vantaggi, B.: On the identification of discrete concentration graph models with one hidden binary variable. Bernoulli **19**, 1920–1937 (2013)
15. van Ours, J.: The locking-in effect of subsidized jobs. J. Comp. Econ. **32**, 37–52 (2004)
16. Zhang, J.L., Rubin, D.B., Mealli, F.: Likelihood-based analysis of causal effects of job-training programs using principal stratification. J. Am. Stat. Assoc. **104**, 166–176 (2009)

Dynamic Segmentation of Financial Markets: A Mixture Latent Class Markov Approach

Francesca Bassi

Abstract

The latent class approach is innovative and flexible and can provide suitable solutions to several problems regarding the development of marketing strategies, because it takes into account specific features of the data, such as their scale of measure (often ordinal or categorical, rather than continuous), their hierarchical structure and their longitudinal component. Dynamic segmentation is of key importance in many markets where it is unrealistic to assume stationary segments due to the dynamics in consumers' needs and product choices. In this paper, a mixture latent class Markov model is proposed to dynamically segment Italian households with reference to financial products ownership.

Keywords

Dynamic segmentation • Financial products • Latent class models • Life cycle • Mover-stayer model

1 Introduction

The latent class (LC) approach is flexible and can provide solutions to several problems regarding the definition and the development of marketing strategies, because it takes into account specific features of the data, such as their scale of measure (often ordinal or categorical, rather than continuous), their hierarchical structure and their longitudinal component. In the recent literature some studies can be found on segmentation performed on hierarchically structured data [3, 4], less attention is devoted to longitudinal data and dynamic segmentation as noted by

F. Bassi (✉)
Statistics Department, University of Padova, Via C. Battisti 241, 35121, Padova, Italy
e-mail: francesca.bassi@unipd.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_20, Published online: 12 November 2014

[15]. Dynamic segmentation instead is of key importance in many markets where it is unrealistic to assume stationary segments due to the dynamics in consumers' needs and product choices.

In this paper a mixture LC Markov (LCM) model is proposed to dynamically segment the Italian market with reference to financial products ownership. The standard LCM model was previously applied successfully in marketing (see, among others, [8]). The estimation of a mixture LCM model, specifically of the mover-stayer model [12, 13], has several advantages with respect to the estimation of a standard LCM model: it allows to achieve a better model fit and to identify customer segments more precisely, improving knowledge of market dynamics, as results in the paper demonstrate. Moreover, a paper by [11] shows that if unobserved heterogeneity in the initial state and in transition probabilities of the latent chain is not taken into account, model measurement component can be estimated with larger bias.

The data used in the paper are collected by the Bank of Italy with the Survey on Household Income and Wealth and refer to a representative sample of Italian families. Ownership of 13 financial products from 2002 to 2010 and household characteristics referring to the same period are considered. The study of household savings has always had relevance in the economic international literature [5] but it became even more important in recent years and especially in Italy, country where the proportion of income devoted by families to savings has always been substantially high. In the last years, however, such relative amount constantly diminished: in our reference period we observe a proportion of savings over income of Italian families of 13.8 % at the beginning of 2002 and of 9.8 % at the end of 2010 [10].

The paper is organized as follows. Section 1.1 introduces the traditional and mixture LCM model. Section 2 describes the survey and the data. Section 3 provides results comparing the traditional and the mixture LCM model and the section "Concluding Remarks" concludes.

1.1 The LC Markov Model

Let us consider the simplest formulation of latent class Markov (LCM) models [16], which assumes that true unobservable transitions follow a first-order Markov chain. As in all standard latent class model specifications, local independence among the indicators is assumed, i.e., indicators are independent conditionally on latent variables.[1]

Let X_{it} denote a latent variable which categories indicate segment belonging at time t for a generic sample household i, $i = 1, .., n$. Y_{ijt} is a binary observed variable assuming value 1 if household i owns financial product j, $j=1,\ldots, J$ at time t, and assuming value 0 otherwise; Y_{ijt} are the LCM model indicators. $P(X_{i1} = l_1)$ is

[1] In the LC model with one indicator per latent variable, the assumption of local independence coincides with the Independent Classification Error (ICE) condition.

the probability of the initial state of the latent Markov chain, i.e., the probability that household i belongs to segment l_1 at time 1 and $P\left(X_{it+1} = l_{t+1} | X_{it} = l_t\right)$ is the transition probability between state l_t and state l_{t+1} from time t to $t+1$, with $t = 1,\ldots,T-1$, where T represents the total number of consecutive, equally spaced time-points over which a household is observed, i.e., the probability that household i moves from segment l_t to segment l_{t+1} over the two survey waves. Besides, let $P\left(Y_{ijt} = 1 | X_{it} = l_t\right)$ be the probability of owning financial product j at time t, given that household i at time t belongs to segment l_t, this is also called model measurement component.

It follows that $P\left(Y(1), \ldots Y(T)\right)$ is the proportion of units observed in a generic cell of the T-way contingency table. For a generic sample household i, a first-order LCM model is defined as:

$$P\left(Y_{i1} = 1, \ldots Y_{iT} = 1\right) = \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} \times P(X_{i1} = l_1) \prod_{t=2}^{T} P(X_{it} = l_t | X_{it-1} = l_{t-1}) \prod_{j=1}^{J} \prod_{t=1}^{T} P(Y_{ijt} = 1 | X_{it} = l_t) \tag{1}$$

where l_t varies over K latent classes.

In a LCM model with concomitant variables, latent class membership and latent transitions are expressed as functions of covariates with known distribution [6]: $P\left(X_{i1} = l_1 | \mathbf{Z}_{i1} = \mathbf{z}_1\right)$, where $\mathbf{z}_1$ is a vector containing the values of covariates for household i at time 1, estimates covariates effects on the initial state and $P\left(X_{it} = l_t | X_{it-1} \mathbf{Z}_{it} = \mathbf{z}_t\right)$, where $\mathbf{z}_t$ is a vector containing the values of covariates for household i at time t, estimates covariates effects on latent transitions. Equation (2) specifies a first-order LCM model with concomitant variables.

$$P\left(\mathbf{Y}_i = \mathbf{y} | \mathbf{Z}_i = \mathbf{z}\right) = \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} P(X_{i1} = l_1 | \mathbf{Z}_{i1} = \mathbf{z}_1) \prod_{t=2}^{T} P(X_{it} = l_t | X_{it-1} = l_{t-1}, \mathbf{Z}_{it} = \mathbf{z}_t) \prod_{j=1}^{J} \prod_{t=1}^{T} P(Y_{ijt} = 1 | X_{it} = l_t) \tag{2}$$

Typically, conditional probabilities are parameterized and restricted by means of logistic regression models. Parameters can be estimated via maximum likelihood using the E-M algorithm [7].

A mixture LCM assumes the existence in the population of non directly observable groups following latent chains with different initial state probabilities and different transition probabilities; the groups can be assumed to have also

different response probabilities; the model can be extended to include time-varying and time-constant covariates [14]. A special case of a mixture LCM model is the mover-stayer model: there is a group of movers who have positive probabilities of moving from one state to another over time, and a group of stayers who do not change latent state. For this last group, transition probabilities between two different states are imposed equal to 0. A two-mixture fist-order LCM model with concomitant variables has the following form:

$$
\begin{aligned}
&P\left(\mathbf{Y}_i = \mathbf{y} | \mathbf{Z}_i = \mathbf{z}\right) \\
&= \sum_{l_1}^{K} \ldots \sum_{l_T}^{K} \sum_{w=1}^{2} P(W = w) P(X_{i1} = l_1 | \mathbf{Z}_{i1} = \mathbf{z}_1,\ W = w) \\
&\prod_{t=2}^{T} P(X_{it} = l_t | X_{it-1} = l_{t-1}, \mathbf{Z}_{it} = \mathbf{z}_t, W = w) \\
&\quad \prod_{j=1}^{J} \prod_{t=1}^{T} P(Y_{ijt} = 1 | X_{it} = l_t, W = w) \qquad (3)
\end{aligned}
$$

where W is a binary latent variable dividing the population into two unobservable groups following over time different latent chains and with a different measurement model component. The mover-stayer model is obtained assuming, for $l_t \neq l_{t-1}$, that $P\left(X_{it} = l_t | X_{it-1} = l_{t-1},\ W = 2\right) = 0$.

2 The Data

The data on financial product ownership by Italian households is collected with the Survey on Household Income and Wealth conducted by the Bank of Italy. The survey gathers information on income, savings, consumption expenditure and real wealth of Italian households, as well as on household composition, demographic characteristics and labour force participation [9]. In the paper the sample of 1,834 households who participate in all waves from 2002 to 2010 is considered.

Information on five equally-spaced (2 years) time points on ownership of 13 financial products and on family characteristics such as form of tenure dwelling, number of household members, number of income recipients, geographical area, gender, age, educational qualification, work status and branch of activity sector of the head of household is used. Previous studies by the Bank of Italy show that financial activities diffusion among Italian families vary with the selected covariates [1].

Table 1 lists the percentage of households holding, in the five measurement occasions, the 13 financial activities: certificates of deposits (CD), repos (CT), post office certificates (BFP), Treasury bills up to 1-year maturity (BOT), fixed-rate long term Treasury bonds (BTP), bonds (OBB), mutual funds (QFC), individually

Table 1 Ownership of financial assets by type, percentage of households, 2002–2010

	2002	2004	2006	2008	2010
CD	2.3	2.4	2.6	2.9	3.2
PCT	0.9	0.7	1.1	1.7	0.9
BFP	5.9	5.8	6.9	6.5	5.9
BOT	10.1	6.9	8.1	10.2	8.5
BTP	2.7	2.9	2.8	2.6	2.9
OBB	7.7	8.7	8.4	10.4	11.2
QFC	12.6	12.5	12.1	8.7	8.8
GP	2.6	2.0	1.7	0.7	1.0
TE	1.3	1.2	1.1	1.4	1.9
COOP	2.2	2.9	2.5	2.8	2.7
DEP	88.5	89.0	90.7	90.9	85.6
SHA	10.6	8.7	8.8	7.4	6.7
CCT	3.9	3.7	3.8	3.0	2.9

managed portfolios (GP), foreign securities (TE), loans to cooperatives (COOP) and bank or postal deposits (DEP), shares and other equities (SHA), floating rate Treasury certificates indexed to BOTs (CCT). The 13 variables used in this study represent all financial investment opportunities aggregating variables with negligible frequencies.

It is interesting to note that our observational interval covers the first period immediately after the economic crisis of 2008. As shown in [2], the crisis has intensified the trends already under way, as confirmed by the further decline in the saving rate and the deterioration in the financial situation of low-income households, young people and tenants. The percentage of income devoted to savings by Italian households decreased from 12.0 % at the beginning of 2008 to 9.8 % at the end of 2010. Between 2002 and 2008, the incidence of ownership of postal and bank deposits rose from 88.5 to 90.9 %, in 2010 it decreased to 85.6 %. The proportion of households owning bonds, investment funds and other risky assets declined from 38 % in 2002 to 33.7 % in 2010. The decline from 2008 to 2010 was especially evident for BOTs.

The distributions of financial assets by family characteristics from 2002 to 2010 show that financial strategies depend on household structure and socio-economic environment and an interaction between ownership, household situation and time, suggesting an analysis of the market in terms of dynamic segmentation.

3 Results

A latent class Markov model with a first order latent Markov chain and 13 binary indicators for each time point was specified. The model assumes a time-constant measurement component in order to ensure that segments remain the same in the market; moreover, this assumption guarantees identifiably. The model was estimated

starting with one latent class per each latent variable and then the number of classes was increased till the Bayesian Criterion Index (BIC) began to raise. In order to avoid local maxima, each model was estimated several times with different sets of starting values.[2] The best fitting model was that with five latent classes for every time point. On this model, the assumption of time-constant transition probabilities was accepted by means of the conditional test ($\Delta L^2 = 70$, $\Delta \text{df} = 80$). Next, covariate effects both on the initial state in 2002 and on transitions probabilities were introduced.

Table 2 presents estimation results for model measurement component: average over the period segments' sizes,[3] segments' sizes in the five survey waves and response probabilities. The largest group (53.44 %) is composed of households owning only a bank or postal deposit and very few forms of other assets showing

Table 2 Standard LCM model estimation: at each wave and average over the period segments' sizes and segments' profiles (percentages)[a,b]

	1	2	3	4	5
Size					
2002	15.56	48.04	7.33	10.36	18.71
2004	12.68	42.39	7.87	9.93	17.12
2006	11.45	54.49	8.21	9.85	16.00
2008	10.81	55.76	8.40	9.90	15.13
2010	10.47	56.51	8.59	9.99	14.44
Average	12.20	53.44	8.08	10.01	16.28
Profile					
DEP	**24.92**	**99.70**	**98.54**	**99.33**	**99.96**
BFP	0.95	1.81	**47.96**	3.33	5.79
COOP	0.00	0.24	**17.69**	4.17	4.10
BOT	0.00	1.19	3.53	**60.27**	11.48
BTP	0.00	0.06	0.77	**12.98**	8.43
CCT	0.09	0.07	1.96	**20.19**	7.52
CD	0.25	0.95	5.32	5.13	**7.30**
PCT	0.00	0.05	0.55	0.98	**5.23**
OBB	0.00	1.25	13.60	17.49	**34.41**
QFC	0.00	2.29	11.37	8.95	**47.11**
GP	0.00	0.67	0.56	0.62	**6.72**
TE	0.00	0.20	1.16	0.54	**6.65**
SHA	0.00	0.87	7.59	3.81	**41.29**

[a]To help interpretation, some meaningful percentages appear in bold

[b]The 13 binary variables are all significant indicators of the five-classes corresponding latent variable

[2]Model estimation was performed with Latent Gold 4.5 Syntax [14].

[3]Latent classes dimensions over time changes due to switching between segments according to the first-order Markov chain.

that they rely heavily on liquid savings forms for transactional purposes; 16.28 % of families owns a deposit and one or more risky financial assets, with a quite diversified portfolio; the group with dimension 12.20 % is that comprising the poorest households: they do not own any kind of financial asset, moreover, a great percentage does not even hold a deposit. A segment with dimension 10.01 % contains households that mainly possess state bonds (as well as deposits). Finally, 8.08 % of families owns a deposit and has made some investment in less risky financial assets such as postal bonds and loans to cooperatives, they avoid risky assets. To help interpretation, we ranked segments in the table in ascending order of product penetration rates, from households owning only bank or postal deposits to households owning more sophisticated financial products. Segment sizes show a dynamics over time: segments 5 (more risky assets) and 1 (only a few with deposits) decrease, segments 2 (only deposits) and 3 (less risky assets) increase while segment 4 (state bonds) slightly decreases.

Table 3 lists transition probabilities. A large percentage of households remained in the same segment over two waves as it is indicated by the numbers on the diagonal of the table. However, the switching that occurs can explain changes in segments' dimension over time.

Estimation results from the standard latent class Markov model show that a high percentage of families do not change segment over two subsequent survey waves. This evidence suggests to specify a mover-stayer latent class Markov model in order to better identify segments, mobility patterns and their determinants. Also in this case, the best fitting model is with five latent classes, a first-order stationary latent chain and constant over time and over the two latent groups response probabilities. Covariates affecting the initial state and transition probabilities of the latent chain were introduced and only those with a significant effect retained. This model has on our data a better fit, in terms of the BIC index equal to 37,335, than the standard LCM model for which the BIC index was equal to 37,595.[4]

Table 3 Standard LCM model estimation: transition probabilities (percentages)[a]

	1	2	3	4	5
1	77.52	21.57	0.91	0.00	0.00
2	2.97	93.06	0.44	2.58	0.95
3	0.00	5.65	94.35	0.00	0.00
4	0.00	10.54	3.10	81.13	5.23
5	0.00	9.23	0.63	2.63	87.50

[a]All effects in the model log-linear representation are statistically significant

[4]Due to the sparse and unbalanced nature of the contingency table, the classical statistics to evaluate model fit, X^2 and L^2, do not have asymptotic χ^2 distribution and then cannot be used. Using various statistics based on the information criterion (BIC, AIC, AIC3) the mixture LCM model performs better than the standard one.

Estimation identifies that 65 % of households are movers and 35 % are stayers over the entire period. Belonging to the subpopulations of movers or stayers significantly depends on head of household's educational condition (less educated heads—no title and elementary school—show a higher probability to be stayers), on head of household's age (youngest heads, till 49, have a higher probability to be movers), on the number of income recipients (families with no, 1 or 3 and more income recipients tend to be stayers), on head of household's professional condition (among professionals, business owners and retired heads we observe more stayers) and economic activity sector (among people not working we observe more stayers).

The five segments are slightly different from those identified by the traditional LCM model: 11.67 % of families belongs to segment 1 (they do not own any kind of financial asset and only a small fraction—20.70 %—has a bank or postal deposit); 53.40 % of families has only a bank or postal deposit and very few other financial products (segment 2); segment 3 (deposits and postal bonds) is composed of 4.86 % of families; 14.89 % of households invests in state bonds and loans to cooperatives (segment 4) and 15.18 % is interested in more risky financial assets. The mixed LCM model identifies segment 3 as that of families who own postal bonds in a great percentage (76.04 %) and bank or postal accounts; the standard LCM model identified segment 3 as that of families who invest in less risky assets such as postal bonds and loans to cooperatives and have a postal or bank account. In the last years the postal service in Italy has become very attractive as an opportunity to invest family savings and it is very plausible that there is a segment of its loyal clients. The identification of the five market segments by the mixed LCM model seems more sensible than that obtained estimating the standard LCM model at least for this country.

The two classes of movers and stayers are highly correlated with the initial state: the χ^2 test has a p-value lower than 0.0001. Movers are more likely to be in segment 2 (59.45 % vs. 42.16 %), stayers are more likely to be in segments 3 (6.59 % vs. 3.93 %) and 5 (32.53 vs. 11.69); the difference among proportions of movers and stayers in segments 1 and 4 is not statistically significant. Magidson et al. [11] states that when the latent classes are highly associated with the initial state of the latent chain, not taking into account unobserved heterogeneity, i.e., estimating a standard LCM model, inflates estimates of model measurement component.

Table 4 lists estimated transition probabilities for the sub-sample of movers. Segment 1 (only a few with deposits) is the most unstable one, segment 2 (only

Table 4 Mixture LCM model estimation: movers' transition probabilities (percentages)[a]

	1	2	3	4	5
1	59.73	38.34	1.93	0.00	0.00
2	4.95	89.07	0.94	3.11	1.93
3	3.33	15.19	79.73	1.00	0.75
4	0.00	12.21	1.97	85.82	0.00
5	0.00	20.64	1.11	3.61	74.64

[a]All effects in the model log-linear representation are statistically significant

Table 5 Mixture LCM model estimation: covariate effects on initial state (effect coding)

Segment	1	2	3	4	5
HOUSE					
Owners	−25.76[a]	−24.00[a]	34.21[a]	40.53[a]	−24.17[a]
Tenants	−24.17[a]	−24.14[a]	33.84[a]	39.82[a]	−25.36[a]
Redemption	75.13[a]	73.65	−102.71	−120.61	74.54
Usufructuaries	−25.20[a]	−24.71[a]	34.65	40.26	−25.00
AGE					
≤ 34	0.18	0.26	0.93[a]	−1.01[a]	−0.37
35–49	−0.52[a]	0.03	0.61[a]	−0.39	0.28
50–64	−0.07	−0.14	−0.70[a]	0.57[a]	0.34[a]
≥ 65	0.41[a]	−0.15	−0.84[a]	0.83[a]	−0.25
GENDER					
Male	−0.46[a]	−0.19[a]	0.55[a]	−0.16	0.25[a]
Female	0.46[a]	0.19[a]	−0.55[a]	0.16	−0.25[a]
AREA					
North	−1.17[a]	−0.31[a]	−0.04	0.63[a]	0.89[a]
Centre	−0.63[a]	−0.17	0.54[a]	0.30	−0.03
South	0.18[a]	0.48[a]	−0.49[a]	−0.93[a]	−0.86[a]
RECIPIENTS					
0	90.60[a]	89.37	−99.64	−81.67	1.33
1	−22.22[a]	−21.61[a]	24.31[a]	20.14[a]	−0.61[a]
2	−23.25[a]	−21.85[a]	24.87[a]	20.49[a]	−0.26
3	−23.02	−21.82	24.40	20.54	−0.09
≥ 4	−22.11	−24.09	26.06	20.50	−0.36
WORK					
Blue-collar	60.39	21.48	−162.33	56.46	24.00
Office	12.61[a]	−26.10	25.77	10.10[a]	−22.39[a]
Manager	10.68[a]	−25.64	25.58	10.91[a]	−21.53[a]
Professional	17.34	−20.43	31.60	16.56	−45.07
Business owner	−125.83	103.37	26.97	−113.76	109.25
Self-employed	11.88	−26-26	26.21	10.01	−21.84
Retired	12.92	−26-43	26.21	9.72	−22.42

[a]significant at 0.05 level

deposits) is the most stable one and also the most attractive one in the sense that it attracts over time clients from all other segments. There is no switching between the segments 4 and 5 and segment 1; these are the most different groups in terms of financial behavior.

Table 5 lists covariates' significant effects on the initial state. Families who are owners or tenants of their house are more likely to be in segments 3 (postal bonds) and 4 (state bonds) and less likely to be in segments 1 (only a few with deposits), 2 (only deposits) and 5 (more risky assets); families with a house with redemption are more likely to be in segment 1; families who are usufructuaries are more likely

to be in segment 1, less likely in segment 2. Households with the youngest heads (≤ 34) are more likely to own postal bonds, less likely to invest in state bonds; when the head is between 35 and 49 the probability to be in segment 3 is higher while it is lower to be in group 1; households with head between 40 and 64 are more likely to invest in state bonds and more risky financial assets and less likely to be in segment 3; households with the oldest heads tend to be in segment 4 and not in segment 3. Families in the North are more prone to invest in state bonds and more risky assets, less likely to belong to segments 1 and 2; families living in the Centre are more likely to be in segment 3, less likely to be in segment 1 and families in the South are more likely not to have any kind of financial asset or to hold only deposits, less likely to be in all other segments. This evidence is quite consistent with socio-economic differences across Italy. For what concerns the gender, with a male head of household the probability is higher to belong to segments 3 and 5 and lower to segments 1 and 2; the opposite is true when the head is a woman. Households financially more active are those with one and two income recipients, they are more likely to be in segments 3 and 5, less likely to be in the other segments. Economic activity sector of the head of household has a small but statistically significant evidence on segment belonging, specifically, households with head in industry and trade are more prone to belong to segments 1 and 4, less prone to belong to segment 5.

The set of covariates that significantly affects switching probabilities is slightly different (Table 6). Families with head of household with no educational title are more likely to move towards segments 1, 2 and 4, not towards segments 3 and 5; households with heads with primary school have a higher probability to move

Table 6 Mixture LCM model estimation: covariate effects on destination state (effect coding)

Segment	1	2	3	4	5
EDUCATION					
No certificate	8.08[a]	6.69[a]	−10.14[a]	4.50[a]	−9.13[a]
Primary school	−0.16	−1.34[a]	2.50[a]	−1.72[a]	0.72
Lower-second.	−1.44[a]	−1.57[a]	2.42[a]	−1.14[a]	1.72[a]
Upper-second.	−2.51[a]	−1.93	2.69	1.20[a]	2.96
University	−3.97[a]	−1.86	2.54	−0.44	3.73
AREA					
North	−0.84[a]	−0.17	−0.02	0.19	0.85[a]
Centre	−0.32	−0.11	−0.33	0.34	0.41
South	1.16[a]	0.28[a]	0.34	−0.53[a]	−1.26[a]
0	17.58	−7.87	10.75	−12.26	−8.19
1	1.58	0.90[a]	−4.67	1.36[a]	0.84
2	1.08	0.72[a]	−4.53	1.62[a]	1.12[a]
3	0.43	0.61	−3.43	1.98	1.29
≥ 4	−19.81	5.65	1.88	7.31	4.97

[a]significant at 0.05 level

towards segment 3, lower towards segments 2 and 4; households with heads with lower-secondary school have a higher probability to move to segments 1, 2 and 4, lower to segments 3 and 5; when the head has a upper-secondary or university degree, the probability is lower to move to segment 1. For area of the country where the family lives, where values on covariates imply a greater probability to belong to an initial state, the model shows that the same covariates' values imply also a greater probability to switch to this state. Only few parameters are statistically significant for the number of income recipients: households with one and two incomes have higher probability to move to segments 2 and 4.

Concluding Remarks

In the paper a mover-stayer first-order LCM model is estimated to dynamically segment the Italian market of financial products ownership. Estimation results show the existence of five different groups of households with different levels of activity in the financial market. Around 11 %, on average in the 8-year period considered, of families belongs to a group owning almost no financial asset (only a small percentage of families has a bank or postal deposit). The largest group (53 %) has only deposits and is almost not at all interested in other financial products. The smallest group (5 %) of households concentrates investment in less risky assets such as postal bonds, beyond having bank or postal deposits. A 15 % of families is interested almost only in states bonds, over deposits. Finally, the most active group in the market (15 % of households) owns mainly more risky financial assets such as shares, certificate of deposits, mutual funds, repos, bonds, individually managed portfolios, foreign securities.

The above groups can be well considered market segments since they satisfy the required properties. Groups are substantial, they are stable since they are the same over time, they are identifiable due to the fact that there is a significant effect of covariates on the initial state. The groups can easily be reached by marketers (accessible) and their characteristics immediately suggest marketing strategies (actionability).

Tenure dwelling, area of the country where the family lives, number of income recipients, age, gender and economic activity sector of the head of household significantly affect segment membership. Results are consistent with previous research [1] and with well known difference of socio-economic conditions across the country and Italian families.

The mover-stayer LCM model takes into account unobserved heterogeneity, identifying two subpopulations: one composed of households who never change segment over the period (35 % of stayers) and one of movers. Belonging to the subpopulations of movers or stayers significantly depends on the number of income recipients and on various characteristics of the head of the household (age, education, profession, economic activity sector).

Acknowledgements Research for this paper was supported by grant CPDA121180/12 financed by the University of Padova for the project with title "Statistical and econometric approach to marketing: applications and developments to customer satisfaction and market segmentation".

References

1. Bank of Italy: Supplement to the Statistical Bulletin. Sample Survey Household Income and Wealth in 2010. Bank of Italy, Rome (2012)
2. Bartiloro, L., Rampazzi, C.: Il risparmio e la ricchezza delle famiglei Italiane durante la crisi. Questioni di Economia e finanza. OP 148. Bank of Italy, Rome (2013)
3. Bassi, F.: Analysing markets within the latent class approach: an application to the pharma sector. Appl. Stoch. Model. Bus. (2012). doi: 10.1102/asmb.1910
4. Bjimolt, T.H.A., Paas, L.J., Vermunt, J.K.: Country and consumer segmentation: multi-level latent class analysis of financial product ownership. Int. J. Res. Mark. **21**, 323–340 (2004)
5. Browning, M., Lusardi, A.: Household saving: micro theories and micro facts. J. Econ. Lit. **34**, 1797–1855 (1996)
6. Dayton, C.M., Mcready, G.B.: Concomitant-variable latent-class models. J. Am. Stat. Assoc. **83**, 173–178 (1988)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**, 1–38 (1977)
8. Dias, J.G., Vermunt, J.K.: Latent class modelling of websites users' search patterns: implications for online market segmentation. J. Retailing Consum. Serv. **14**, 359–368 (2007)
9. Giraldo, A., Rettore, E., Trivellato, U.: Attrition bias in the Bank of Italy's survey of household income and wealth, WP 41, Progetto di ricerca . Occupazione e disoccupazione in Italia: misura e analisi dei comportamenti. Department of Statistics, University of Padova, Italy (2001)
10. Istat: Household Income and Savings and Non-financial Corporations Profits. Istat, Rome (2012)
11. Magidson, J., Vermunt, J.K., Tran, B.: Using a mixture latent Markov model to analyze longitudinal U.S. employment data involving measurement error. In: Shigemasu, K., Okada, A., Imaizumi, T., Hoshino, T. (eds.) New Trend in Psychometrics, pp. 235–242. Universal Academy Press, Tokio (2007)
12. Paas, L.J., Vermunt, J.K., Bijmolt, T.H.A.: Discrete time discrete class latent Markov modeling for assessing and predicting household acquisitions of financial products. J. R. Stat. Soc. A **170**, 955–974 (2007)
13. Van de Pol, F., Langeheine, R.: Mixed Markov latent class models. Sociol. Methodol. **33**, 231–247 (1990)
14. Vermunt, J., Magidson, J.: LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module. Statistical Innovations, Belmont (2008)
15. Wedel, M., Kamakura, W.A.: Market Segmentation: Concepts and Methodological Foundations. Kluwer Academic, Boston (2000)
16. Wiggins, L.M.: Panel Analysis: Latent Probability for Attitude and Behavior Processes. Elsevier Scientific, New York (1973)

Latent Class Markov Models for Measuring Longitudinal Fuzzy Poverty

Giovanni Marano, Gianni Betti, and Francesca Gagliardi

Abstract

The traditional approach to poverty measurement utilises only monetary variables as indicators of individuals' intensity of the state of deprivation, causing measurement errors of the phenomenon under investigation. Moreover, when adopted in a longitudinal context, this approach tends to overestimate transition poverty. Since poverty is not directly observable, a latent definition can be adopted: in such a conception is possible to use Markov chain models in their latent acceptation. This chapter proposes to use Latent class Markov models which allow taking into account more observed (manifest) variables. We define those variables via monetary and non-monetary fuzzy indicators.

Keywords

Poverty • Fuzzy sets • Markov models

Abbreviations

AIC	Akaike Information Criterion
EU-SILC	EU Statistics on Income and Living Conditions
EM	Expectation-Maximization
FM	Fuzzy Monetary
FS	Fuzzy Supplementary
TFR	Totally Fuzzy and Relative

G. Marano • G. Betti (✉) • F. Gagliardi
Department of Economics and Statistics, University of Siena, Siena, Italy
e-mail: giovannimarano4@gmail.com; gianni.betti@unisi.it; gagliardi10@unisi.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_4, Published online: 28 October 2014

1 Introduction

The conceptualisation and assessment of poverty involves at least four dimensions: multiple aspects covering both monetary and non-monetary deprivation; diverse statistical measures in order to capture different facets of poverty; the time-dimension; and comparability over space and time. This chapter addresses the time-dimension of poverty, which involves two types of measures: (1) measures of poverty trend over time at the aggregate level; (2) measures of persistence or otherwise of poverty at the micro level. In particular, attention is focused on the distinction between transitory and persistent poverty, which cannot be distinguished in a traditional cross-sectional analysis, see Hoy and Zheng [9], and Costa and De Angelis [8] for the most recent contributions to the issue.

Moreover, according to Verma and Gagliardi [15] "*...The main measurement problem in the assessment of poverty trends is the definition of the poverty threshold and its consistency over time. The main measurement problem in the assessment of persistence of poverty concerns the effect of random measurement errors on the consistency of the individuals' observed poverty situation at different times...*"

In fact, poverty is not directly observable from sample surveys, especially when monetary variables are collected as indicators of individuals' intensity of the state of deprivation: this causes measurement errors in the phenomenon under investigation. Moreover, when adopted in a longitudinal context, the traditional approach tends to overestimate transition poverty, since small changes in the measured income or consumption expenditure can result in an individual crossing the poverty line. To overcome these problems Betti [2] adopted a latent definition for poverty: in such a conception it is possible to use Markov chain models to correct measurement error in panel data.

The novelty of the chapter consists of: (i) using Latent class Markov models which allow taking into account more observed (manifest) variables; and (ii): defining those variables via monetary and non-monetary fuzzy indicators.

The chapter is composed of five sections. After the present introduction, Sect. 2 describes the latent class Markov models recently proposed in the literature and used for analysing poverty dynamics in this chapter. Section 3 describes the fuzzy set approach to measure monetary and non-monetary (multidimensional) poverty cross-sectionally. Section 4 reports the results of the empirical analysis conducted on the 4 year balanced panels form EU-SILC waves 2006–2009 for France, Italy and United Kingdom, while the final section "Concluding Remarks and Further Research" concludes the chapter.

2 Latent Class Markov Models

A Markov chain is a discrete-time stochastic process characterized by the basic assumption that the next state only depends on the current state, regardless of the whole past history: these processes are suitable whenever one wants to model a

situation where the future is independent of the past, given the present. A natural extension of Markov chains are the Hidden Markov Models or Latent Class Markov Models [5]: such models are characterized by an unobservable discrete Markov process and by a sequence of observable responses, which could be defined on a discrete or continuous support.

The unobservable process is called "hidden" or "latent" since we only observe the output responses, and each of these responses is a deterministic or stochastic function of the current latent state. Therefore, Latent Class Markov Models may be seen as a bivariate stochastic process where two basic assumptions hold:

$$P(X_t|X_{t-1},\ X_{t-2}\ \ldots\ X_1,\ Y_t) = P(X_t|X_{t-1}) \tag{1}$$

$$P(Y_t|X_t,\ X_{t-1}\ \ldots\ X_1,\ Y_{t-1},\ Y_{t-2}\ \ldots\ Y_1) = P(Y_t|X_t) \tag{2}$$

the latent state at the current moment, given the one at the immediately previous moment, is independent of the all previous states, and the observed output at the current moment, given the corresponding latent state, is independent of all previous observations.

In our study, latent variables are considered as the real individual's states, i.e. the latent meaning of variable under investigation, whilst output variables are considered as responses affected by errors, because they are collected by a survey.

Latent Class Markov Models are divided into discrete and continuous, according to their support. In discrete Latent Class Markov Models, two fundamental matrices, along with an initial vector, need to be estimated: a matrix for transition probabilities between latent states, a matrix for response probabilities (i.e. probabilities of observing responses, given the latent state) and a vector for initial probabilities of being in each latent class. Rabiner and Juang [12] give an exhaustive overview of methods and problems related to these models, providing applications on speech recognition.

In continuous Latent Class Markov Models, the matrix of conditional response probabilities is replaced by a matrix containing the response distribution parameters, conditionally on each latent state.

In this chapter we focus attention on application of Latent Class Markov Models to panel data [1, 10], i.e. an extension of the hidden Markov structure for univariate time series towards models describing more individual processes. This feature marks the difference in terminology between Hidden Markov Models and Latent Class Markov Models. The first typically refers to data structures with long time series and small number of individuals, whereas the second typically refers to data structures with a large number of individuals and few instants of time.

Moreover, we will handle the effects of some covariates on the output responses, by using these models in a regression contest. Generalized Linear Models [11] are a generalization of ordinary linear regression, which allows for non-normally distributed response variables: regression can be fitted for response variables belonging to exponential family, by means of some link functions transforming

the expected value in a linear predictor. We will extend these procedures to Latent Class Markov Models for multiple output sequences by supposing that the observed responses of Markov process are affected by a set of known variables; finally, an expected value of output variable given the latent state and the value of covariates will be obtained. Above models will be applied to longitudinal data from a panel survey, i.e. a survey repeated on the same sample over time, suitable whenever the goal is to study the dynamic population processes along with estimates of particular parameters such as net and gross changes in the population. Some device such as a rotating scheme may be used in a panel in order to maintain representativeness of the sample.

We will use the Expectation-Maximization (EM) algorithm in order to estimate model parameters. It is a two-steps algorithm in which at first a conditional expectation of model log-likelihood is computed, according to the current values of parameters, and then a new estimation of parameters is obtained by maximizing this function; the two steps are iteratively repeated till convergence.

3 Fuzzy Measures of Monetary and Non-Monetary Deprivation

In the traditional approach, poverty is characterized by a simple dichotomization of the population into poor and non-poor defined in relation to some chosen poverty line. This poverty line may represent a certain percentage of the mean or the median of the income distribution. This approach presents two main limitations: firstly, it is unidimensional, since it refers to only one proxy of poverty; and secondly, it divides the population into a simple dichotomy.

However, poverty is a complex phenomenon that cannot be reduced solely to monetary dimension but it must also take account of non–monetary indicators of living conditions; moreover, it is not an attribute that characterises an individual in terms of presence or absence, but is rather a predicate that manifests itself in different shades and degrees.

The fuzzy approach considers poverty as a matter of degree rather than an attribute that is simply present or absent for individuals in the population. An early attempt to incorporate the concept of poverty as a matter of degree at methodological level was made by Cerioli and Zani [6] who drew inspiration from the theory of Fuzzy Sets. Subsequently, Cheli and Lemmi [7] proposed the so called Totally Fuzzy and Relative (TFR) approach in which the membership function—quantitative specification of individuals' or households' degrees of poverty and deprivation—is defined as the distribution function of income, normalised (linearly transformed) so as to equal 1 for the poorest and 0 for the richest person in the

population. Betti and Verma [4] defined the membership function of monetary and non-monetary deprivation for any individual i as:

$$\mu_{i,K} = \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma | X_\gamma > X_i}{\sum_{\gamma=2}^{n} w_\gamma | X_\gamma > X_1} \right)^{\alpha_K - 1} \left(\frac{\sum_{\gamma=i+1}^{n} w_\gamma X_\gamma | X_\gamma > X_i}{\sum_{\gamma=2}^{n} w_\gamma X_\gamma | X_\gamma > X_1} \right) \tag{3}$$

where X is the equivalised income in the monetary deprivation, or the overall score s in the non-monetary deprivation; w_γ is the sample weight of individual of rank γ ($\gamma = 1, \ldots, n$) in the ascending distribution, and α_K (K = 1, 2) are two parameters. Each parameter α_K is estimated so that the mean of the corresponding membership function is equal to the at-risk-of-poverty rate computed on the basis the official Eurostat poverty line (60 % of the median income). The monetary based indicator has been termed as Fuzzy Monetary (FM), while the non-monetary indicator as Fuzzy Supplementary (FS).

4 Empirical Analysis

In the present chapter we have used data from the EU Statistics on Income and Living Conditions (EU-SILC) survey, distributed by Eurostat. The EU-SILC survey is designed to collect detailed information on the income of each household member, on various aspects of the material and demographic situation of the household, and for producing structural indicators on social cohesion. EU-SILC surveys involve a rotational panel design conducted annually in each country. The national sample designs and sizes have been determined primarily for the purpose of estimation and reporting of indicators at the national level, with limited breakdown by major socio-demographic subgroups of the population. It provides two types of annual data: (i) cross-sectional data; and (ii) longitudinal data, pertaining to individual-level changes over time, observed annually over a 4 year period for each person.

In the present work we have used the longitudinal 2009 data set covering years from 2006 to 2009. For each wave, we have calculated the FM and the FS measures according to Eq. (1); moreover, following Betti et al. [3], we have reduced the total number of items (supplementary variables) adopted for constructing the overall non-monetary measures (FS) into meaningful dimensions. Exploratory and confirmatory factor analyses, as proposed in Whelan et al. [16], allow us to achieve this objective. We have applied such procedures to three EU countries—France, Italy and United Kingdom—obtaining six dimensions:

1. FS1 = Basic lifestyle;
2. FS2 = Consumer durables;
3. FS3 = Housing amenities;

Table 1 Mean poverty and deprivation measures in cross-sectional and balanced panel samples

	H_cs	*FM*	*FS*	*FS1*	*FS2*	*FS3*	*FS4*	FS5	FS6
FR_06	13.2	12.7	13.4	10.1	6.1	5.2	29.8	11.2	17.0
FR_07	13.1	12.4	13.1	9.8	5.6	5.2	30.1	10.9	16.5
FR_08	12.7	11.9	12.8	9.7	5.3	6.0	7.0	10.5	16.2
FR_09	12.9	11.9	12.7	9.8	5.1	5.9	8.1	7.0	17.4
IT_06	19.6	19.8	18.5	15.7	6.9	5.8	24.3	14.4	21.5
IT_07	19.8	19.8	17.6	14.6	6.2	5.6	26.9	14.8	22.8
IT_08	18.7	18.6	17.0	15.0	6.0	5.4	26.3	13.7	21.6
IT_09	18.4	18.4	16.8	13.9	5.3	6.0	9.4	7.3	21.1
UK_06	19.0	17.9	18.0	12.8	6.5	7.1	27.3	13.3	17.4
UK_07	18.6	16.9	16.7	13.1	5.9	6.1	22.8	9.4	16.2
UK_08	18.7	17.7	17.9	13.3	6.1	6.0	26.4	13.5	14.8
UK_09	17.3	17.0	17.7	13.3	5.8	5.8	27.4	13.8	15.0

4. FS4 = Financial situation;
5. FS5 = Work & Education;
6. FS6 = Health related.

Table 1 reports average poverty and deprivation measures for the three countries under study; column *H_cs* shows the percentage of poor individuals (the at-risk-of-poverty rate) based on the cross-sectional SILC samples and published by Eurostat. All the remaining columns are based on the balanced 4 year panels:[1] those averages are weighted by SILC target variable RB064 (Longitudinal weight, 4 year duration), which lets the balanced panels be representative of the population present continuously in the country over the period. This variable is constructed via a calibration procedure based on a set of post-stratification variables (see [13] for details). It is possible to observe that, in the case of France and UK, FM (and FS) differs significantly from *H_cs*: the tests have been performed by estimating measures' standard errors with JRR method [14]. There are at least two explanations: (i) "longitudinal" population may differ from cross-sectional population; (ii) longitudinal weights RB064 have been calibrated by Eurostat using variables which are not so highly correlated to monetary poverty. This leads to a new source of measurement errors, which could be corrected by the latent dimension intrinsic in the Markov models.

Generalized Linear Models have been used to fit the data: we have started by modelling FM and FS singularly as Gamma-distributed response variables with inverse canonical link function, and then we have put them together in a bivariate model, i.e. a model with a shared hidden Markov chain and two response variables.

[1] These balanced panels are the base for the analysis via Latent class Markov Models.

Table 2 Best set of covariates

Country	Sex	Age	Education level	Marital status	Macro-region
FR	√	√	√	√	–
IT	√		√		
UK		√	√		–

We also have tried to insert FS1–FS6, but their distributions are not suitable to be fitted as Gamma variable, therefore they need some appropriate transformations to be implemented in further research work. Latent class Markov models need to be initialised with starting values for parameters: as initial values we have used parameters obtained from a similar analysis carried out according to the traditional approach. Covariates introduced into the models were: sex, age, education level and marital status; macro-regions have also been introduced for Italy.

For each state, we have tested all combinations of covariates, in order to find the best model by minimising Akaike Information Criterion (AIC). The influence of covariates is measured on the bivariate FM-FS response variable. In Table 2 the best set of covariates is described.

The final model for France includes sex, education level, age and marital status: a lowest education level and a marital status in the set of "widowed", "separated" and "divorced" have a strong negative effect on both FM and FS poverty status; the female gender has also a double negative effect, although lower; conversely, an age between 35 and 64 has a positive effect on FM and a negative one on FS poverty status.

The final model for Italy includes sex and education level: particularly, female gender and lowest education level have a negative effect on both FM and FS poverty status. Models with macro-region covariate were also tried, but they have a higher AIC. The final model for UK includes the education level again, along with the age; the first has the same effect described before, while the latter has two different effects, one for 35–64 years old people and one for 65 or more years old people. They both have a negative effect on FM poverty status, compared to the 0–34 baseline category; on FS poverty status the effect of 35–64 category is still negative, but the effect of 65+ is positive instead, compared to the 0–35 baseline category. Latent state only affects the magnitude of coefficients, but it does not affect their sign.

Table 3 reports initial and transition probabilities estimated via Latent class Markov models. These refer to bivariate models with shared hidden Markov chain and with covariates included. Initial non-poverty probability is higher in France (85.7 %) than in UK and Italy (81.8 % and 79.7 % respectively). Transition matrices show high probabilities of remaining in the non-poor state (around 98 %) and low probabilities of moving towards the poor state for all countries. Conversely, transition probabilities from poor to non-poor state vary from 12.2 % for Italy to 14.4 % for UK.

Table 3 Initial and transition probabilities, Latent class Markov models

Country	Latent states	Initial probabilities	Transition probabilities	
			Poor	Non-poor
FR	Poor	0.143	0.871	0.129
	Non-poor	0.857	0.020	0.980
IT	poor	0.203	0.878	0.122
	Non-poor	0.797	0.020	0.980
UK	poor	0.182	0.856	0.144
	Non-poor	0.818	0.029	0.971

Concluding Remarks and Further Research

The above empirical evidence shows the benefit of the use of Latent class Markov models in longitudinal poverty analysis, since this result was not clear from simply observing cross-sectional analysis along time. However, a possible limitation in utilising Markov models in analysing poverty dynamics consist in the fact that a Markov chain status at time t depends only on status at time t-1, while the status of a poor unit at time t may depend on its situation in previous periods as well.

Since this is the first attempt to apply Latent class Markov models for measuring longitudinal fuzzy poverty, we aim to improve the present chapter from a theoretical point of view by identifying a larger set of link functions which could better fit the manifest variable distributions, and improve the work from an empirical point of view by performing a full comparative analysis among the 27 EU countries, and including in the models the non-monetary dimensions FS1–FS6.

Bibliography

1. Bartolucci, F., Farcomeni, A., Pennoni, F.: Latent Markov Models for Longitudinal Data. Chapman and Hall/CRC Press, Boca Raton (2012)
2. Betti, G.: A longitudinal approach to poverty analysis: the latent class Markov model. Statistica **56**, 345–359 (1996)
3. Betti, G., Gagliardi, F., Lemmi, A., Verma, V.: Sub-national indicators of poverty and deprivation in Europe: methodology and applications. Camb. J. Regions Econ. Soc. **5**, 149–162 (2012)
4. Betti, G., Verma, V.: Fuzzy measures of the incidence of relative poverty and deprivation: a multi-dimensional perspective. Stat. Meth. Appl. **17**, 225–250 (2008)
5. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state Markov chain. Ann. Math. Stat. **37**, 1554–1563 (1966)
6. Cerioli, A., Zani, S.: A fuzzy approach to the measurement of poverty. In: Dagum, C., Zenga, M. (eds.) Income and Wealth Distribution, Inequality and Poverty, pp. 272–284. Springer, Berlin (1990)
7. Cheli, B., Lemmi, A.: A totally fuzzy and relative approach to the multidimensional analysis of poverty. Econ. Notes **24**, 115–134 (1995)

8. Costa, M., De Angelis, L.: A dynamic latent model for poverty measurement. Comm. Stat. Theor. Meth. (2014)
9. Hoy, M., Zheng, B.: Measuring lifetime poverty. J. Econ. Theory **146**, 2544–2562 (2011)
10. Maruotti, A.: Mixed hidden Markov models for longitudinal data: an overview. Int. Stat. Rev. **79**, 427–454 (2011)
11. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman & Hall/CRC, Boca Raton (1989)
12. Rabiner, L.R., Juang, B.H.: An introduction to hidden Markov models. IEEE ASSP Mag. **3**(1), 4–16 (1986)
13. Verma, V., Betti, G., Ghellini, G.: Cross-sectional and longitudinal weighting in a rotational household panel: applications to EU-SILC. Stat. Transit. **8**, 5–50 (2007)
14. Verma V., Betti G.: Taylor linearization sampling errors and design effects for poverty measures and other complex statistics, *Journal of Applied Statistics*, **38**(8), 1549–1576 (2011)
15. Verma, V., Gagliardi, F.: On assessing the time-dimension of poverty. In: Betti G., Lemmi A. (eds.) Poverty and Social Exclusion: New Methods of Analysis, pp. 109–127. Routledge, London/New York (2013)
16. Whelan, C.T., Layte, R., Maitre, B., Nolan, B.: Income, deprivation and economic strain: an analysis of the European community household panel. Eur. Socio. Rev. **17**, 357–372 (2001)

A Latent Class Approach for Allocation of Employees to Local Units

Davide Di Cecco, Danila Filipponi, and Irene Rocchetti

Abstract

In 2011, the Italian Business Register has been reshaped as a database of single workers microdata. Determining the workplace of each individual provides the National Statistical Institute (NSI) with huge information potential. Unfortunately, the administrative sources at our disposal do not always allow a reliable determination of the workplace of each worker. We present a probabilistic methodology to assign a workplace to each employee by assigning him to one of the local units of the enterprise he works for. We used a Latent Class Model to estimate the probability of each employee to belong to each local unit. We assumed the total number of employees per local unit as a constraint. A computationally intensive optimization problem has been solved for each of the ca. 200 thousands multilocated enterprises. The results refer to year 2011.

Keywords

Administrative data • Latent class analysis • Linear programming

1 Introduction

The 2011 Census Industry provided an opportunity for a methodological and conceptual revision of the Italian Business Register (hereafter BR). That process resulted in widening the BR information content, thanks to the use of new administrative sources, that made possible to release data on individual workers, while formerly only enterprise data were available. By doing so, we created a so-called Linked Employer-Employee Database (LEED). This data structure

D. Di Cecco (✉) • D. Filipponi • I. Rocchetti
ISTAT, via Oceano Pacifico 171, Roma, Italy
e-mail: dicecco@istat.it; dafilipp@istat.it; irocchetti@istat.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_23, Published online: 12 November 2014

defines a direct relationship between employees and enterprises. The traditional BR information concerning employment and employment composition at enterprise level (e.g. gender and age composition) can then be reconstructed by summing over the individual information.

In the traditional BR, our possibilities in terms of geographical information were limited to statistics on enterprises by local areas. In the new setting, where a worker-level variable indicating the workplace is available, we can associate a geographical information to all variables in that record, thus enhancing substantially our information capability. At the very least, we could disseminate statistics on demographic characters of the employees (gender, age, ...) and of their job (type of contract, salary, ...) by local areas (municipalities or even finer areas). At a deeper level, it is easy to see the potential for spatial–temporal statistical analysis of such a database of microdata.

However, while the workplace determination is immediate for workers of enterprises consisting of a sole local unit, this is not the case for multilocated enterprises, which represent over one third of the total employment. In fact, as we will see in the next section, none of the available administrative sources can be used to identify with certainty the local unit to which each individual belongs.

As a consequence, the process is split into two steps: at a macro level, the NSI produces the usual information on local units through the Local Units Register (LUR), using information from administrative sources and statistical surveys on local unit structure of enterprises. Then, at a micro level, we use a probabilistic methodology to assign each worker to a single local unit in consistency with the aggregated information provided by the LUR.

In the present work we focus on the second step. We illustrate the probabilistic approach based upon Latent Class Analysis (see [1]), to associate to each individual, his probability of working in each local unit. Then, we describe the optimization algorithm that is used to assign the employees so to conform with the LUR information.

This paper is organized as follows: All the statistical and administrative sources used in the process are described in Sect. 2. The probabilistic method developed is presented in Sect. 3. In Sect. 4 some conclusions are drawn and possible future works are discussed.

2 Statistical and Administrative Sources

Since 2011, some new administrative sources are being considered for the realization of the BR, and integrated in its production process.

Of these, the most important is the EMENS database, managed by the Social Security Authority. The EMENS records monthly employer declarations on job positions for all employees. Each record refers to a single job contract and reports a number of information on the type of contract (fixed term, permanent, ...), as

well as relevant events that occurred during the year (number of worked days per month, number of days in maternity leave, sick leave, ...). In particular, the number of worked days is used to assign to each job contract a weight $q \in [0, 1]$, which is defined as the proportion of worked days during the year. The importance of these weights will be explained in the next section. The available information for the identification of the workplace is the municipality where the employee works.

Secondly, the National Insurance Agency (INAIL) records declarations made by the enterprises about their employees, as beneficiaries of compulsory insurance against work-related accidents and diseases. The variable of interest for us is the workplace (address) where each worker is insured.

Finally, the Tax Register provides us all legal places of residence of persons holding an Italian tax code. The addresses of the workers of the multilocated enterprises have been geocoded.

Summarizing, the administrative variables of interest are the following:

- individual workplace's municipality (according to EMENS);
- individual workplace's address (according to INAIL);
- individual permanent address (sourced by the Tax Register).

An exploratory analysis, carried out by comparing EMENS with INAIL, revealed the existence of many mismatches with regards to the workplace. We observed two kind of issues:

coverage issues: a large number of employees is not covered by any of the two sources. INAIL, in particular, covers just about 60 % of the workers recorded by the BR;
coherence issues: the two sources do not agree on about 40 % of the individual workplaces.

As said, the NSI produces, in a separate process, a yearly enterprises database, the Local Unit Register (LUR), based on administrative data and on a survey addressed to the larger enterprises. The LUR reports, for each enterprise, the active local units, their geocoded locations, and their yearly mean number of employees.

The production of the LUR is a long-standing well-established process, based on reliable information on local units structure. In fact, it includes clerical review of the survey results for the larger enterprises. For these reasons, we bind the employees workplace allocation process to the LUR; that is, the allocation is forced to fit the yearly mean number of employees by local units as reported by the LUR.

The EMENS and INAIL information, aggregated at a local unit level, differs considerably from the values provided by the LUR. For example, some minor municipalities having local units according to the LUR are absent from both INAIL

and EMENS. The low quality of the administrative sources, together with the requirement of consistency with the LUR, convinced us of the necessity of a probabilistic approach.

3 Methodology

Although we deal, essentially, with a problem of imputation of partial missing information, the approach we developed is strictly related to Record Linkage works, (see e.g. [2] and [3]). Given a multilocated enterprise, the BR provides the list of workers of that enterprise, while the LUR provides the list of the active local units. We consider all the possible couples, employee-potential local unit of membership, and choose the "correct" ones (we would say the correct links in a Record Linkage context). To formalize things, in a multilocated enterprise with n employees and k active local units, let us denote as

- U_j, $j = 1, \ldots, k$, its local units;
- P_i, $i = 1, \ldots, n$, its employees;
- q_i, $i = 1, \ldots, n$, the yearly mean weights associated to the employees.

For each enterprise, we construct the cartesian product of $\{P_i\}_{i=1,\ldots,n}$ and $\{U_j\}_{j=1,\ldots,k}$, and, for each of the $n \times k$ pairs $\left(P_i, U_j\right)$, we record the values of the following three indicators:

- $D_1(i, j)$—indicating whether workplace's municipality of employee P_i, recorded in EMENS, coincides with address municipality of local unit U_j recorded in the LUR. In detail, $D_1(i, j) \in \{1, 2, 3\}$ where $D_1(i, j) = 1$ if the Emens municipality is missing, 2 if the municipality is different, and 3 if the municipality is the same.
- $D_2(i, j)$—measuring the coherence between the address of Employee's workplace, as reported by Inail, and local unit's address, as recorded by LUR. $D_2(i, j) \in \{1, 2, 3, 4\}$, where 1 indicates that Inail information is missing, 2 indicates that the municipality is different, 3 indicates that the municipality is the same, and 4 indicates that local unit's address is recorded as the workplace by Inail.
- $D_3(i, j)$—measuring the distance between the geocoded employee's residence and the geocoded local unit's address. In this case, we categorize the distance into four classes corresponding to indicator values 1,2,3,4 from the farthest to the closest distance respectively.

The probability for the i-th employee P_i to belong to the j-th local unit U_j, i.e., the probability of the couple (P_i, U_j) of being correct, has been estimated through a *Latent Class Model*. In this model, each couple (P_i, U_j) represents a unit. The number of latent classes has been fixed to 2, under the assumption that a two-dimensional latent structure underlies the analyzed phenomenon, one class indicating the individual "membership" and the other one indicating the individual "non membership".

We could have chosen a different number of Latent Classes, as in [3], with the aim of investigating whether different latent structure assumptions could be made. But, in our case, the decision framework of the problem is somehow different from a typical Record Linkage problem as described in Fellegi–Sunter's theory. In fact, we do not have to define a "critical area", as we do not have the possibility of a clerical review of the so-called "uncertain links". We just have to assign each employee to a local unit in such a way that the yearly average total number of employees, given by the sum of the weights of the employees assigned to that local unit, coincides with that reported in the LUR. For this reason, once the probabilities of each couple of being correct have been estimated, we approached the allocation problem simply as an optimization problem, where the quantity to be maximized is the sum of the probabilities of the accepted couples (P_i, U_j). The idea of using algorithms from Operational Research for the optimal allocation in a Record linkage problem dates back at least to [2], but there the author still use them in combination with thresholds deriving from Fellegi–Sunter's theory, to define the certain and uncertain links.

3.1 The Latent Class Model

If we denote with

- L the latent variable with two classes $c = 1, 2$;
- γ_c the prior probability to belong to the latent class c;
- $\rho_{s,r_s|c}$ the probability for variable D_s, $s = 1, 2, 3$ to assume modality r_s conditionally to $L = c$

then, the posterior probability $p_{i,j}(c)$ for the pair (P_i, U_j) to belong to the class c conditionally to the observed values $d(i, j) = (d(i, j)_1, d(i, j)_2, d(i, j)_3)$ of the variables $D = (D_1, D_2, D_3)$ is

$$p_{i,j}(c) = P(L = c \mid D = d(i, j)) = \frac{\left(\prod_s^3 \prod_{r_s^{R_s}} \rho_{(s,r_s|c)}^{I\{d_s(i,j)=r_s\}}\right)\gamma_c}{\sum_{c=1}^2 \left(\prod_s^3 \prod_{r_s^{R_s}} \rho_{(s,r_s|c)}^{I\{d_s(i,j)=r_s\}}\right)\gamma_c}$$

Each triple of possible values for the indicators D_1, D_2, D_3 is called a *pattern*. Since the three indicators have respectively 3, 4, and 4 modalities, we have 48 possible patterns. The total number of free parameters to be estimated is equal to the number of free (ρ parameters) for each Latent Class, plus the number of free prior probability (γ parameters), that is, in our model, $(2 + 3 + 3) \cdot 2 + 1 = 17$.

As the BR has millions of local units and millions of employees, we have several billions of possible couples (P_i, U_j). For that reason, we drawn a sample of enterprises and estimated the model on the subset of couples coming from the enterprises of that sample. The model has been estimated various times on different samples of different sizes to test for the robustness of the estimates, and, as a matter of fact, they resulted to be very stable across different samples.

We used an EM algorithm to find the ML estimates of the 17 model parameters. However, we just report in Table 1 the posterior probabilities obtained for each possible pattern, as they have a clearer role in the interpretation of the model, and are central for the second part of our methodology.

In fact, analyzing Table 1, we can see that higher values of the indicators correspond to higher posterior probabilities for Latent Class 1. That is, whenever D_1 and D_2 indicate a coherence between the workplace reported in our sources and the local unit's location, and D_3 indicates that the local unit is close to the individual residence, we have a higher posterior probability for that class. Moreover, the estimated probabilities in many cases are very close to 0 or very close to 1. This fact denotes a good "latent class separation" (see [1]). A common approach to quantify latent class separation in posterior classification is based on entropy indexes. One of the most common indexes is proposed by [5] and is defined as a weighted average of individual's posterior probabilities:

$$E = 1 - \frac{\sum_{i=1}^{n} \sum_{c=1}^{C} -p_i(c) \log p_i(c)}{n \log C},$$

where $p_i(c)$ is unit i's posterior probability of membership in latent class c, and C is the number of latent classes. E ranges between 0 and 1, with larger values indicating better latent class separation. In our results we get $E = 0.9$, indicating a very good grade of separation between the two classes. Secondly, the (ranking of the) values $\hat{p}_{i,j}(1)$ obtained for the patterns are perfectly consistent with common intuition. In fact, couples of individual and local units confirmed as correct by EMENS and INAIL have higher posterior probabilities, which decrease as the two sources disagree. The farther the local unit from the employees residence, the lower the probability. Moreover, the importance of that distance, i.e., of D_3, is far lower than that of the other two indicators D_1 and D_2, and that is highly desirable.

These evidences allowed us to heuristically identify Latent Class 1 as the one characterizing "membership" of the employee to the local unit. Hence, the estimated posterior probability $\hat{p}_{i,j}(1)$ of a couple (P_i, U_j) of belonging to Latent Class 1 has been interpreted as the probability that the worker P_i belongs to the local unit U_j,

Table 1 The estimated posterior probabilities of belonging to the two Latent Classes conditionally on the values of (D_1, D_2, D_3).

D_1	D_2	D_3	$\hat{p}_{i,j}(1)$	$\hat{p}_{i,j}(2)$	D_1	D_2	D_3	$\hat{p}_{i,j}(1)$	$\hat{p}_{i,j}(2)$
1	1	1	0.3734	0.6265	2	3	1	0.5788	0.4211
1	1	2	0.0545	0.9454	2	3	2	0.1173	0.8826
1	1	3	0.2284	0.7715	2	3	3	0.4058	0.5941
1	1	4	0.6687	0.3312	2	3	4	0.8231	0.1768
1	2	1	0.1035	0.8964	2	4	1	0.4952	0.5047
1	2	2	0.0112	0.9889	2	4	2	0.0866	0.9133
1	2	3	0.0542	0.9457	2	4	3	0.3277	0.6722
1	2	4	0.2812	0.7187	2	4	4	0.7687	0.2312
1	3	1	0.9441	0.0558	3	1	1	0.9641	0.0358
1	3	2	0.6206	0.3793	3	1	2	0.7224	0.2775
1	3	3	0.8936	0.1063	3	1	3	0.9304	0.0695
1	3	4	0.9828	0.0171	3	1	4	0.9891	0.0108
1	4	1	0.9235	0.0764	3	2	1	0.8391	0.1608
1	4	2	0.5387	0.4612	3	2	2	0.3353	0.6646
1	4	3	0.8571	0.1428	3	2	3	0.7215	0.2784
1	4	4	0.9761	0.0238	3	2	4	0.9464	0.0535
2	1	1	0.0461	0.9538	3	3	1	0.9986	0.0013
2	1	2	0.0046	0.9953	3	3	2	0.9866	0.0133
2	1	3	0.0234	0.9765	3	3	3	0.9973	0.0026
2	1	4	0.1409	0.8590	3	3	4	0.9996	0.0003
2	2	1	0.0092	0.9907	3	4	1	0.9981	0.0018
2	2	2	0.0009	0.9992	3	4	2	0.9813	0.0186
2	2	3	0.0046	0.9953	3	4	3	0.9963	0.0036
2	2	4	0.0308	0.9691	3	4	4	0.9994	0.0005

that is, as the probability of (P_i, U_j) of being a correct couple. In the following, these probabilities will be denoted simply as $p_{i,j}$.

3.2 The Allocation Process

Formally, a *Generalized Assignment Problem* (GAP) needs to be solved in the aim of assigning individuals to a sole local unit. For each enterprise, we aim at finding the $x_{i,j} \in \{0, 1\}\ i = 1, \ldots, n\ j = 1, \ldots, k$ maximizing

$$\sum_{j=1}^{k} \sum_{i=1}^{n} p_{i,j}\ x_{i,j}$$

subject to the constraints:

$$\sum_{i=1}^{n} x_{i,j} q_i = n_j \qquad \forall j = 1, \ldots, k \tag{1}$$

and

$$\sum_{j=1}^{k} x_{i,j} = 1 \qquad \forall i = 1, \ldots, n \tag{2}$$

where $x_{i,j} = 1$ if employee P_i is assigned to the j-th local unit U_j and n_j represents the *fixed* number of employees to be assigned to the j-th local unit given by the LUR. The problem has been solved separately for each of the about 200 thousand multilocated enterprises utilizing a specific SAS routine (see [4]). The biggest enterprises, leading to a cartesian product of billions of possible couples (P_i, U_j), required a form of blocking strategy; that is, we restricted the optimization problem to the local units of a single region (or province or smaller areas when necessary) and went on iteratively. In detail, let $n_{r_1}, n_{r_2}, \ldots$ be the total number of employees to be assigned to the local units of region 1,2,... according to the LUR, and let $k_1, k_2, \ldots$ be the corresponding number of local units per region. Then we solve the optimization problem for the first region with the following constraints:

$$\sum_{i=1}^{n} x_{i,j} q_i = n_{r_1} \qquad \forall j = 1, \ldots, k_1$$

$$\sum_{j=1}^{k} x_{i,j} \leq 1 \qquad \forall i = 1, \ldots, n$$

substituting (1) and (2) above, and go on with the second region with the remaining $n - n_{r_1}$ employees.

4 Discussion and Future Works

Our method provides an estimate of employees' workplace using an unsupervised probabilistic model conditioned to an accurate information coming from the LUR. As a result, we integrated our LEED database with an individual variable indicating the workplace. This information can be used to produce several statistics on demographic and job characters of the employees by geographical areas.

We show that the latent class approach to estimate the probability that a worker belongs to a specific local unit works well. However, accurate estimation of error rates for different statistical domains are not available. Here, resembling the Record Linkage literature, the error rate is composed by "false match", i.e., workers

assigned to a wrong local unit, and "false non-match", i.e., workers not assigned to the correct local units. If the set $\hat{M}$ contains the couples (P_i, U_j) identified as matching, the set $\hat{U}$ contains the couples identified as non-matching, and U and M are, respectively, the sets containing the false and the true matches, then the false match rate is given by $P(U \mid \hat{M})$ and the false non-match rate is given by $1 - P(M \mid \hat{U})$. Winkler in [6] demonstrated that, if a properly chosen training dataset is available, then we could obtain accurate estimates of error rates. At the moment of this writing, no benchmark is available. As data from the 2011 Population Census will be available, it would be possible to estimate the error rate.

The Latent Class approach proposed in this paper has been applied to 2011 data. However, we need to produce the same information yearly. The allocation process described in Sect. 3.2 does not ensure that each single worker would be assigned to the most probable local unit, because of the assumed constraints. This fact may give room to implausible changes of workplace for the same individual from one year to another. To insure coherent individual data over time we are exploring the possibility of using longitudinal information. This could be handled by modeling the latent class membership over time with a Hidden Markov Models (see Latent Transition Analysis in [1]) to estimate not only the latent class membership probabilities, but also the transitions probabilities in latent class membership.

References

1. Collins, L.M., Lanza, S.: Latent Class and Latent Transition Analysis. Wiley Series in Probability and Statistics. Wiley, Hoboken (2010)
2. Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. J. Am. Stat. Assoc. **84**, 414–420 (1989)
3. Larsen, M.D., Rubin, D.B.: Iterative automated record linkage using mixture models. J. Am. Stat. Assoc. **96**, 32–41 (2001)
4. Pratt, R., Hughes, E.: Linear Optimization in SAS/OR Software: Migrating to the OPTMODEL Procedure. SAS Global Forum (2011)
5. Ramaswamy, V., DeSarbo, W.S., Reibstein, D.J., Robinson, W.T.: An empirical pooling approach for estimating marketing mix elasticities with PIMS data. Market. Sci. **12**, 103–124 (1993)
6. Winkler, W.E.: Record Linkage and Bayesian Networks. American Statistical Association. In: Proceedings of the Section on Survey Research Methods (2002)

Finding Scientific Topics Revisited

Martin Ponweiser, Bettina Grün, and Kurt Hornik

Abstract

The publication of statistical results based on the use of computational tools requires that the data as well as the code are provided in order to allow to reproduce and verify the results with reasonable effort. However, this only allows to rerun the exact same analysis. While this is helpful to understand and retrace the steps of the analysis which led to the published results, it constitutes only a limited proof of reproducibility. In fact for "true" reproducibility one might require that the essentially same results are obtained in an independent analysis. To check for this "true" reproducibility of results of a text mining application we replicate a study where a latent Dirichlet allocation model was fitted to the document-term matrix derived for the abstracts of the papers published in the Proceedings of the National Academy of Sciences from 1991 to 2001. Comparing the results we assess (1) how well the corpus and the document-term matrix can be reconstructed, (2) if the same model would be selected and (3) if the analysis of the fitted model leads to the same main conclusions and insights. Our study indicates that the results from this study are robust with respect to

M. Ponweiser
Department of Finance, Accounting and Statistics, Institute for Statistics and Mathematics, WU (Wirtschaftsuniversität Wien), Welthandelsplatz 1, 1020 Wien, Austria
e-mail: M.Ponweiser@gmail.com

B. Grün (✉)
Department of Applied Statistics, Johannes Kepler University Linz, Altenbergerstraße 69, 4040 Linz, Austria
e-mail: Bettina.Gruen@jku.at

K. Hornik
Department of Finance, Accounting and Statistics, Institute for Statistics and Mathematics, WU (Wirtschaftsuniversität Wien), Welthandelsplatz 1, 1020 Wien, Austria
e-mail: Kurt.Hornik@wu.ac.at

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_11, Published online: 12 November 2014

slightly different preprocessing steps and the use of a different software to fit the model.

Keywords
Latent Dirichlet allocation • Replication • Reproducibility • Topic model

1 Introduction

Reproducibility of research results is a topic which has recently received increased interest [6, 8]. To ensure easy reproducibility of statistical analyses, data and code are often made available. This allows to rerun the exact same procedures using in general the complete same software environment in order to arrive at the same results [9]. However, as Keiding points out "it ridicules our profession to believe that there is a serious check on reproducibility in seeing if somebody else's computer reaches the same conclusion using the same code on the same data set as the original statistician's computer did [7, p. 377]." True reproducibility therefore would require that an independent analysis arrives at the same results and conclusions, i.e., one might only claim that a result is reproducible, when approximately the same results are obtained if the data preprocessing as well as the model fitting steps are essentially the same, but not necessarily identical. This would imply that the results are robust to small changes in the data preprocessing and model fitting process.

In the following we perform an independent reanalysis of the text mining application published by Griffiths and Steyvers in 2004 [4, in the following referred to as GS2004]. GS2004 use the latent Dirichlet allocation [LDA, 1] model with collapsed Gibbs sampling to analyze the abstracts of papers published in the Proceedings of the National Academy of Sciences (PNAS) from 1991 to 2001. LDA was introduced by Blei and co-authors as a generative probabilistic model for collections of discrete data such as text corpora. Because PNAS is a multidisciplinary, peer-reviewed scientific journal with a high impact factor, this corpus should allow to discover some of the topics addressed by scientific research in this time period.

We try to reproduce the results presented in GS2004 using open-source software with respect to (1) retrieving and preprocessing the corpus to construct the document-term matrix and (2) fitting the LDA model using collapsed Gibbs sampling. In our approach we rerun the analysis without access to the preprocessed data and use different software for model fitting and different random number generation. This allows us to assess if the results are robust to changes in the data retrieval and preprocessing steps as well as the model fitting.

2 Retrieving and Preprocessing the Corpus

In order to reconstruct the corpus web scraping techniques were employed to download the abstracts from the PNAS web page. We ended up with 27,292 abstracts in the period 1991–2001 and with 2,456 in 2001, compared to 28,154 in

Table 1 Summary of the document-term matrices constructed from the abstracts of the PNAS from years 1991 to 2001 by GS2004 and in our replication study

	GS2004	Replication
Vocabulary size	20,551	20,933
Total occurrence of words	3,026,970	2,924,594
Average document length (in terms)	107.51	107.16

1991–2001 and 2,620 in 2001 used by GS2004. This means that we essentially were able to obtain the same number of abstracts. The slight deviations might be due to the fact that our data collection omitted (uncategorized) commentaries, corrections and retractions.

GS2004 did not provide any information if only the abstracts were used or the abstracts combined with the titles. We decided to leave out the paper title for each document because this led to a document-term matrix closer to the one in GS2004. In a first preprocessing step we transformed all characters to lowercase. GS2004 used any delimiting character, including hyphens, to separate words and deleted words which belonged to a standard "stop" list used in computational linguistics, including numbers, individual characters and some function words. We built the document-term matrix with the R [12] package **tm** [2, 3] and a custom tokenizing function which we deduced from the few exemplary terms in the original paper. Our tokenizer treats non-alphanumeric characters, i.e., characters different from "a"–"z" and "0"–"9", as delimiters. This step also implicitly strips non-ASCII characters from our downloaded corpus in Unicode encoding, thereby marginally reducing the information in abstracts which contain characters that are widely used in scientific publications, such as those from the Greek alphabet. The minimum word length was set to two and numbers and words in the "stop" list included in package **tm** were removed. GS2004 further reduced the vocabulary by omitting terms which appeared in less than five documents and we also performed this preprocessing step.

The characteristics of the final document-term matrices are compared in Table 1. Despite the fact that the original set of documents was not the same and a number of preprocessing steps were not clearly specified or slightly differently performed, the final document-term matrices are quite similar with respect to vocabulary size, total occurrence of words and average document length.

3 Model Fitting

GS2004 fit the model using their own software [13]. We use the R package **topicmodels** [5] with the same settings with respect to number of topics, number of chains, number of samples, length of burn-in interval and sample lag. The implementation of the collapsed Gibbs sampler in the package was written by Xuan-Hieu Phan and coauthors [10].

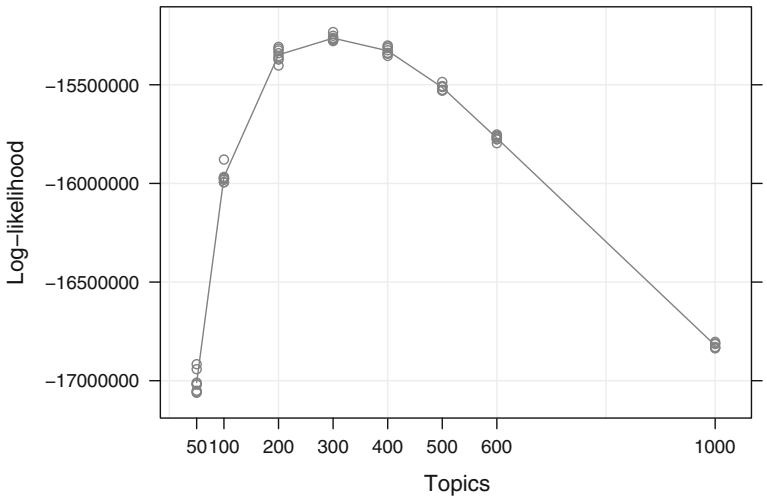

Fig. 1 Estimated marginal log-likelihoods for each number of topics and chain (*circles*). The average marginal log-likelihoods are joint with *lines*

3.1 Model Selection

The number of topics are selected by GS2004 using the marginal log-likelihoods determined by the harmonic mean method. Their results are shown in Fig. 3 of their paper and they decide that 300 topics are a suitable choice. For comparison our results are given in Fig. 1. The figure essentially looks quite similar and would lead to the same decision. In the following the topic model fitted with 300 topics is used for further analysis.

3.2 Scientific Topics and Classes

GS2004 used the 33 minor categories which are assigned to each paper to validate whether these class assignments correspond to the differences between the abstracts detected using the statistical analysis method. Using only the abstracts from 2001 we determined the mean topic distribution for each minor category. The most diagnostic topic was then determined as the one where the ratio of the mean value for this category divided by the sum over the mean values of the other categories was greatest. The results are shown in Fig. 2, which corresponds to Fig. 4 in GS2004. Note that our figure includes all 33 minor categories, whereas in the figure in GS2004 category "Statistics" is missing. Again a high resemblance between the two results can be observed. For comparison the five most probable words for the topic assigned to minor category "Ecology" are "species", "global", "climate", "co2" and "water" in GS2004 and "species", "diversity", "marine", "ecological"

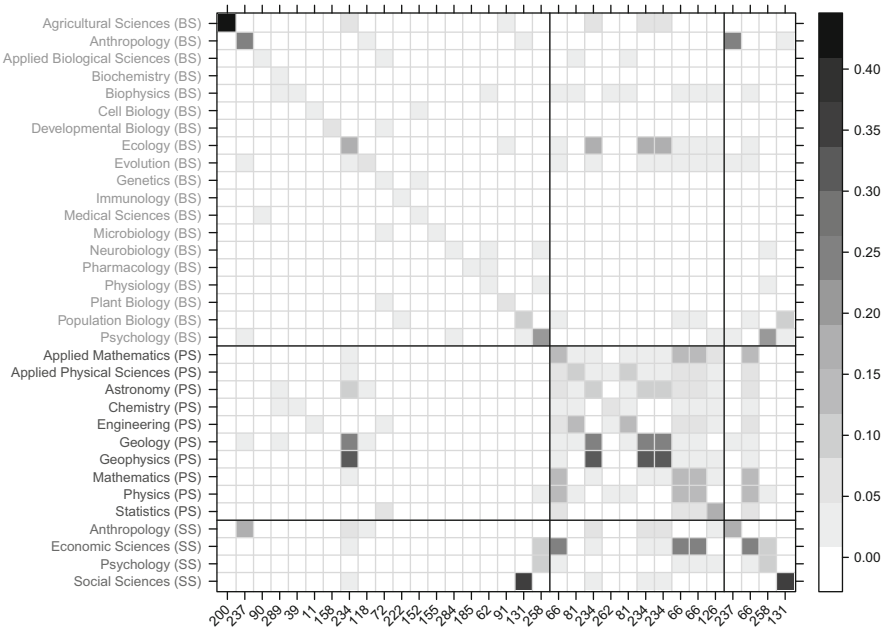

Fig. 2 Mean values of the topic assigned to each of the 33 minor categories based on all abstracts published in 2001. Higher probabilities are indicated with *darker cells*. The abbreviations "BS", "PS" and "SS" denote the major categories Biological, Physical and Social Sciences

and "community" in our replication study. This topic also has high mean values for the minor categories "Geology" and "Geophysics" in both solutions.

3.3 Hot and Cold Topics

In a next step GS2004 analyze the dynamics of the topics using a post hoc examination of the mean topic distribution estimates for each year from 1991 to 2001. A linear trend was fitted to each topic over time and the estimated slope parameters were used to identify "hot" and "cold" topics. The five topics with the largest positive and negative slopes in our model are given in Fig. 3. This figure corresponds to Fig. 5 in GS2004 except that they only show the three "hottest" and "coldest" topics. A comparison of results using the twelve most probable words of each topic indicates that matches for the three topics in GS2004 can be identified among the five topics identified by our model, even though the order of the topics is not identical. The coldest topic detected in each of the analyses is remarkably similar, as indicated by a comparison of the twelve most probable words, which are given in Table 2.

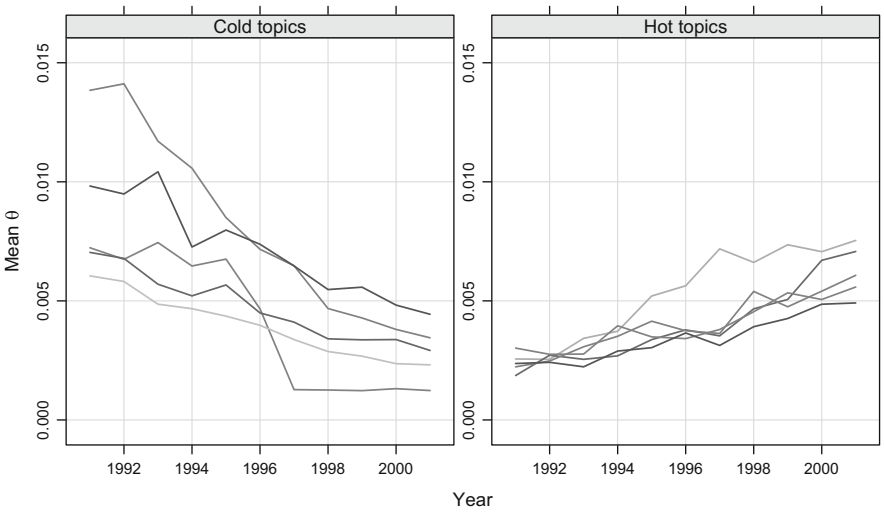

Fig. 3 Dynamics of the five hottest and five coldest topics from 1991 to 2001, defined as those topics that showed the strongest positive and negative linear trends

Table 2 The twelve most probable words for the coldest topic

GS2004	Replication
cdna	cdna
Amino	Sequence
Sequence	Amino
Acid	Acid
Protein	Protein
Isolated	Isolated
Encoding	Encoding
Cloned	Cloned
Acids	Expressed
Identity	Identity
Clone	Clone
Expressed	Deduced

3.4 Tagging Abstracts

Each sample of the collapsed Gibbs sampling algorithm consists of a set of assignments of words to topics. These assignments can be used to identify the role words play in documents. In particular this allows to tag each word in the document with the topic to which it was assigned. Our results are given in Fig. 4. The assignments are indicated by the superscripts. Words which do not have a superscript were not included in the vocabulary of the document-term matrix. The shading was determined by averaging over several samples how often the word was assigned to the most prevalent topic of the document. This should be a reasonable

A generalized[66] fundamental[66] theorem[66] of natural[22] selection[22] is derived[118] for populations[22] incorporating[22] both genetic[22] and cultural[22] transmission[22]. The phenotype[106] is determined[180] by an arbitrary[66] number of multiallelic[22] loci[106] with two-factor epistasis[22] and an arbitrary[66] linkage[106] map[106], as well as by cultural[22] transmission[22] from the parents[22]. Generations[22] are discrete[66] but partially[70] overlapping[280], and mating[22] may be nonrandom[22] at either the genotypic[22] or the phenotypic[22] level[175] (or both). I show that cultural[22] transmission[22] has several important implications[22] for the evolution[22] of population[22] fitness[22], most notably[175] that there is a time[66] lag[40] in the response[10] to selection[22] such that the future[234] evolution[22] depends[22] on the past[234] selection[22] history[22] of the population[22].

Fig. 4 A PNAS abstract tagged according to topic assignments. The *shading* indicates how often a word was assigned to the most prevalent topic of the document. Higher frequencies are indicated by *darker shades*

estimate even in the presence of label switching. Again a comparison to Fig. 6 in GS2004 indicates that both taggings strongly resemble each other.

Conclusions

The complete analysis presented in GS2004 was reproduced by collecting the data using web scraping techniques, applying preprocessing steps to determine the document-term matrix and fitting the LDA model using collapsed Gibbs sampling. The fitted model was analyzed in the same way as in GS2004: the topic distributions of the minor categories were determined and the most prevalent topics for each minor category are compared with respect to their weight assigned to the minor categories. In addition time trends of the topics were fitted and words in documents were tagged based on the topic assignments from the LDA model. Further results from this replication study and a detailed description of the code used for this analysis are given in [11], except for the use of a slightly different tokenizer. The tokenizer used in [11] is the default in package **tm** 0.5.1.

Certainly small deviations can be observed between the two results obtained in each of the analyses. However, in general the conclusions drawn as well as the overall assessment are essentially the same. This leads to the conclusion that the study could be successfully reproduced despite the use of completely different tools and a different text database.

Acknowledgements This research was supported by the Austrian Science Fund (FWF) under Elise-Richter grant V170-N18.

Computational Details

For the automated document retrieval from the public PNAS archive (http://www.pnas.org/) we employed Python 2.6.6 with the web scraping framework **Scrapy** 0.10.3, and additional libraries **pycurl** 7.19.0-3+b1 and **BeautifulSoup** 3.1.0.1-2. Texts that were only available as PDF files were converted to plain text with **pdftotext** 0.12.4.

The main programming and data analysis were conducted in R 2.15.3 with packages **tm** 0.5-8.3, **topicmodels** 0.1-9, **lattice** 0.20-13, **xtable** 1.7-1 and **Rmpfr** 0.5-1.

Calculations for model selection and model fitting were delegated to a computer cluster running the Sun Grid Engine at WU (Wirtschaftsuniversität Wien).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Feinerer, I.: **tm**: Text Mining Package (2013). URL http://CRAN.R-project.org/package=tm. R package version 0.5-8.3
3. Feinerer, I., Hornik, K., Meyer, D.: Text mining infrastructure in R. J. Stat. Softw. **25**(5), 1–54 (2008). URL http://www.jstatsoft.org/v25/i05/
4. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. U S A **101**, 5228–5235 (2004)
5. Grün, B., Hornik, K.: **topicmodels**: An R package for fitting topic models. J. Stat. Softw. **40**(13), 1–30 (2011). URL http://www.jstatsoft.org/v40/i13/
6. Hothorn, T., Leisch, F.: Case studies in reproducibility. Brief. Bioinform. **12**(3), 288–300 (2011)
7. Keiding, N.: Reproducible research and the substantive context. Biostatistics **11**(3), 376–378 (2010)
8. Koenker, R., Zeileis, A.: On reproducible econometric research. J. Appl. Econ. **24**, 833–847 (2009)
9. de Leeuw, J.: Reproducible research: the bottom line. Technical Report 2001031101, Department of Statistics Papers, University of California, Los Angeles (2001). URL http://repositories.cdlib.org/uclastat/papers/2001031101/
10. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: Proceedings of the 17th International World Wide Web Conference (WWW 2008), pp. 91–100. Beijing, China (2008)
11. Ponweiser, M.: Latent Dirichlet allocation in R. Diploma thesis, Institute for Statistics and Mathematics, WU (Wirtschaftsuniversität Wien), Austria (2012). URL http://epub.wu.ac.at/id/eprint/3558
12. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2012). URL http://www.R-project.org/. ISBN:3-900051-07-0
13. Steyvers, M., Griffiths, T.: MATLAB **Topic Modeling** Toolbox 1.4 (2011). URL http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

A Dirichlet Mixture Model for Compositions Allowing for Dependence on the Size

Andrea Ongaro and Sonia Migliorati

Abstract

The Dirichlet is the most well known distribution for compositional data, i.e. data representing vectors of proportions. The flexible Dirichlet distribution (FD) generalizes the Dirichlet one allowing to preserve its main mathematical and compositional properties. At the same time, it does not inherit its lack of flexibility in modeling the dependence concepts appropriate for compositional data. The present paper introduces a new model obtained by extending the basis of positive random variables generating the FD by normalization. Specifically, the new basis exhibits a more sophisticated mixture (latent) representation, which leads to a twofold result. On the one side, a more general distribution for compositional data, called EFD, is obtained by normalization. In particular, the EFD allows for a significantly wider differentiation among the clusters defining its mixture representation. On the other side, the generalized basis induces a tractable model for the dependence between composition and size: the conditional distribution of the composition given the size is still an EFD, the size affecting it in a simple fashion through the cluster weights.

Keywords

Basis size • Clusters • Compositional invariance • Dirichlet mixture

A. Ongaro • S. Migliorati (✉)
Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Piazza Ateneo Nuovo 1, Milano, Italy
e-mail: andrea.ongaro@unimib.it; sonia.migliorati@unimib.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_13, Published online: 12 November 2014

1 Introduction

In many disciplines data consist of vectors of proportions, thus being subject to unit-sum constraints. This entails many new issues, among which the search for a suitable model defined on the simplex which allows for appropriate forms of dependence. Some proposals for managing such type of data can be found in [1–5, 7]. Recently a new distribution on the simplex has been proposed (see [6]): the flexible Dirichlet (FD). Such distribution generalizes the well known Dirichlet one overcoming its main drawback, i.e. its extreme independence structure, but preserving many of its mathematical and compositional properties.

The FD shows an important independence property for compositional data modeling: compositional invariance. Given a D-dimensional random vector $\mathbf{Y}$ with positive components (basis) and the corresponding normalized version $\mathbf{X} = \mathbf{Y}/Y^+$ (composition) where $Y^+ = \sum_{i=1}^{D} Y_i$ (size), the basis is compositionally invariant if $\mathbf{X}$ is independent of Y^+. Such property is relevant from an applicative point of view as often in real data, compositions are not affected by size, e.g. rock compositions or chemical composition of goods or of biological tissues. Moreover it is very useful from a mathematical one since properties of the simplex distribution are more easily derived from the basis ones, for example joint moments and explicit expressions for the density.

Nevertheless there are setups where it is of interest to study some forms of dependence between the composition and the size. This is the case, for example, of family household budgets: when focusing on the proportions of total expenditure spent on different commodity groups (i.e. housing, foodstuffs, clothing or luxury goods), it is reasonable to expect that the composition is significantly affected by the size.

In the present paper we propose a generalization of the FD which allows to incorporate such type of dependence. Our aim is to perform a preliminary investigation of the model to understand some key features and evaluate its potential. In particular, after briefly recalling the definition of the FD (Sect. 2), we introduce the new model in Sect. 3. This is defined by extending the mixture structure of the FD basis (Sect. 3.1). Then, we analyze the distribution of the corresponding size and composition focusing on the benefits in terms of added flexibility in cluster modeling (Sect. 3.2). Section 4 is devoted to the study of the effect of the size on the composition within the new model. A final discussion is given in Sect. 5.

2 The Flexible Dirichlet Distribution

The FD distribution derives from the normalization of a basis $\mathbf{Y}$ of positive dependent random variables obtained by starting from the usual basis of independent equally scaled gamma random variables (i.e. the Dirichlet basis) and randomly allocating to the ith component a further independent gamma random variable. Formally:

$$Y_i = W_i + Z_i\ U \qquad i = 1, \ldots, D \tag{1}$$

where the random variables $W_i \sim Ga(\alpha_i, \beta)$ are independent, $U \sim Ga(\tau, \beta)$ is an independent gamma random variable with the same scale parameter as the W_i's. Moreover $\mathbf{Z} = (Z_1, \ldots, Z_D) \sim Mu(1, \mathbf{p})$ is a multinomial random vector independent of U and of the W_i's which is equal to $\mathbf{e}_i$ with probability p_i, where $\mathbf{e}_i$ is a vector whose elements are all equal to zero except for the ith element which is one. Here the vector $\mathbf{p} = (p_1, \ldots, p_D)$ is such that $0 \leq p_i < 1$ and $\sum_{i=1}^{D} p_i = 1$, $\alpha_i > 0$ and $\tau > 0$.

The FD distribution is then defined as the distribution of the composition $\mathbf{X} = \mathbf{Y}/Y^+$ and its density function can be expressed as:

$$f_{FD}(\mathbf{x}; \boldsymbol{\alpha}, \mathbf{p}, \tau) = \frac{\Gamma\left(\alpha^+ + \tau\right)}{\prod_{r=1}^{D} \Gamma(\alpha_r)} \left(\prod_{r=1}^{D} x_r^{\alpha_r - 1} \right) \sum_{i=1}^{D} p_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau)} x_i^{\tau} \tag{2}$$

with $\mathbf{x} \in \mathscr{S}^D = \left\{ \mathbf{x} : \ x_i > 0 \, , i = 1, \ldots, D, \ \sum_{i=1}^{D} x_i = 1 \right\}$ and $\alpha^+ = \sum_{i=1}^{D} \alpha_i$. The FD contains the Dirichlet distribution as the inner point $\tau = 1$ and $p_i = \alpha_i / \alpha^+$, $\forall i = 1, \ldots, D$. Notice that the distribution of $\mathbf{X}$ does not depend on the scale parameter β as a consequence of well known properties of the gamma random variable. Moreover, it can be shown that under the FD model the composition $\mathbf{X}$ and the size Y^+ are independent (compositional invariance).

A key feature of the FD is that its distribution function $FD^D(\mathbf{x}; \boldsymbol{\alpha}, \mathbf{p}, \tau)$ can be written as a finite mixture of Dirichlet distributions $\mathscr{D}^D(\mathbf{x}; \boldsymbol{\alpha} + \tau \mathbf{e}_i)$, i.e.:

$$FD^D(\mathbf{x}; \boldsymbol{\alpha}, \mathbf{p}, \tau) = \sum_{i=1}^{D} p_i \mathscr{D}^D(\mathbf{x}; \boldsymbol{\alpha} + \tau \mathbf{e}_i). \tag{3}$$

Such mixture representation, among other aspects, allows for a variety of different shapes for the density, including multi-modality.

Many relevant properties of the FD can be found in [6], namely the distribution of marginals, conditionals and subcompositions, some fruitful representations and expressions of joint and conditional moments. Moreover the model is closed under components permutation and amalgamation. The latter property, implies that the composition obtained by amalgamating (i.e. summing up) some components is still FD distributed. A further and particularly relevant feature of the FD model is its ability of modeling most types of independence relevant for compositional data.

3 An Extension of the Flexible Dirichlet

3.1 The Basis

The random allocation scheme defining the FD basis (see Sect. 2) can be extended assigning a different gamma random variable U_i to each component of the basis. Thus, let us consider the basis

$$Y_i = W_i + Z_i U_i \tag{4}$$

where the gamma random variables $U_i \sim Ga(\tau_i, \beta)$ are independent and independent of $\mathbf{W} = (W_1, \ldots, W_D)$ and $\mathbf{Z} = (Z_1, \ldots, Z_D)$, which are defined as in (1). Here the vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_D)$ has positive elements.

By conditioning on $\mathbf{Z}$, it is possible to prove that the distribution function of the basis (4) $F_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\alpha}, \mathbf{p}, \boldsymbol{\tau}, \beta)$ admits the following representation:

$$F_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\alpha}, \mathbf{p}, \boldsymbol{\tau}, \beta) = \sum_{i=1}^{D} \left\{ Ga(y_i; \alpha_i + \tau_i, \beta) \prod_{r \neq i} Ga(y_r; \alpha_r, \beta) \right\} p_i \tag{5}$$

i.e. it is a finite mixture of random vectors with independent gamma components.

This allows to derive the corresponding density, which has the form:

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\alpha}, \mathbf{p}, \boldsymbol{\tau}, \beta) = \frac{\beta^{\alpha^+}}{\prod_{r=1}^{D} \Gamma(\alpha_r)} e^{-\beta y^+} \prod_{r=1}^{D} y_r^{\alpha_r - 1} \sum_{i=1}^{D} (\beta y_i)^{\tau_i} \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau_i)} p_i . \tag{6}$$

Furthermore, as the gamma moments have a simple explicit form, i.e. if $W \sim Ga(\alpha, \beta)$ then:

$$E(W^n) = \frac{\alpha^{[n]}}{\beta^n},$$

where $x^{[n]}$ is the factorial $x(x+1)\ldots(x+n-1)$ with $x^{[0]} = 1$, the joint moments of $\mathbf{Y}$ can be explicitly computed:

$$E\left[\prod_{i=1}^{D} Y_i^{n_i} \right] = \beta^{-n^+} \prod_{r=1}^{D} \alpha_r^{[n_r]} \sum_{i=1}^{D} \frac{(\alpha_i + \tau_i)^{[n_i]}}{\alpha_i^{[n_i]}} p_i \tag{7}$$

where n_i's are arbitrary nonnegative integers and $n^+ = \sum_{i=1}^{D} n_i$.

3.2 The Size and the Composition

The mixture representation (5) allows to derive an explicit expression for the corresponding composition and size densities. More specifically, conditionally on $\mathbf{Z} = \mathbf{e}_i$, $\mathbf{X} = \mathbf{Y}/Y^+$ is a $\mathscr{D}^D(\boldsymbol{\alpha}_i)$ where $\boldsymbol{\alpha}_i = \boldsymbol{\alpha} + \tau_i \mathbf{e}_i$ independent of $Y^+ \sim Ga(\alpha^+ + \tau_i, \beta)$. Thus the distribution of $(\mathbf{X}, Y^+)$ can be written as

$$F_{\mathbf{X},Y+}(\mathbf{x}, y^+) = \sum_{i=1}^{D} \mathscr{D}^D(\mathbf{x}; \boldsymbol{\alpha}_i) Ga(y^+; \alpha^+ + \tau_i, \beta) p_i . \tag{8}$$

From (8), marginal and conditional distributions of $\mathbf{X}$ and Y^+ can be easily derived. In particular, the distribution of the size Y^+ is:

$$F_{Y+}(y^+) = \sum_{i=1}^{D} Ga(y^+; \alpha^+ + \tau_i, \beta) p_i .$$

This is a quite general parametric family being a mixture of gamma random variables with arbitrary shape parameters and a common scale parameter β. The latter still represents a scale parameter for the mixture model. Furthermore, by varying α^+, $\boldsymbol{\tau}$ and $\mathbf{p}$, many different shapes can be obtained. For example, multimodality can be accounted for, as for $\alpha > 1$ the random variable $W \sim Ga(\alpha, \beta)$ is unimodal. The moments of Y^+ can be easily computed as mixtures of gamma moments:

$$E[(Y^+)^{n_i})] = \sum_{i=1}^{D} \frac{(\alpha^+ + \tau_i)^{[n_i]}}{\beta^{n_i}} p_i .$$

In particular, mean and variance take the form:

$$E(Y^+) = \frac{\alpha^+ + \bar{\tau}}{\beta}$$

$$Var(Y^+) = \frac{\alpha^+ + \bar{\tau} + s_\tau^2}{\beta^2}$$

where $\bar{\tau} = \sum_{i=1}^{D} \tau_i p_i$ and $s_\tau^2 = \sum_{i=1}^{D} (\tau_i - \bar{\tau})^2 p_i$.

Moreover, (8) implies that the distribution of $\mathbf{X}$, called extended flexible Dirichlet and denoted by $EFD(\boldsymbol{\alpha}, \mathbf{p}, \boldsymbol{\tau})$, has a Dirichlet mixture representation

$$\sum_{i=1}^{D} \mathscr{D}^D(\mathbf{x}; \boldsymbol{\alpha}_i) p_i , \tag{9}$$

with $\boldsymbol{\alpha}_i = \boldsymbol{\alpha} + \tau_i \mathbf{e}_i$, which leads to the density

$$f_{EFD}(\mathbf{x};\boldsymbol{\alpha},\mathbf{p},\boldsymbol{\tau}) = \frac{1}{\prod_{r=1}^{D}\Gamma(\alpha_r)}\left(\prod_{r=1}^{D}x_r^{\alpha_r-1}\right)\sum_{i=1}^{D}\frac{\Gamma(\alpha_i)\Gamma(\alpha^+ + \tau_i)}{\Gamma(\alpha_i+\tau_i)}x_i^{\tau_i}\,p_i\,. \tag{10}$$

The main difference with respect to the FD density (2) is that the EFD model entails different exponents τ_i for the x_i power terms.

Analogously to the FD model, joint moments of the EFD can be straightforwardly computed from the Dirichlet ones thanks to the mixture representation (9):

$$E\left[\prod_{i=1}^{D}X_i^{n_i}\right] = \prod_{i=1}^{D}\alpha_i^{[n_i]}\sum_{r=1}^{D}\frac{(\alpha_r+\tau_r)^{[n_r]}}{\alpha_r^{[n_r]}(\alpha^+ + \tau_r)^{[n^+]}}\,p_r \tag{11}$$

where n^+ and $x^{[n]}$ are defined in Sect. 3.1.

For example, the first moment of the EFD takes the simple form

$$E(X_i) = \alpha_i\sum_{r=1}^{D}\frac{p_r}{\alpha^+ + \tau_r} + \tau_i\frac{p_i}{\alpha^+ + \tau_i} \qquad i = 1,\ldots,D.$$

Another relevant property of the FD distribution carries over to the EFD, namely closure under permutation: if $\mathbf{X} \sim EFD(\boldsymbol{\alpha},\mathbf{p},\boldsymbol{\tau})$ then any permutation of (the components of) $\mathbf{X}$ is EFD with parameters obtained by applying the same permutation to $\boldsymbol{\alpha}$, $\mathbf{p}$ and $\boldsymbol{\tau}$. This implies a symmetry of the distribution with respect to the components which allows the statistical analyses to be unaffected by the particular order chosen to form the composition. Notice that many widespread models for compositional data do not share such property, see for example [1,3,7].

A first important advantage of the EFD over the FD is a larger flexibility in modeling the implied cluster structure. The FD model can be viewed as composed by D (Dirichlet distributed) clusters, each with weight p_i. The parameter τ regulates the extent of the common differences among all clusters: the larger τ the larger each ith component of the ith cluster with respect to the ith component of the other clusters ($i = 1,\ldots,D$). Thus, only one parameter determines how far apart the D clusters are from each other in a completely symmetric fashion.

By introducing the τ_i's a much larger variety of clusters is clearly reachable, as the cluster structure can be modeled separately to some extent. Specifically, let us consider the effect of the τ_i's on the cluster means. In the EDF model the mean vector $\boldsymbol{\mu}_i$ of the generic ith cluster is $(\boldsymbol{\alpha}+\tau_i\mathbf{e}_i)/(\alpha^+ + \tau_i)$. It follows that each τ_i only affects the corresponding ith cluster mean. Moreover, by increasing τ_i such mean vector varies, monotonically and continuously componentwise, from $\boldsymbol{\alpha}/(\alpha^+)$ to the ith vertex $\mathbf{e}_i$. In other words, τ_i dictates how far the ith cluster mean is from the common barycenter $\boldsymbol{\alpha}/(\alpha^+)$ in the direction of the ith vertex of the simplex.

To illustrate, compare the FD model with equal α_i's and p_i's shown in Fig. 1 (corresponding to an EFD with equal τ_i's) with the EFD model with the same α_i's and p_i's but different τ_i's shown in Fig. 2.

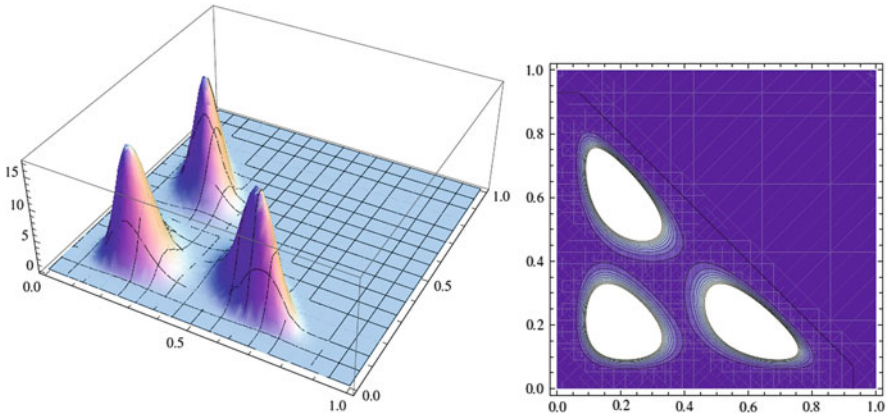

Fig. 1 FD density (*left*) and contour plots (*right*) with $\boldsymbol{\alpha} = (10, 10, 10)$, $\mathbf{p} = (1/3, 1/3, 1/3)$ and $\boldsymbol{\tau} = (20, 20, 20)$

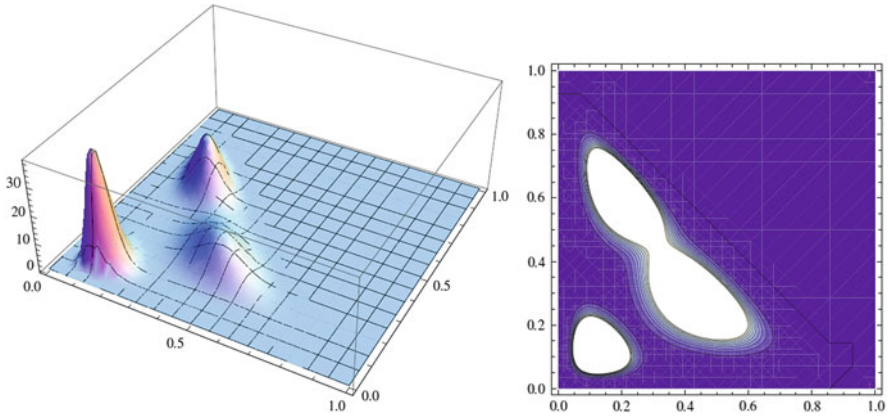

Fig. 2 EFD density (*left*) and contour plots (*right*) with $\boldsymbol{\alpha} = (10, 10, 10)$, $\mathbf{p} = (1/3, 1/3, 1/3)$ and $\boldsymbol{\tau} = (5, 20, 50)$

In particular, different τ_i's allow to break the symmetry of the clusters so that the distance between the vertexes and the cluster barycenters can be quite different.

4 Dependence Between Composition and Size

Within the $EFD(\boldsymbol{\alpha}, \mathbf{p}, \boldsymbol{\tau})$ model the dependence between composition and size can be investigated computing the distribution of $\mathbf{X}|Y^+ = y^+$. Remarkably, from (8) one can see that such conditional distribution is still an EFD. Furthermore it displays a simple dependence on the size. Specifically, we have that $\mathbf{X}|Y^+ = y^+ \sim EFD(\boldsymbol{\alpha}, \mathbf{p}'(y^+), \boldsymbol{\tau})$ where

$$p_i'(y^+) \propto \frac{(y^+)^{\tau_i} p_i}{\Gamma(\alpha^+ + \tau_i)}, \quad i = 1, \ldots, D. \tag{12}$$

It follows that the cluster structure is unaltered in the sense that the cluster barycenters remain the same. The size influences only the probabilities $p_i'(y^+)$'s defining the weights of the clusters and such influence is regulated by the τ_i's.

The form (12) of the updated weights allows to show that equality of the τ_i's is the only case of compositional invariance. More precisely, if $\tau_i = \tau \ \forall i$, it is immediate to see that the $p_i'(y^+)$'s are independent of y^+ (and coincide with the p_i's). On the other hand, if the basis is compositional invariant, then $p_i'(y^+)$ does not depend on y^+ and, therefore, the ratio $p_i'(y^+)/p_r'(y^+)$ does not either $\forall i \neq r$. Since such ratio is proportional to $(y^+)^{\tau_i - \tau_r}$, then $\tau_i = \tau_r \ \forall i \neq r$.

The relevant hypothesis of compositional invariance can be therefore easily expressed and tested within the EFD model and it is equivalent to testing appropriateness of the FD model (2).

Let us now investigate what type of dependence is implied by the EFD when the τ_i's are not all equal. In particular, consider the relative effect of y^+ on the weights of two generic clusters as measured by the probability ratio $p_i'(y^+)/p_r'(y^+)$. The size has no influence, i.e. $p_i'(y^+)/p_r'(y^+) = p_i/p_r$, if and only if $\tau_i = \tau_r$. Otherwise such ratio is a monotone function of y^+ with range $(0, \infty)$, being increasing if $\tau_i > \tau_r$ and decreasing in the opposite case. Therefore the τ_i's dictate how y^+ affects the structure of weights in the following simple form: the larger y^+ the higher the weights associated with high values of the τ_i's.

The effect of the size on the updating mechanism of the weights can be better illustrated graphically. So let us consider a simple situation where the τ_i's are increasing: $\boldsymbol{\tau} = (3, 4, 5)$. Then, Fig. 3 shows how the total weight moves from the cluster on the right bottom (corresponding to the lowest τ_i) to the cluster on the left bottom (characterized by the highest τ_i) as y^+ increases.

To get a concrete interpretation of the size effect entailed by the model, consider the following example relative to household budgets. Suppose that commodities are grouped into three types from the most essential to the more luxury ones. Then, we may figure there are three expenditure patterns (clusters), the ith pattern being characterized by families with a higher proportion of expenditure on the ith type than the other families. The EDF model reasonably assumes that an increase in the total absolute expenditure Y^+ implies a shift of the importance of the patterns towards the more luxurious ones.

The dependence between composition and size can be further investigated through the conditional mean (regression on y^+) which takes the form:

$$E\left[X_i | Y^+ = y^+\right] = \frac{1}{c}\left(\alpha_i \sum_{r=1}^{D} (y^+)^{\tau_r} \ d_r + \tau_i (y^+)^{\tau_i} \ d_i\right)$$

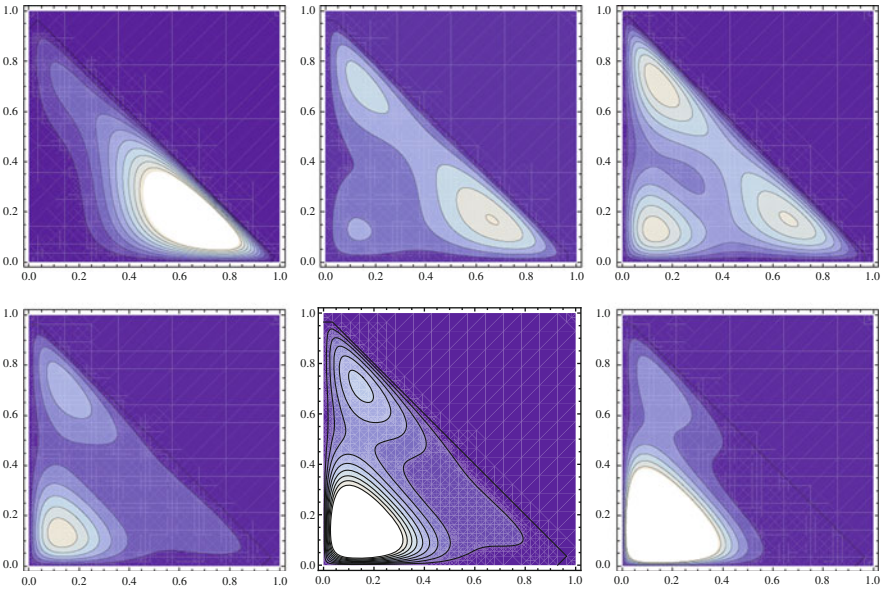

Fig. 3 EFD contour plots with $\boldsymbol{\alpha} = (2, 2, 2)$, $\mathbf{p} = (1/3, 1/3, 1/3)$, $\boldsymbol{\tau} = (3, 4, 5)$ for increasing values of the size: $y^+ = 1, 5, 8, 15, 30, 70$

where

$$d_i = \frac{p_i}{\Gamma(\alpha^+ + \tau_i)(\alpha^+ + \tau_i)} \quad i = 1, \ldots, D$$

and

$$c = \sum_{r=1}^{D} \frac{(y^+)^{\tau_r}\ p_r}{\Gamma(\alpha^+ + \tau_r)}.$$

Figure 4 shows the behavior of such regression functions for the same parameter configuration of Fig. 3.

The regression functions have a pattern similar to the weights $p_i'(y^+)$: as y^+ increases the relevance of the various components changes, shifting from the first to the second and then to the third. In particular, the first component (corresponding to the smallest τ_i) and the last one (corresponding to the largest τ_i) are monotone while the second is bell-shaped. To better understand the regression function behavior, we report an example with more components in Fig. 5.

The "shifting" effect is confirmed as well as a common "bell-shaped" form for all the middle components.

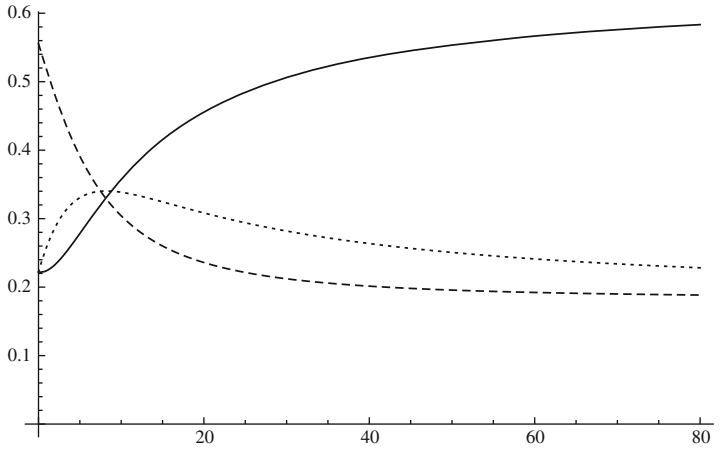

Fig. 4 Conditional mean of X_1 (*dashed line*), X_2 (*dotted line*) and X_3 (*solid line*) with $\boldsymbol{\alpha} = (2, 2, 2)$, $\mathbf{p} = (1/3, 1/3, 1/3)$, $\boldsymbol{\tau} = (3, 4, 5)$

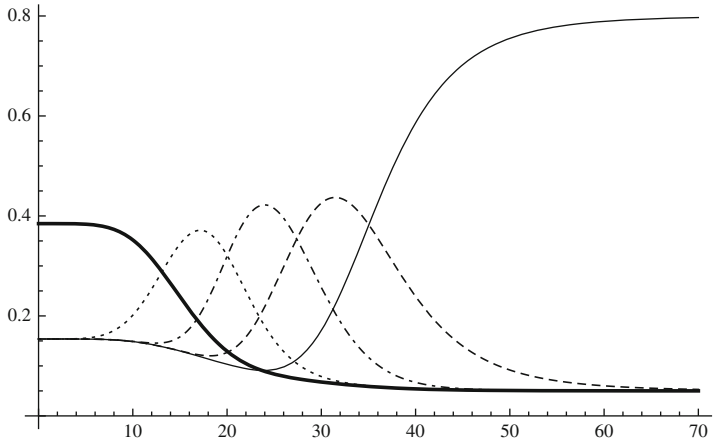

Fig. 5 Conditional mean of X_1 (*solid thick line*), X_2 (*dotted line*), X_3 (*dotdashed line*), X_4 (*dashed line*) and X_5 (*solid thin line*) with $\boldsymbol{\alpha} = (2, 2, 2, 2, 2)$, $\mathbf{p} = (1/5, 1/5, 1/5, 1/5, 1/5)$, $\boldsymbol{\tau} = (3, 8, 15, 22, 30)$

5 Discussion

The EFD model loses some convenient properties of the FD such as closure under amalgamation and various simple representations. Nevertheless, our analysis shows that it remains mathematically rather tractable and sufficiently easy to interpret.

Furthermore, the EFD exhibits relevant advantages over the FD in at least two directions: modeling cluster structure and dependence on the size. In particular, the EFD cluster structure is shown to substantially extend the FD one, by removing

symmetry constraints entailed by the latter. Moreover, the EFD, unlike the FD, enables to model both independence and dependence on the size. The former case corresponds to the FD. It follows that the relevant hypothesis of independence (compositional invariance) can be conveniently formulated and tested within the EFD. The dependence form implied by the EFD as well as the meaning of the influential parameters are quite easy to grasp and to deal with. This is partly due the useful property that, under the EFD model, the conditional distribution of the composition given the size still has an EFD distribution.

Further investigation is needed to fully understand the EFD behavior in terms of theoretical properties such as distributions of marginals and conditionals and dependence structure.

Even more importantly, inferential aspects should be tackled. In particular, direct maximization of the likelihood is not feasible due to the presence of several local maxima. Yet, the finite mixture structure of the EFD allows the estimation to be fulfilled via E–M algorithm, where the usual M-step is implemented by means of a Newton–Raphson scheme. However, a preliminary analysis shows that the choice of the starting values for the E–M is crucial, requiring at least a multiple points starting strategy. Indeed, a substantial improvement of the initial values choice can be expected by devising an ad hoc initial clustering procedure which exploits the particular features of the context, i.e. the compositional nature of data and the peculiar mixture structure implied by the model. The first aspect could be dealt with by adopting suitable transformations of data, such as the symmetric representation of the D-dimensional simplex as a regular simplicial polytope in $\mathbb{R}^{D-1}$. The second one could be used to the correct labeling of the initial groups by associating them to the appropriate corresponding mixture components.

Finally, an analysis of the EFD as prior for categorical data is worth exploring, as it is easily seen to be conjugate with respect to multinomial sampling.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. The Blackburn Press, London (2003)
2. Barndorff-Nielsen, O.E., Jorgensen, B.: Some parametric models on the simplex. J. Multivariate Anal. **39**, 106–116 (1991)
3. Connor, J.R., Mosimann, J.E.: Concepts of independence for proportions with a generalization of the Dirichlet distribution. J. Am. Stat. Assoc. **64**, 194–206 (1969)
4. Favaro, S., Hadjicharalambous, G., Prunster, I.: On a class of distributions on the simplex. J. Stat. Plan. Infer. **141**, 2987–3004 (2011)
5. Gupta, R.D., Richards, D.St.P.: Multivariate Liouville distributions. J. Multivariate Anal. **23**, 233–256 (1987)
6. Ongaro, A., Migliorati, S.: A generalization of the Dirichlet distribution. J. Multivariate Anal. **114**, 412–426 (2013)
7. Rayens, W.S., Srinivasan, C.: Dependence properties of generalized Liouville distributions on the simplex. J. Am. Stat. Assoc. **89**, 1465–1470 (1994)

A Latent Variable Approach to Modelling Multivariate Geostatistical Skew-Normal Data

Luca Bagnato and Marco Minozzo

Abstract

In this paper we propose a spatial latent factor model to deal with multivariate geostatistical skew-normal data. In this model we assume that the unobserved latent structure, responsible for the correlation among different variables as well as for the spatial autocorrelation among different sites is Gaussian, and that the observed variables are skew-normal. For this model we provide some of its properties like its spatial autocorrelation structure and its finite dimensional marginal distributions. Estimation of the unknown parameters of the model is carried out by employing a Monte Carlo Expectation Maximization algorithm, whereas prediction at unobserved sites is performed by using closed form formulas and Markov chain Monte Carlo algorithms. Simulation studies have been performed to evaluate the soundness of the proposed procedures.

Keywords

Closed skew-normal distribution • Factor model • Geostatistics • Monte Carlo EM • Spatial process

L. Bagnato (✉)
Università Cattolica del Sacro Cuore, Milano, Italy
e-mail: luca.bagnato@unicatt.it

M. Minozzo (✉)
Dipartimento di Scienze Economiche, Università degli Studi di Verona, Via dell'Artigliere 19, 37129 Verona, Italy
e-mail: marco.minozzo@univr.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_14, Published online: 28 October 2014

1 Introduction

Although a large variety of spatial data sets (on radioactive contamination, rainfalls, winds, etc.) contain measurements with a considerable amount of skewness, its modelling still remains an issue. For instance, with regard to radiological monitoring, in [5], disregarding any physically-based modelling approach, it is argued on the necessity of developing mapping algorithms for emergency detection taking into consideration the skewness in the data. A boost to these developments came from the Spatial Interpolation Comparison (SIC) 2004 (see [11]) in which, whereas the routine scenario could easily be modelled using a Gaussian random field, the emergency scenario, which mimics an accidental release of radioactivity, needs to be modelled taking properly into account that, due to the presence of extreme measurements, the data are positively skewed. Just to cite a few works, to deal with skewed measurements coming from radioactive monitoring, [18] and [15] propose copula-based geostatistical approaches, whereas [9] argues that the structuring of extreme values can be faced in a coherent manner by using the class of Hermitian isofactorial models. Moreover, [4] proposes a Gaussian anamorphosis transformation to deal with skewed data coming from contaminated facilities, and [19] argues in favor of a Bayesian approach pointing out that both the Gaussian copula and the non-Gaussian χ^2-copula models are inappropriate to model strongly skewed radioactivity measurements. Other works dealing with skewed radiological measurements are [27], which is concerned with the estimation of the variogram and the development of optimal sampling plans, [7], which proposes a dynamic spatial Bayesian model for non-Gaussian measurements from radioactivity deposition, as well as the works in [22, 28] and [32]. On the other hand, a general approach developed to cope with some types of univariate non-Gaussian spatial data (including skew data) has been proposed in [8] by defining a family of transformed Gaussian random fields that provides an alternative to trans-Gaussian kriging.

Whereas in the univariate case, that is, in presence of just one regionalized variable, spatial modelling and prediction have been extensively studied for different types of non-Gaussian data, in particular skew data, in a multivariate non-Gaussian context only a limited number of works have been published. Among these, [26] and [25] extend to multivariate geostatistical non-Gaussian data the modelling approach of [10], whereas [6] proposes a hierarchical Bayesian approach to model Gaussian, count, and ordinal variables, by designing a Gibbs sampler with Metropolis-Hastings steps. Other works dealing with multivariate spatial data are those in [33], which explores the use of the Bayesian Maximum Entropy approach in presence of both continuous and categorical regionalized variables, and in [31], which uses Markov chain Monte Carlo methods for the Bayesian modelling of multivariate counts.

In this paper, to model skewness in a multivariate (that is, in presence of more than one regionalized variable) geostatistical context, we propose an alternative approach based on the use of the skew-normal distribution. Our modelling approach, which extends some of the ideas in [24] (see also [35]), is based on the skew-normal distribution [2, 3] and on a latent Gaussian factor structure. Just to give some examples, this approach might prove useful in the modelling of the radiological data in [16] or the data related to the Fukushima disaster (data are available from TEPCO at http://www.tepco.co.jp) where more than one radiological measurement has been collected for each sampling site. Apart from providing a much greater flexibility with respect to the traditional Gaussian random fields, it is possible to show that our model has all its finite-dimensional marginal distributions belonging to the family of the closed skew-normal distribution [13, 14]. It must be mentioned that the modelling construction proposed here is substantially different from some of the most popular constructions based on the skew-normal distribution that have recently appeared in the literature to model univariate skewed spatial data, like those, for instance, of [1, 20] and [17] (for a critical discussion on these constructions see [24]).

The paper is organized as follows. The model and its properties are presented in Sect. 2 and in Sect. 3, respectively. In Sect. 4 we present the estimation and prediction procedures and some simulation results, and in section "Conclusions" we make some final comments. More technical results are presented in the Appendix.

2 A Multivariate Closed Skew-Normal Geostatistical Model

In the following we define a model for geostatistical multivariate skewed data exploiting the ideas in [24] and in [25], by building the model on an unobserved latent Gaussian spatial factor structure. Let $y_i(\mathbf{x}_k)$, $i = 1, \ldots, m$, $k = 1, \ldots, K$, be a set of geo-referenced data measurements relative to m regionalized variables, gathered at K spatial locations $\mathbf{x}_k$. Each of these m measured variables can be viewed as a partial realization of a particular stochastic process $Y_i(\mathbf{x})$, $i = 1, \ldots, m$, $\mathbf{x} \in \mathbb{R}^2$. We assume that these stochastic processes are given by

$$Y_i(\mathbf{x}) = \beta_i + Z_i(\mathbf{x}) + \omega_i S_i(\mathbf{x}), \quad i = 1, \ldots, m, \tag{1}$$

where β_i and ω_i are unknown constants, representing, respectively, an intercept and a scale parameter, and $Z_i(\mathbf{x})$ and $S_i(\mathbf{x})$ are latent processes. In particular, for every $i = 1, \ldots, m$, $Z_i(\mathbf{x})$ is a mean zero stationary Gaussian process, whereas for every $i = 1, \ldots, m$, and for each $\mathbf{x} \in \mathbb{R}^2$, $S_i(\mathbf{x})$ is an independent random variable distributed as a skew-normal [2], that is, $S_i(\mathbf{x}) \sim SN(0, 1, \alpha_i)$, which means that, for every $\mathbf{x} \in \mathbb{R}^2$, the density of $S_i(\mathbf{x})$ is given by $f_{S_i}(s) = 2\phi_1(s; 1)\Phi(\alpha_i s)$, for $-\infty < s < \infty$, where $\alpha_i \in \mathbb{R}$, $\phi_1(\cdot; 1)$ is the scalar normal density function with zero mean and unit variance, and $\Phi(\cdot)$ is the scalar N(0, 1) distribution function.

Let us note that, for each $i = 1, \ldots, m$, and for every $\mathbf{x} \in \mathbb{R}^2$, conditionally on $Z_i(\mathbf{x})$, the random variable $Y_i(\mathbf{x})$ has a skew-normal distribution, that is,

$$Y_i(\mathbf{x}) \,|\, Z_i(\mathbf{x}) \sim SN\left(\beta_i + Z_i(\mathbf{x}), \omega_i^2, \alpha_i\right), \tag{2}$$

which means that we can write its density as

$$f(y_i(\mathbf{x})|z_i(\mathbf{x})) = 2\,\phi_1\left(y_i(\mathbf{x}) - \beta_i - z_i(\mathbf{x}); \omega_i^2\right) \Phi\left(\frac{\alpha_i}{\omega_i}\left(y_i(\mathbf{x}) - \beta_i - z_i(\mathbf{x})\right)\right),$$

where $\phi_1(\,\cdot\,; \sigma^2)$ is the scalar normal density function with zero mean and positive variance σ^2. Moreover, for each $i = 1, \ldots, m$, and for every $\mathbf{x} \in \mathbb{R}^2$, the (scalar) random variable $Y_i(\mathbf{x})$ has a (marginal) skew-normal distribution, that is,

$$Y_i(\mathbf{x}) \sim SN\left(\beta_i, \varsigma_i^2 + \omega_i^2, \alpha_i \omega_i \Big/ \sqrt{\varsigma_i^2\left(1 + \alpha_i^2\right) + \omega_i^2}\right), \tag{3}$$

where $\varsigma_i^2 = \mathrm{Var}\left[Z_i(\mathbf{x})\right]$.

A similar result holds also for the other marginal distributions of the process. Indeed, with some algebra it is possible to show that all finite dimensional marginal distributions of the (weakly and strongly stationary) multivariate spatial process $(Y_1(\mathbf{x}), \ldots, Y_m(\mathbf{x}))^T$, for $\mathbf{x} \in \mathbb{R}^2$, are closed skew-normal (CSN). This implies, for instance, that, for each $i = 1, \ldots, m$, the univariate spatial process $Y_i(\mathbf{x})$, for $\mathbf{x} \in \mathbb{R}^2$, has all its finite-dimensional marginal distributions belonging to the CSN family (see the Appendix), and that, for any fixed spatial location $\mathbf{x} \in \mathbb{R}^2$, the random vector $(Y_1(\mathbf{x}), \ldots, Y_m(\mathbf{x}))^T$ has a multivariate CSN distribution [13, 14]. In principle, these results make the approach very appealing since they allow, due to the stationarity of the processes, to empirically check some of the distributional properties of the model. For instance, for a given set of observations, the empirical distribution of $y_i(\mathbf{x}_k)$, $k = 1, \ldots, K$, for any given $i = 1, \ldots, m$, can be compared with the marginal skew-normal distribution in (3).

For the latent part of the model, that is, for the stationary Gaussian processes $Z_i(\mathbf{x})$, $i = 1, \ldots, m$, we assume that

$$Z_i(\mathbf{x}) = \sum_{p=1}^{P} a_{ip} F_p(\mathbf{x}), \tag{4}$$

where a_{ip} are $m \times P$ real coefficients, and $F_p(\mathbf{x})$, $p = 1, \ldots, P$, are $P \leq m$ non-observable spatial processes (*common factors*) responsible for the cross-correlations in the model. The processes $F_p(\mathbf{x})$, $p = 1, \ldots, P$, are assumed zero mean, stationary, and Gaussian with covariance function

$$\mathrm{Cov}\left[F_p(\mathbf{x}), F_q(\mathbf{x}+\mathbf{h})\right] = \begin{cases} \rho(\mathbf{h}), & p = q, \\ 0, & p \neq q, \end{cases}$$

where $\mathbf{h} \in \mathbb{R}^2$ and $\rho(\mathbf{h})$ is a real spatial autocorrelation function common to all factors with $\rho(\mathbf{0}) = 1$ and $\rho(\mathbf{h}) \to 0$, as $\|\mathbf{h}\| \to \infty$. Similarly to the classical linear factor model, this latent linear structure is responsible for a specific correlation structure among the processes $Z_i(\mathbf{x})$. In particular, for each $i = 1, \ldots, m$, the covariance functions are given by $\mathrm{Cov}[Z_i(\mathbf{x}), Z_i(\mathbf{x}+\mathbf{h})] = \sum_{p=1}^{P} a_{ip}^2 \rho(\mathbf{h})$, whereas the cross-covariance functions are given by $\mathrm{Cov}\left[Z_i(\mathbf{x}), Z_j(\mathbf{x}+\mathbf{h})\right] = \sum_{p=1}^{P} a_{ip} a_{jp} \rho(\mathbf{h})$. Taking $\mathbf{h} = \mathbf{0}$, we find that $\mathrm{Var}[Z_i(\mathbf{x})] = \sum_{p=1}^{P} a_{ip}^2$ and $\mathrm{Cov}\left[Z_i(\mathbf{x}), Z_j(\mathbf{x})\right] = \sum_{p=1}^{P} a_{ip} a_{jp}$.

3 Variograms and Cross-Variograms

Let us consider here the correlation structure of the observable processes, induced by the latent factor model. For the observable stochastic processes $Y_i(\mathbf{x}), i = 1, \ldots, m$, we can show that

$$\mathrm{E}[Y_i(\mathbf{x})] = \beta_i + \omega_i \delta_i \left(\frac{2}{\pi}\right)^{\frac{1}{2}}, \qquad \mathrm{Var}[Y_i(\mathbf{x})] = \varsigma_i^2 + \omega_i^2 \left[1 - \frac{2}{\pi}\delta_i^2\right],$$

where $\delta_i = \alpha_i / \sqrt{1+\alpha_i^2}$, and, for $\mathbf{h} \neq \mathbf{0}$,

$$C_{ii}(\mathbf{h}) = \mathrm{Cov}[Y_i(\mathbf{x}), Y_i(\mathbf{x}+\mathbf{h})] = \varsigma_i^2 \rho(\mathbf{h}). \tag{5}$$

Note that if $\rho(\mathbf{h}) = \rho(-\mathbf{h})$ we have that $C_{ii}(\mathbf{h}) = C_{ii}(-\mathbf{h})$. Furthermore, $C_{ii}(\infty) = 0$ and $C_{ii}(\mathbf{0}) \neq C_{ii}(\mathbf{0}^+) = \varsigma_i^2$, that is, the covariance function $C_{ii}(\mathbf{h})$ is discontinuous at the origin.

On the other hand, for $\mathbf{h} \neq \mathbf{0}$, the variogram of the observable $Y_i(\mathbf{x})$ takes the form

$$\gamma_{ii}(\mathbf{h}) = \frac{1}{2}\mathrm{Var}[Y_i(\mathbf{x}+\mathbf{h}) - Y_i(\mathbf{x})] = \omega_i^2 \left[1 - \frac{2}{\pi}\delta_i^2\right] + \varsigma_i^2 [1 - \rho(\mathbf{h})], \tag{6}$$

which is, similarly to the covariance function, discontinuous in zero. In fact, we have that $\gamma_{ii}(\mathbf{0}) = 0$ and $\gamma_{ii}(\mathbf{0}^+) = \omega_i^2[1 - (2/\pi)\delta_i^2]$. Note that $\gamma_{ii}(\infty) = C_{ii}(\mathbf{0})$. To visually asses Formula (6), Fig. 1 shows the form taken by the variogram $\gamma_{ii}(\mathbf{h})$

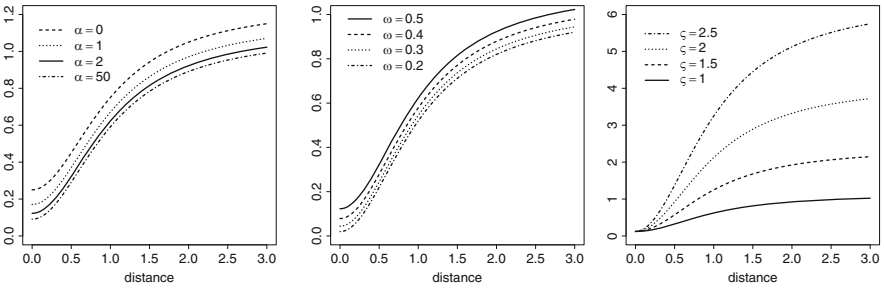

Fig. 1 The graphs show the shape of the theoretical variogram $\gamma_{ii}(\mathbf{h})$ given in Formula (6), for a Cauchy autocorrelation function with both parameters equal to 1, and for different values of the other parameters: (*left*) $\omega = 0.5$, $\varsigma = 1$; (*middle*) $\alpha = 2$, $\varsigma = 1$; (*right*) $\alpha = 2$, $\omega = 0.5$. The *solid line* in the three graphs corresponds to the same set of parameter values. The *line* in the first graph corresponding to $\alpha = 0$ gives the variogram in the case of a Gaussian process

for different values of the parameters, in the case of a Cauchy spatial autocorrelation function $\rho(\mathbf{h}) = \left[1 + (\|\mathbf{h}\|/\gamma)^2\right]^{-\eta}$, with $\gamma = 1$ and $\eta = 1$. As we can see, the nugget of the variogram decreases for decreasing values of ω and for values of the skewness parameter α departing from zero.

For any two stochastic processes $Y_i(\mathbf{x})$ and $Y_j(\mathbf{x})$, with $i \neq j$, it is easy to show that

$$C_{ij}(\mathbf{h}) = \mathrm{Cov}\left[Y_i(\mathbf{x}), Y_j(\mathbf{x}+\mathbf{h})\right] = \mathrm{Cov}\left[Z_i(\mathbf{x}), Z_j(\mathbf{x}+\mathbf{h})\right] = \varsigma_{ij}\rho(\mathbf{h}), \tag{7}$$

where $\varsigma_{ij} = \sum_{p=1}^{P} a_{ip}a_{jp} = \mathrm{Cov}\left[Z_i(\mathbf{x}), Z_j(\mathbf{x})\right]$. Note that $C_{ij}(\mathbf{h}) = C_{ji}(\mathbf{h})$ and that if $\rho(\mathbf{h}) = \rho(-\mathbf{h})$, then $C_{ij}(\mathbf{h}) = C_{ij}(-\mathbf{h})$.

For the cross-variogram between $Y_i(\mathbf{x})$ and $Y_j(\mathbf{x})$, with $i \neq j$, we obtain

$$\gamma_{ij}(\mathbf{h}) = \frac{1}{2}\mathrm{Cov}\left[Y_i(\mathbf{x}+\mathbf{h}) - Y_i(\mathbf{x}), Y_j(\mathbf{x}+\mathbf{h}) - Y_j(\mathbf{x})\right] = \varsigma_{ij}\left[1 - \rho(\mathbf{h})\right]. \tag{8}$$

4 Estimation and Prediction

Assuming to know the number P of common factors and the spatial autocorrelation function $\rho(\mathbf{h})$, the model depends on the parameter vector $\boldsymbol{\vartheta}^* = (\boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\omega}, \boldsymbol{\alpha})$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^T$, $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m)^T$ with $\mathbf{a}_i = (\mathrm{a}_{i1}, \ldots, \mathrm{a}_{iP})^T$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)^T$, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T$. Note that, similarly to the classical factor model, our model is not identifiable. Indeed, there are two groups of orthogonal transformations of the matrix $\mathbf{A}$, given by permutation matrices and by some special reflection matrices, that leave the model unchanged [30]. However, this is the only indeterminacy in the model and can easily be faced.

In the following, we will further assume to know the parameters $\boldsymbol{\omega}$ and $\boldsymbol{\alpha}$. In this case, by resorting to Markov chain Monte Carlo (MCMC), and in particular to the Metropolis-Hasting algorithm, a likelihood based estimation procedure for the parameter $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \mathbf{A})$ can be developed by exploiting the Monte Carlo Expectation Maximization (MCEM) algorithm. Let $\mathbf{F} = (\mathbf{F}_1, \ldots, \mathbf{F}_P)^T$, where $\mathbf{F}_p = \big(F_p(\mathbf{x}_1), \ldots, F_p(\mathbf{x}_K)\big)^T$, $p = 1, \ldots, P$, and let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)^T$, where $\mathbf{y}_i = (y_i(\mathbf{x}_1), \ldots, y_i(\mathbf{x}_K))^T$, $i = 1, \ldots, m$. Whereas the marginal log-likelihood $l(\boldsymbol{\vartheta}) = \ln f(\mathbf{y}; \boldsymbol{\vartheta})$ is not available due to the presence of multidimensional integrals in the derivation of the marginal density $f(\mathbf{y}; \boldsymbol{\vartheta})$, the complete log-likelihood based on the joint distribution $f(\mathbf{y}, \mathbf{F}; \boldsymbol{\vartheta})$ is easily given by

$$
\begin{aligned}
l_c(\boldsymbol{\vartheta}) &= \ln f(\mathbf{y}, \mathbf{F}; \boldsymbol{\vartheta}) = \ln\big(f(\mathbf{y}|\mathbf{F}; \boldsymbol{\vartheta}) \cdot f(\mathbf{F})\big) \\
&= \ln\left\{\left(\prod_{i=1}^{m}\prod_{k=1}^{K} f(y_{ik}; Z_{ik}, \beta_i)\right) \cdot f(\mathbf{F})\right\} \\
&= \ln\left\{\left(\prod_{i=1}^{m}\prod_{k=1}^{K} 2\,\phi_1\big(y_{ik} - \beta_i - Z_{ik}; \omega_i^2\big)\,\Phi\left(\frac{\alpha_i}{\omega_i}(y_{ik} - \beta_i - Z_{ik})\right)\right) \cdot \left(\prod_{p=1}^{P} f(\mathbf{F}_p)\right)\right\},
\end{aligned}
\tag{9}
$$

where $y_{ik} = y_i(\mathbf{x}_k)$ and $Z_{ik} = Z_i(\mathbf{x}_k)$. In this situation, the marginal log-likelihood $l(\boldsymbol{\vartheta}) = \ln f(\mathbf{y}; \boldsymbol{\vartheta})$ can be maximized by resorting to the Monte Carlo Expectation Maximization (MCEM) algorithm (see, for instance, [23] and [12]).

At the sth iteration, the MCEM algorithm involves three steps: S-step, E-step and M-step. In the first step (S-step), R_s samples $\mathbf{F}^{(r)}$, $r = 1, \ldots, R_s$, are drawn from the (filtered) conditional distribution $f(\mathbf{F}|\mathbf{y}; \boldsymbol{\vartheta}_{s-1})$, where $\boldsymbol{\vartheta}_{s-1}$ is the guess of the parameter $\boldsymbol{\vartheta}$ after the $(s-1)$th iteration. These samples can be collected by using some Markov chain Monte Carlo (MCMC) procedure based on the Metropolis-Hustings algorithm. In the second step (E-step) the following approximation of the conditional expectation of the complete log-likelihood is computed

$$
Q_s(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}_{s-1}) = \hat{\mathrm{E}}\left[\ln f(\mathbf{y}, \mathbf{F}; \boldsymbol{\vartheta})|\mathbf{y}\right] = \frac{1}{R_s}\sum_{r=1}^{R_s} \ln f\big(\mathbf{y}, \mathbf{F}^{(r)}; \boldsymbol{\vartheta}\big).
$$

The last step (M-step) supplies as the new guess $\boldsymbol{\vartheta}_s$ the value of $\boldsymbol{\vartheta}$ which maximizes $Q_s(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}_{s-1})$.

Although convergence results for this algorithm are not available, it is nevertheless possible to show that the "average" complete likelihood which is maximized in the M-step of the MCEM algorithm is concave and admits a unique local (and global) maximum. This result allows to safely implement standard numerical maximization techniques.

Assuming as known all parameters of the model, prediction of the observable processes Y_i (**x**) at an unobserved spatial location (or at an unobserved set of spatial locations) can be carried out either by exploiting some of the properties of the CSN distribution, or by implementing some MCMC algorithm. On the other hand, for the prediction of the unobserved common factors F_p (**x**), we need to resort to MCMC algorithms. In the case in which we are interested in predicting a common factor on a large set of spatial locations (maybe on a grid), instead of carring out an MCMC run at each spatial location, we can carry out an MCMC run only at the sampling points (that is, only at those points for which we gathered observations), and then exploit a linear property similar to Kriging, and also similar to that found by [34] in a univariate framework, to obtain predictions at all other spatial locations.

To asses the goodness of the MCEM estimation procedure we performed some simulation studies. To give some examples, in Fig. 2 we show the results of some simulation analyses. For these analyses we considered $m = 2$ and $P = 1$, that is, two observable variables and one latent common factor F (**x**). In the first two simulation experiments we considered a powered exponential (stable) spatial autocorrelation function $\rho(\mathbf{h}) = \exp\left[-(\gamma \|\mathbf{h}\|)^{\eta} \right]$, with $\gamma = 10^{-5}$ and $\eta = 1.5$, whereas in the last two experiments we considered a Cauchy autocorrelation function with $\gamma = 7{,}000$ and $\eta = 1$. For any given set of parameter values $\boldsymbol{\vartheta}^*$ and a given spatial autocorrelation function $\rho(\mathbf{h})$, we simulated 50 realizations from the model over $K = 25$ equally spaced fixed sampling points located on the nodes of a grid. For each simulated realization, we run the MCEM estimation algorithm, assuming as unknown only the parameters a_{11}, a_{21}, β_1 and β_2. Each time, we considered 800 iterations of the MCEM algorithm, and at each step of the algorithm we considered 800 MCMC samples (of which 400 burn-in). As shown in Fig. 2, despite some possible distortion (which could be due to the modest sample size), the sampling distributions look quite reasonable. However, though our simulation experiments gave us reassuring results, we feel that more efforts should be made to fully investigate the theoretical inferential properties of the proposed inferential procedure.

As far as the computational load of our estimation procedure is concerned, implementing our algorithm with the help of the OpenBUGS software [21] using the package R2WinBUGS in R [29], and using standard commercial personal computers, the computing times are still demanding. Just to give an example, with 25 observations on a grid simulated assuming the powered exponential autocorrelation function and the value of the parameters used to obtain the simulated distributions in the second row of Fig. 2, one iteration of the MCEM algorithm (with an MCMC sample size of 800) took 41 s. Increasing the size of the grid to 49 observations, the computing time increases to 102 s. Let us note that much of the time is needed for the maximization step of the MCEM algorithm. In the former case, the time needed to generate the MCMC sample was less than 1 second, whereas the time needed by the maximization step was 40 s. Thus, to obtain one MCEM estimate, using 800 iterations of the MCEM, takes more than 9 h, and to obtain a simulated distribution, based on 50 replicates, of the MCEM estimator (that is, one row of Fig. 2) takes several days.

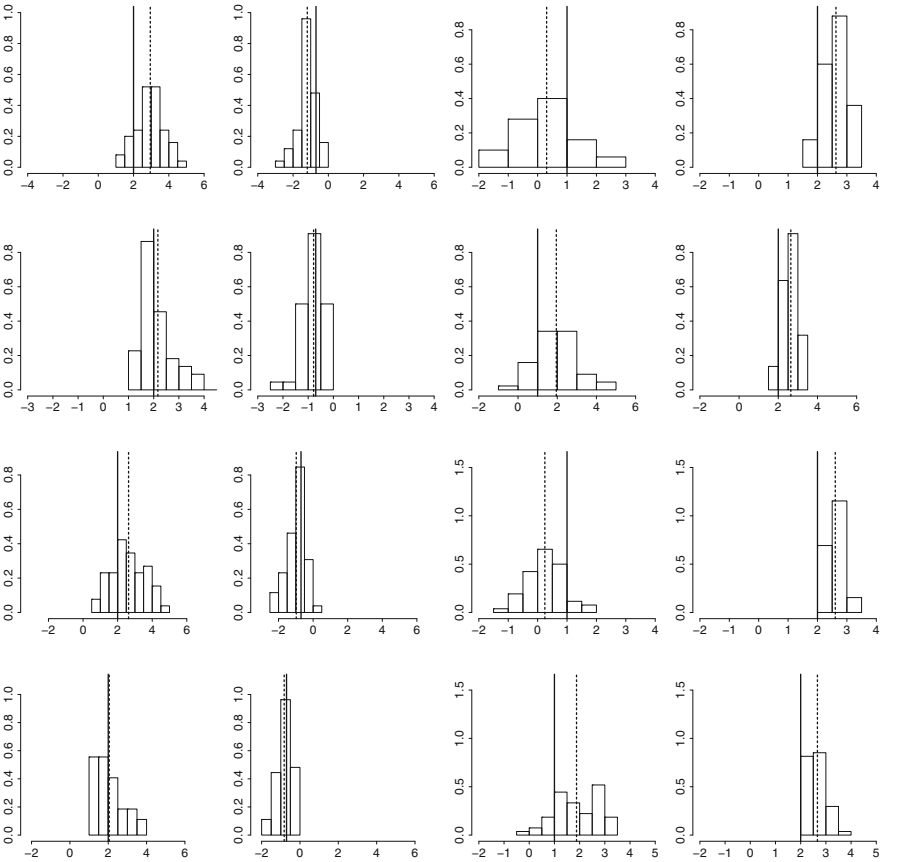

Fig. 2 The histograms show the simulated univariate marginal sampling distributions of the MCEM estimator of the parameters a_{11}, a_{21}, β_1 and β_2 (*from left to right*) in a model with $m = 2$ and $P = 1$ obtained in four simulation experiments (*from top to bottom*). The *vertical solid lines* represent the true parameter values, whereas the *vertical dashed lines* represent the empirical means over the 50 simulated realizations. For the spatial autocorrelation function $\rho(\mathbf{h})$ we chose a powered exponential model with $\gamma = 0.00001$ and $\eta = 1.5$ in the first two simulation experiments (*first two rows*), and a Cauchy model with $\gamma = 7{,}000$ and $\eta = 1$ in the last two simulation experiments (*last two rows*). The parameters α_1 and α_2 were fixed equal to: -1 and 1 (*first row*); 2 and 2 (*second row*); -1 and 1 (*third row*); 2 and 2 (*fourth row*). For all four simulation experiments, the other parameters where equal to: $a_{11} = 2, a_{21} = -0.7, \beta_1 = 1, \beta_2 = 2, \omega_1 = 1, \omega_2 = 1$

Conclusion

In this work we have proposed and studied a model for the analysis of multivariate geostatistical data showing some degree of skewness. Our geostatistical model based on latent factors can be considered as an extension to skewed non-Gaussian data of the classical geostatistical proportional covariance model.

By framing our model in a hierarchical context, that is, by extending to the multivariate case the model-based geostatistical approach in [10], it would be possible to extend the present work to deal with regionalized variables of different kind. Instead of assuming that the conditional distributions of $Y_i(\mathbf{x})$ given $Z_i(\mathbf{x})$ are all skew-normal, we might assume, for different values of $i = 1, \ldots, m$, that they are of different type. For instance, [25] considers a model in which some of the (conditional) distributions, of the observable regionalized variables, are Poisson whereas some others are Gamma. In this way, we could obtain a model for non-Gaussian data flexible enough to account for observable regionalized variables showing different departures from normality.

On the other hand, a generalization in a different direction might involve the introduction of more spatial scales as in the classical linear model of coregionalization. This would supply a more flexible spatial autocorrelation structure in which the latent processes $Z_i(\mathbf{x})$, which are behind the level of the observable regionalized variables $Y_i(\mathbf{x})$, are not constrained to have proportional covariance and cross-covariance functions. However, the high level of complexity of this generalization would require a large amount of data to be detected and would pose serious inferential problems.

As regard to the model presented in this work, we presented a computationally intensive likelihood based inferential procedure, exploiting the capabilities of the MCEM algorithm. It must be noted that with this procedure we estimated just some of the parameters of the model, assuming the others as known. In particular, we assumed as known the parameters $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)^T$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T$ that characterize the shape of the skew-normal (conditional) distributions. In this way we avoided many of the well known inferential problems posed by the estimation of the parameters of the skew-normal distribution. Although in this work we did not discuss any inferential procedure for these parameters, these can nevertheless be calibrated comparing the theoretical marginal distributions and the theoretical variograms with the corresponding empirical counterparts. From a computational perspective, although we checked the feasibility of our estimation procedure for reasonable sample sizes and for different parameter values, it must be remarked that in more complex situations the computational burthen might increase considerably.

Acknowledgements We gratefully acknowledge funding from the Italian Ministry of Education, University and Research (MIUR) through PRIN 2008 project 2008MRFM2H.

Appendix

In this appendix we report some distributional results regarding the observable processes $Y_i(\mathbf{x})$. Let us first recall some definitions. Following, for instance, [2], we say that a random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ has an *extended skew-normal distribution* with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\alpha}$ and τ, and we write $\mathbf{Y} \sim \mathrm{ESN}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \tau)$, if it has probability density function of the form

$$f(\mathbf{y}) = \phi_n(\mathbf{y} - \boldsymbol{\mu}; \boldsymbol{\Sigma}) \cdot \Phi(\alpha_0 + \boldsymbol{\alpha}^T \mathbf{D}^{-1}(\mathbf{y} - \boldsymbol{\mu}))/\Phi(\tau), \quad \text{for} \quad \mathbf{y} \in \mathbb{R}^n, \tag{10}$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ is a vector of location parameters, $\phi_n(\cdot\,; \boldsymbol{\Sigma})$ is the n-dimensional normal density function with zero mean vector and (positive-definite) variance-covariance matrix $\boldsymbol{\Sigma}$ having elements σ_{ij}, $\Phi(\cdot)$ is the scalar N(0,1) distribution function, $\mathbf{D} = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{nn})^{1/2}$ is the diagonal matrix formed with the standard deviations of the scale matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\alpha} \in \mathbb{R}^n$ is a vector of skewness parameters, and $\tau \in \mathbb{R}$ is an additional parameter. Moreover, $\alpha_0 = \tau(1 + \boldsymbol{\alpha}^T \mathbf{R} \boldsymbol{\alpha})^{1/2}$ where $\mathbf{R}$ is the correlation matrix associated to $\boldsymbol{\Sigma}$, that is, $\mathbf{R} = \mathbf{D}^{-1} \boldsymbol{\Sigma} \mathbf{D}^{-1}$. Clearly, this distribution extends the multivariate normal distribution through the parameter vector $\boldsymbol{\alpha}$, and for $\boldsymbol{\alpha} = 0$ it reduces to the latter. When $\tau = 0$, also $\alpha_0 = 0$ and (10) reduces to

$$f(\mathbf{y}) = 2 \cdot \phi_n(\mathbf{y} - \boldsymbol{\mu}; \boldsymbol{\Sigma}) \cdot \Phi(\boldsymbol{\alpha}^T \mathbf{D}^{-1}(\mathbf{y} - \boldsymbol{\mu})), \quad \text{for} \quad \mathbf{y} \in \mathbb{R}^n. \tag{11}$$

In this case we simply say that $\mathbf{Y}$ has a *skew-normal distribution* and we write, more concisely, $\mathbf{Y} \sim \mathrm{SN}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$.

According to [13] and [14], we say that the n-dimensional random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ has a multivariate *closed skew-normal distribution*, and we write $\mathbf{Y} \sim \mathrm{CSN}_{n,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}_c, \boldsymbol{\nu}, \boldsymbol{\Delta})$, if it has probability density function of the form

$$f(\mathbf{y}) = \frac{1}{\Phi_m(\mathbf{0}; \boldsymbol{\nu}, \boldsymbol{\Delta} + \mathbf{D}_c^T \boldsymbol{\Sigma} \mathbf{D}_c)} \cdot \phi_n(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \Phi_m(\mathbf{D}_c^T(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \boldsymbol{\Delta}), \quad \text{for} \quad \mathbf{y} \in \mathbb{R}^n, \tag{12}$$

where: m is an integer greater than 0; $\boldsymbol{\mu} \in \mathbb{R}^n$; $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is a positive-definite matrix; $\mathbf{D}_c \in \mathbb{R}^{n \times m}$ is an $n \times m$ matrix; $\boldsymbol{\nu} \in \mathbb{R}^m$ is a vector; $\boldsymbol{\Delta} \in \mathbb{R}^{m \times m}$ is a positive-definite matrix; and $\phi_n(\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_n(\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are the probability density function and the cumulative distribution function, respectively, of the n-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Though, as we have already noticed, the multivariate finite-dimensional marginal distributions of the multivariate spatial process $(Y_1(\mathbf{x}), \ldots, Y_m(\mathbf{x}))^T$, for $\mathbf{x} \in \mathbb{R}^2$, are not skew-normal (in the sense of [2]), it is possible to show that they are closed skew-normal, according to the definition of [13]. This implies that, for any given $i = 1, \ldots, m$, each univariate spatial process $Y_i(\mathbf{x})$ has all its finite-dimensional marginal distributions that are closed skew-normal. To see this (see also [24]), consider n spatial locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and the corresponding n-dimensional random vector $\mathbf{Y} = (Y_i(\mathbf{x}_1), \ldots, Y_i(\mathbf{x}_n))^T$. Recalling that for any given $\mathbf{x} \in \mathbb{R}^2$ we can write $Y_i(\mathbf{x}) = \beta_i + Z_i(\mathbf{x}) + \omega_i S_i(\mathbf{x})$, the vector $\mathbf{Y}$ can be written as $\mathbf{Y} = \beta_i \mathbf{1}_n + \mathbf{Z} + \mathbf{D}_\omega \mathbf{S} = \mathbf{W} + \mathbf{V}$, where $\mathbf{W} = \beta_i \mathbf{1}_n + \mathbf{Z}$, $\mathbf{V} = \mathbf{D}_\omega \mathbf{S}$, $\mathbf{Z} = (Z_i(\mathbf{x}_1), \ldots, Z_i(\mathbf{x}_n))^T$, $\mathbf{S} = (S_i(\mathbf{x}_1), \ldots, S_i(\mathbf{x}_n))^T$ and $\mathbf{D}_\omega$ is the $n \times n$ diagonal matrix with ω_i on the diagonal. Now, since $S_i(\mathbf{x})$, for $\mathbf{x} \in \mathbb{R}^2$, are independently and identically distributed as $\mathrm{CSN}_{1,1}(0, 1, \alpha_i, 0, 1)$, according to Theorem 3 of [14], we have that $\mathbf{S} \sim \mathrm{CSN}_{n,n}(0, \mathbf{I}_n, \mathbf{D}_\alpha, 0, \mathbf{I}_n)$, where $\mathbf{D}_\alpha$ is the $n \times n$ diagonal matrix with α_i on the diagonal. On the other hand, since $\mathbf{Z}$ follows a multivariate normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}_Z$ with entries given by $\mathrm{Cov}\big[Z_i(\mathbf{x}), Z_i(\mathbf{x} + \mathbf{h})\big] = \varsigma_i^2 \rho(\mathbf{h})$, we also have that $\mathbf{Z} \sim \mathrm{CSN}_{n,1}(0, \boldsymbol{\Sigma}_Z, 0, 0, 1)$. Moreover, being $\mathbf{W}$ distributed as a multivariate normal with mean $\beta_i \mathbf{1}_n$ and covariance matrix $\boldsymbol{\Sigma}_Z$, we can write that $\mathbf{W} \sim \mathrm{CSN}_{n,1}(\beta_i \mathbf{1}_n, \boldsymbol{\Sigma}_Z, 0, 0, 1)$, and using Theorem 1 of [14] we can also write that $\mathbf{V} \sim \mathrm{CSN}_{n,n}(0, \mathbf{D}_{\omega^2}, \mathbf{D}_{\alpha/\omega}, 0, \mathbf{I}_n)$, where $\mathbf{D}_{\omega^2}$ is the $n \times n$ diagonal matrix with ω_i^2 on the diagonal, and $\mathbf{D}_{\alpha/\omega}$ is the $n \times n$ diagonal matrix with α_i / ω_i on the diagonal. Thus, considering that $\mathbf{Y} = \mathbf{W} + \mathbf{V}$, we can conclude, using Theorem 4 of [14], that $\mathbf{Y} \sim \mathrm{CSN}_{n,n+1}(\beta_i \mathbf{1}_n, \boldsymbol{\Sigma}_Z + \omega_i^2 \mathbf{I}_n, \mathbf{D}^*, 0, \boldsymbol{\Delta}^*)$, for some matrices $\mathbf{D}^*$ and $\boldsymbol{\Delta}^*$.

References

1. Allard, D., Naveau, P.: A new spatial skew-normal random field model. Commun. Stat. Theory **36**, 1821–1834 (2007)
2. Azzalini, A.: The skew-normal distribution and related multivariate families. Scand. J. Stat. **32**, 159–188 (2005)
3. Azzalini, A., Dalla Valle, A.: The multivariate skew-normal distribution. Biometrika **83**, 715–726 (1996)
4. Bechler, A., Romary, T., Jeannée, N., Desnoyers, Y.: Geostatistical sampling optimization of contaminated facilities. Stoch. Env. Res. Risk A. **27**, 1967–1974 (2013)
5. Brenning, A., Dubois, G.: Towards generic real-time mapping algorithms for environmental monitoring and emergency detection. Stoch. Env. Res. Risk A. **22**, 601–611 (2008)
6. Chagneau, P., Mortier, F., Picard, N., Bacro, J.-N.: A hierarchical Bayesian model for spatial prediction of multivariate non-Gaussian random fields. Biometrics **67**, 97–105 (2011)
7. De, S., Faria, Á.E.: Dynamic spatial Bayesian models for radioactivity deposition. J. Time Ser. Anal. **32**, 607–617 (2011)
8. De Oliveira, V., Kedem, B., Short, D.A.: Bayesian prediction of transformed Gaussian random fields. J. Am. Stat. Assoc. **92** 1422–1433 (1997)

9. Desnoyers, Y., Chilès, J.-P., Dubot, D., Jeannée, N., Idasiak, J.-M.: Geostatistics for radiological evaluation: study of structuring of extreme values. Stoch. Env. Res. Risk A. **25**, 1031–1037 (2011)
10. Diggle, P.J., Moyeed, R.A., Tawn, J.A.: Model-based geostatistics (with discussion). Appl. Stat. **47**, 299–350 (1998)
11. Dubois, G., Galmarini, S.: Spatial interpolation comparison (SIC) 2004: introduction to the exercise and overview of results. In: Dubois, G. (ed.) Automatic Mapping Algorithms for Routine and Emergency Monitoring Data - Spatial Interpolation Comparison 2004, Office for Official Publication of the European Communities (2005)
12. Fort, G., Moulines, E.: Convergence of the Monte Carlo expectation maximization for curved exponential families. Ann. Stat. **31**, 1220–1259 (2003)
13. González-Farías, G., Domínguez-Molina, J.A., Gupta, A.K.: The closed skew-normal distribution. In: Genton, M.G. (ed.) Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality, pp. 25–42. Chapman & Hall/CRC, London (2004)
14. González-Farías, G., Domínguez-Molina, J.A., Gupta, A.K.: Additive properties of skew-normal random vectors. J. Stat. Plan. Infer. **126**, 521–534 (2004)
15. Gräler, B.: Modelling skewed spatial random fields through the spatial vine copula. Spat. Stat. (2014). http://dx.doi.org/10.1016/j.spasta.2014.01.001
16. Herranz, M., Romero, L.M., Idoeta, R., Olondo, C., Valiño, F., Legarda, F.: Inventory and vertical migration of ^{90}Sr fallout and ^{137}Cs/^{90}Sr ratio in Spanish mainland soils. J. Env. Radioact. **102**, 987–994 (2011)
17. Hosseini, F., Eidsvik, J., Mohammadzadeh, M.: Approximate Bayesian inference in spatial GLMM with skew normal latent variables. Comput. Stat. Data Anal. **55**, 1791–1806 (2011)
18. Kazianka, H., Pilz, J.: Copula-based geostatistical modeling of continuous and discrete data including covariates. Stoch. Env. Res. Risk A. **24**, 661–673 (2010)
19. Kazianka, H., Pilz, J.: Bayesian spatial modeling and interpolation using copulas. Comput. Geosci. **37**, 310–319 (2011)
20. Kim, H.-M., Mallick, B.K.: A Bayesian prediction using the skew Gaussian distribution. J. Stat. Plan. Infer. **120**, 85–101 (2004)
21. Lunn, D., Spiegelhalter, D., Thomas, A., Best, N.: The BUGS project: evolution, critique and future directions. Stat. Med. **28**, 3049–3067 (2009)
22. Maglione, D.S., Diblasi, A.,M.: Exploring a valid model for the variogram of an isotropic spatial process. Stoch. Env. Res. Risk A. **18**, 366–376 (2004)
23. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley, New York (2007)
24. Minozzo, M., Ferracuti, L.: On the existence of some skew-normal stationary processes. Chil. J. Stat. **3**, 157–170 (2012)
25. Minozzo, M., Ferrari, C.: Multivariate geostatistical mapping of radioactive contamination in the Maddalena Archipelago (Sardinia, Italy). AStA Adv. Stat. Anal. **97**, 195–213 (2013)
26. Minozzo, M., Fruttini, D.: Loglinear spatial factor analysis: an application to diabetes mellitus complications. Environmetrics **15**, 423–434 (2004)
27. Oliver, M.A., Badr, I.: Determining the spatial scale of variation in soil radon concentration. Math. Geol. **27**, 893–922 (1995)
28. Pilz, J., Spöck, G.: Why do we need and how should we implement Bayesian kriging methods. Stoch. Env. Res. Risk A. **22**, 621–632 (2008)
29. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (2012)
30. Ren, Q., Banerjee, S.: Hierarchical factor models for large spatially misaligned data: a low-rank predictive process approach. Biometrics **69**, 19–30 (2013)

31. Schmidt, A.M., Rodriguez, M.A.: Modelling multivariate counts varying continuously in space. Bayesian Stat. **9** (2011). doi:10.1093/acprof:oso/9780199694587.003.0020
32. Spöck, G.: Spatial sampling design with skew distributions: the special case of trans-Gaussian kriging. Ninth International Geostatistical Congress, Oslo, Norway, 11–15 June 2012, 20 pages
33. Wibrin, M., Bogaert, P., Fasbender, D.: Combining categorical and continuous spatial information within the Bayesian maximum entropy paradigm. Stoch. Env. Res. Risk A. **20**, 423–433 (2006)
34. Zhang, H.: On estimation and prediction for spatial generalized linear mixed models. Biometrics **58**, 129–136 (2002)
35. Zhang, H., El-Shaarawi, A.: On spatial skew-Gaussian processes and applications. Environmetrics **21**, 33–47 (2010)

Modelling the Length of Stay of Geriatric Patients in Emilia Romagna Hospitals Using Coxian Phase-Type Distributions with Covariates

Adele H. Marshall, Hannah Mitchell, and Mariangela Zenga

Abstract

The attention placed on healthcare systems has been constantly increasing in recent years. This is especially true for geriatric services: older people often have complex medical and social needs and the proportion of elderly in the population is currently rising. In this paper we apply the Coxian phase-type distribution to model the length of stay of geriatric patients admitted to 19 geriatric wards at hospitals in the Emilia-Romagna region in Italy for the years 2008–2011. The results confirm previous research carried out on patients in the UK and extends the research by allowing the influence of patient characteristics, available on admission, to be taken into account as covariates.

Keywords

Coxian phase type distribution • Covariates • Geriatric wards • Emilia Romagna Region

A.H. Marshall (✉) • H. Mitchell
Centre for Statistical Science and Operational Research (CenSSOR), Queen's University, Belfast, Northern Ireland, UK
e-mail: a.h.marshall@qub.ak.uk; hmitchell03@qub.ac.uk

M. Zenga
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano, Italy
e-mail: mariangela.zenga@unimib.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_21, Published online: 12 November 2014

1 Introduction

In the last 15 years the proportion of elderly people has increased across all European countries. This means a growth of the service for the health care system dedicated to the people aged more than 65 years, in particular an increase in the expenditure due to an overall increase in patient length of stay (LoS) in hospital. In Italy in 2012 the elderly people comprise 37 % of the admissions to hospital consuming nearly half (49 %) of the LoS days. It has been estimated [4] that in 2050 the ageing of the population will produce an increase of 4–8 % of the GDP across Europe. In contrast, recent years in Italy has also seen the closure of several pediatric wards replaced by geriatric wards. The modeling of hospital wards and patient activity can be addressed by focusing on techniques that consider the length of stay that patients experience in hospital. In particular, the study of duration of stay of geriatric patients in hospital has led to the modeling of survival data using a two term mixed exponential distribution. Further work has resulted in the successful representation of this distribution using a Coxian phase-type distribution along with key patient characteristics known on admission to hospital [11]. The purpose of the research presented in this paper is to use the Coxian phase-type distribution to consider patient length of stay of the elderly population in hospitals in the Emilia-Romagna region in Italy for data recorded more recently between 2008 and 2011. By doing so, the survival data will be represented as different stages of care presenting an opportunity to investigate the different streams of care through the system, and the characteristics involved that influence this. The results can also offer insight into the needs and behavior of this growing cohort of elderly patients found in almost all European hospitals.

2 The Coxian Phase-Type Distribution

Past investigations of modelling length of stay have led to the discovery that a two-term mixed exponential model produces a good representation of patient length of stay [13, 14]. Since then further research has endeavoured to improve the mixed exponential models with the incorporation of more complex compartmental systems and more sophisticated stochastic models such as the Coxian phase-type distribution. The Coxian phase-type distributions are a subset of the widely used phase-type distributions introduced by Neuts in 1975, they have the benefit of overcoming the problem generality within phase-type distributions by only requiring $2n - 1$ parameters to describe a distribution requiring n phases, whereas the general phase-type distribution requires $n^2 + 2$ [15]. The Coxian phase-type distributions have been used in a variety of settings: from component failure data [6] and the length of treatment for patients at risk of suicide to prisoner remand times and the lifetime of male rats [5]. Marshall et al. [12] used the Coxian phase-type distribution to model the career progression of students at university where the process can be thought of as a series of transitions through latent phases

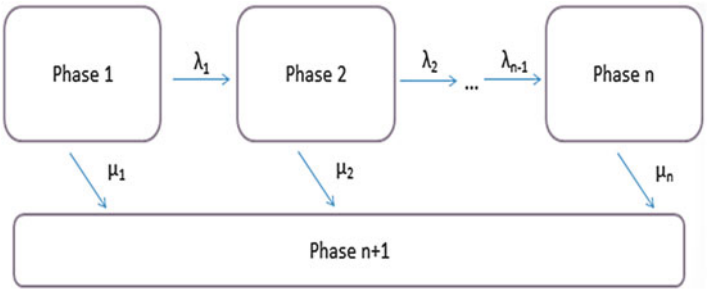

Fig. 1 Coxian phase-type distribution

until the event of leaving the university occurs. However most applications of the Coxian phase-type distribution has been made in modelling the length of time spent in hospitals in particular McClean et al. [9] showed that the Coxian phase-type distribution was appropriate for describing the length of time United Kingdom geriatric patients spent in care. The distributions have also been used to model the stages of progression of the patients from first entering the hospital through to the individual leaving due to recovery or death. The transitions through the ordered transient states could correspond to the stages in patient care such as diagnosis, assessment, rehabilitation and long-stay care where the patients eventually will then reach the absorbing state of the Coxian phase-type distribution which will correspond to them leaving the hospital through discharge, transfer or death [5].

Methodologically speaking, the Coxian phase-type distribution represents the time to absorption of a finite latent Markov chain in continuous time where there is a single absorbing state and the stochastic process starts in a transient state. It describes the probability $P(t)$ that the process is still active at time t and differs from the general phase-type distribution in that the transient states (phases) of the model are ordered (see Fig. 1). The process begins in the first phase and either moves sequentially through the phases or into the absorbing state. In other terms, a Coxian phase-type distribution results when the transient states have a natural order and only forward transitions between them may occur. These phases may be used to describe the stages of a process until termination, in which the transition rates need to be estimated.

The transient states in this model could have some real world meaning attached to them, for example within a hospital environment each of the stages could be thought of as the progression of treatment: the first state could be admittance, followed by diagnosis, treatment and rehabilitation. During each state the individual can leave hospital due to discharge, transfer or death.

Definition 1 Let $(X(t); t \geq 0)$ be a latent Markov chain in continuous time with states $1, 2, \ldots, n, n+1$ and $X(0) = 1$. For $i = 1, 2, \ldots, n-1$ the probability that a patient will move from one phase to the next phase in the system, during the time

interval δt may be written as

$$\text{prob}\{X(t+\delta t) = i+1 | X(t) = i\} = \lambda_i \delta t + o(\delta t) \tag{1}$$

and likewise for i=1,2,...,n, the probability that a patient, during the time interval δt, will leave the system completely and enter the absorbing phase may be written as

$$\text{prob}\{X(t+\delta t) = n+1 | X(t) = i\} = \mu_i \delta t + o(\delta t). \tag{2}$$

The states 1,2,...,n are latent (transient) states of the process and state n+1 is the absorbing state, while λ_i represents the rates of movement from state i to state (i+1) and μ_i is the rates of transition from state i to the absorbing state (n+1).

The time until the absorption $T = \{t \geq 0 | X(t) = n+1\}$ is said to have a Coxian Phase Type distribution, and the probability density function of T can be written as follow:

$$f(t) = \mathbf{p}\exp(\mathbf{Q}t)\mathbf{q} \tag{3}$$

where $\mathbf{Q}$ is the matrix of transition rates between states[1],

$$\mathbf{Q} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \dots & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -(\lambda_{n-1}+\mu_{n-1}) & \lambda_{n-1} \\ 0 & 0 & 0 & \dots & 0 & -\mu_n \end{pmatrix} \tag{4}$$

$\mathbf{q} = -\mathbf{Q}\mathbf{e} = (\mu_1, \mu_2, \dots, \mu_n)^T$, $\mathbf{p} = (1, 0, \dots, 0)$ and $\mathbf{e} = (1, 1, \dots, 1)$
The survival function is given by

$$S(t) = \mathbf{p}\exp(\mathbf{Q}t)\mathbf{e}. \tag{5}$$

The number of parameters in a Coxian phase-type distribution is equal to $2n-1$. The Coxian family is dense in the class of all distributions on $[0, \infty]$ and is appropriate for estimating long-tailed distributions.

Gardiner et al. [8, 18] reformulate the probability density function (3) in the following terms.

Let

$$\delta_k = \lambda_k + \mu_k \tag{6}$$

be the hazard rates in transient states, for $k = 1, 2, \dots, n-1$ and $\delta_n = \mu_n$.

[1]The expression $exp(A)$ denotes the so-called matrix exponential $exp(A) = \sum_{k=0}^{\inf} \frac{A^k}{k!}$ linked to the solutions of differential equations defining Markov chains in continuous time.

Moreover, let

$$p_{k,k+1} = \frac{\lambda_k}{\lambda_k + \mu_k} \tag{7}$$

be the probabilities of transition from $k \longrightarrow k+1$, $k = 1, 2, \ldots, n-1$, and $p_{n,n+1} = 1$.

Suppose we observe the event times of m individuals $\mathbf{t} = (t_1, \ldots, t_m)$ from a Coxian phase-type distribution with n transient states.

Let $\mathbf{X}$ be the covariate information matrix: $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_m})$ where $\mathbf{x_i} = [x_{1i}, x_{2i}, \ldots, x_{li}]^T$. In previous work the covariates were incorporated in generalised linear models, see for example Faddy and McClean [7] and McClean et al. [10]. In Gardiner's approach, it is possible to introduce the covariate effects into the distribution.

Let

$$\delta_{ki} = \delta_k(\mathbf{x}_i) = \delta_{0k}\exp(-\mathbf{x}_i^T\boldsymbol{\beta}) \tag{8}$$

be the hazard rate of the i-th individual into the k-th phase, where $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_l]^T$ is the coefficient vector. In this way the conditional mean time is log-linear in $\mathbf{x}$, that is $log(E(T|\mathbf{x_i}) = a_0 + \mathbf{x}_i^T\boldsymbol{\beta}$, where $a_0 = \frac{1}{\delta_{01}} + \sum_{k=2}^{n}(\prod_{j=1}^{k-1}\frac{p_{j,j+1}}{\delta_{0k}})$.

The likelihood function of (2) becomes:

$$L(t|\mathbf{X}, \boldsymbol{\delta_0}, \boldsymbol{\beta}, \mathbf{p}) = \prod_{i=1}^{m}\mathbf{p}[\exp(\boldsymbol{\Lambda_i}\mathbf{P}t_i)](-\boldsymbol{\Lambda_i}\mathbf{Pe}) = \prod_{i=1}^{m}\mathbf{p}\exp(\tilde{\mathbf{Q}}_i t_i)\tilde{\mathbf{q}}_i \tag{9}$$

where

$$\Lambda_i = \begin{pmatrix} -\delta_{1i} & 0 & \ldots & 0 \\ 0 & -\delta_{2i} & \ldots & 0 \\ : & : & : & : \\ 0 & 0 & \ldots & -\delta_{ni} \end{pmatrix}$$

and

$$P = \begin{pmatrix} 1 & -p_{12} & 0 & \ldots 0 \\ 0 & 1 & -p_{23} & \ldots 0 \\ : & : & : & : \\ 0 & 0 & 0 & \ldots 1 \end{pmatrix}$$

$\tilde{\mathbf{Q}}_i = \boldsymbol{\Lambda}_i\mathbf{P}$ and $\tilde{\mathbf{q}}_i = -\boldsymbol{\Lambda}_i\mathbf{Pe} = (p_{1,n+1}\delta_{1i}, p_{2,n+1}\delta_{2i}, \ldots, p_{n,n+1}\delta_{ni})^T$. The likelihood function (9) is given by:

$$\prod_{i=1}^{m} \mathbf{p} \exp\left(\exp(-\mathbf{x}_i^T \boldsymbol{\beta}) \begin{pmatrix} -\delta_{01} & p_{12}\delta_{01} & \dots & 0 \\ 0 & -\delta_{02} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & -\delta_{0n} \end{pmatrix} t_i \right) \times$$

$$\times \left(\exp(-\mathbf{x}_i^T \boldsymbol{\beta}) \begin{pmatrix} (1-p_{12})\delta_{01} \\ (1-p_{23})\delta_{02} \\ \vdots \\ \delta_{0n} \end{pmatrix} \right). \tag{10}$$

The stability of the maximum likelihood estimates is in doubt unless some structural simplifications can be made. The authors used the Bayesian methods proposed in Ausín [1–3], incorporating covariates into the model and extending the Bayesian method to fit the Coxian phase-type regression model.

It was then assumed without loss of generality that $\delta_{01} \geq \delta_{02} \geq \ldots \geq \delta_{0m}$. One way of incorporating this ordering restriction was to represent the hazard rates of the model as follows :

$$\delta_{0k} = \delta_{01}\upsilon_2\upsilon_3 \ldots \upsilon_k, \ 0 < \upsilon_j \leq 1, \ j,k = 2,3,\ldots,n \tag{11}$$

as proposed in [1–3]. The transition probabilities can be obtained from $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ where η_k is the total probability of exiting from transient state k.

$$\begin{aligned} \eta_1 &= 1 - p_{12} \\ \eta_2 &= p_{12}(1 - p_{23}) \\ &\vdots \\ &\vdots \\ \eta_{n-1} &= p_{12}p_{23} \ldots p_{n-2,n-1}(1 - p_{n-1,n}) \\ \eta_m &= p_{12}p_{23} \ldots p_{n-2,n-1}p_{n-1,n}. \end{aligned} \tag{12}$$

Instead of the parameters $(n, \delta_0 1, \boldsymbol{\beta}, \mathbf{p})$, the new parametrization $(n, \boldsymbol{\eta}, \delta_0 1, \boldsymbol{\beta}, \boldsymbol{\upsilon})$ allows us to find a solution more stable using a MCMC framework, as shown in Tang et al. [18]. For example, for a Coxian phase-type distribution with two phases, the survival function with the new parametrization becomes:

$$\begin{aligned} S(t) &= (1-\eta_2)\exp(-\delta_{01}t) + \frac{\eta_2}{1-\upsilon_2}\{\exp(-\delta_{01}\upsilon_2 t) - \upsilon_2 \exp(-\delta_{01}t)\} = \\ &= \left(1 - \frac{\eta_2}{1-\upsilon_2}\right)\exp(-\delta_{01}t) + \left(\frac{\eta_2}{1-\upsilon_2}\right)\exp(-\delta_{01}\upsilon_2 t). \end{aligned}$$

Several routines are implemented for the Coxian phase-type distribution. A well known optimisation function is the Nelder Mead algorithm using maximum likelihood techniques to determine the transition rates within the distribution. The EM-algorithm is also used having the advantage of preserving the structures of zeroes. In general, Matlab, C, SAS and R softwares can be used to fit the Coxian phase-type distribution. Payne et al. [16] investigated the efficiency of fitting the Coxian phase-type distributions to healthcare data using these programs. They concluded that SAS was their software package of choice but that the Matlab and EMpht programs consistently had a high rate of convergence. Gardiner [8] coded their approach in SAS using PROC SEVERITY [17].

3 The Data

The data used in this paper consists of the ordinary admissions of 66,728 patients aged 65 years or older to every geriatric ward of the acute care hospitals (19 geriatric wards in total) operating in the North-East Italian Region of Emilia-Romagna between 2008 and 2011. The data was provided by the Italian Health Care Ministry and it is a subset of a large administrative data set covering all of Italy's geriatric wards. Individual Hospital Discharge Charts (HDC) are reported in the data set including patient information (gender, age, and residence), the hospital (regional code), the treatments received during hospitalization including information such as Disease Related Group (DRG), principal and secondary diagnoses and procedures, data of admission and so on. Patients aged 85 years or older represents 50 % of all patients in the data set. Approximately 42 % (27,838 patients) of patients were male with only 2.7 % of patients not living in the Emilia-Romagna Region. Eighty-three percent (55,401 patients) of patients were admitted to the department of geriatric medicine from emergency admission; 12 % (8,078) had emergency GP admission, 2 % (1,451) were transferred from another institute and the remaining 2.4 % (1,798) had a planned admission or other. Approximately one quarter of patients were admitted for surgery (16,878 patients). Chronic patients represent approximately 30 % of all patients. Moreover 31 % (20,756 patients) of the admitted patients had a principal diagnosis of circulatory system, 20 % (12,921 patients) as respiratory system problems, 9 % (5,676 patients) problems of the digestive system, and 7 % (4,787 patients) were admitted with cancer. The destination of patients on departure from hospital could be one of several possibilities: the patient may return home; transfer to a nursing home, residential home, another ward, or other hospital; or may die while in hospital. Outcome was coded to describe three locations: home, transfer, or death. Approximately 70 % of patients left the geriatric ward to return home; 13 % died while the remaining 17 % transferred. The highly right skewed patient length of stay distribution is illustrated in Fig. 2. The average length of stay is nine days (SD = 5.502) and median eight. It is interesting to note that the maximum value is 30 days with no further tail of the distribution beyond this one month point. A logrank test was conducted to compare the survival distributions

Table 1 Results for log-rank test

Variable	Chi-Squared	Df	Sig
Gender	18.944	1	< 0.0001
Outcome	1490.325	2	< 0.0001
Way of Admission	172.559	4	< 0.0001
Chronic	56.403	1	< 0.0001
Surgery	654.236	1	< 0.0001
Principal Diagnostic Group	616.764	4	< 0.0001
Ward	7046.536	18	< 0.0001

for length of stay according to the separate patient characteristics. Table 1 shows that, for all variables (except for place of residence), we reject the null hypothesis of equality for the survival function across all categories.

4 The Results

We first estimate Coxian phase-type distributions to model the patient length of stay of the geriatric patients in Emilia-Romagna. The second part of the work is devoted to extending previous work by introducing the covariates into the Coxian phase-type distribution using the Gardiner approach. The EM-algorithm was used to estimate the simple Coxian phase-type distribution, and a SAS routine implemented to estimate the model with covariates.

4.1 The Coxian Phase-Type Distribution for the Distribution of the Length of Stay of Emilia-Romagna's Geriatric Patients

We implemented the actual fitting of the distribution by coding the EM-algorithm to perform iteratively in C. In each iteration, the new parameter estimates are calculated by solving a system of homogeneous linear differential equations using the Runge-Kutta method of fourth order. The programme was stopped when there was no further significant contribution made with the addition of another phase. The Akaike information criterion (AIC) was then calculated and used to find the most appropriate model to represent the data.The results show that the four phase Coxian distribution most suitably represents length of stay of geriatric patients within the Emilia-Romagna Region. At this stage, we carried out a verification of the fitted Coxian phase-type distribution to ensure it is the best model to describe the data set. This was done by separately fitting the alternative Lognormal, Weibull, Log-Logistic, Generalized Beta, Burr XII, Dagum distributions and Coxian phase-type distribution (with 4 phases) to the data set. A comparison of the fitted distributions and corresponding Akaike information criterion (AIC) values was carried out as reported in Table 2.

Table 2 Comparison of the commonly used distributions against Coxian phase-types to describe length of stay at the hospital

Distribution	N.Parameters	−2Loglikelihood	AIC
Lognormal	2	407177.40	407181.40
Weibull	2	404885.60	404889.60
Log-Logistic	2	406026.00	406030.00
Generalized Beta	2	403022.40	403026.40
Burr XII	2	403309.80	403313.80
Dagum	3	404190.00	404196.00
Coxian	7	384870.00	384884.00

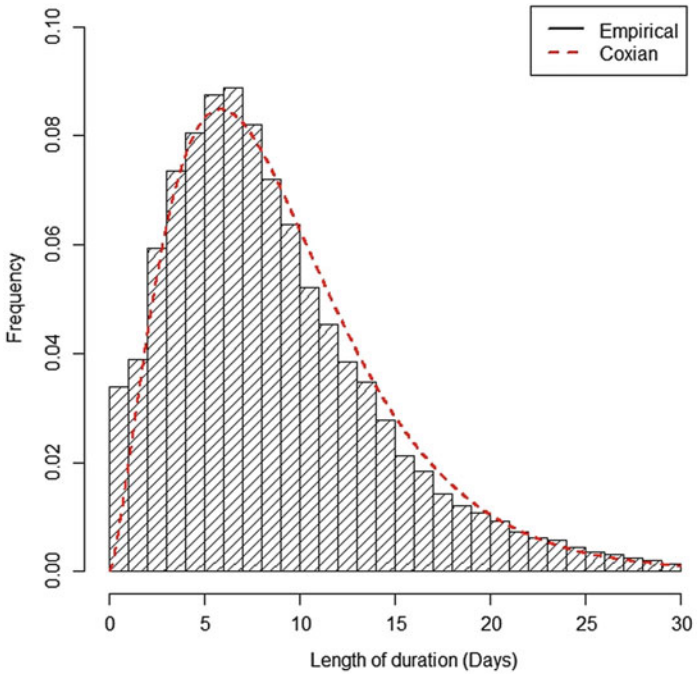

Fig. 2 Length of stay of geriatric patients in the Emilia-Romagna region, 2008–2011

The lowest AIC value corresponds to the Coxian phase-type distribution with four phases thus indicating that it fits the data better than the other models.

Table 2 and Fig. 2 show that the fitted distribution represents the length of stay data well[2]. By illustrating the four phase Coxian distribution with the parameter values, it can be seen that no patients leave the first phase through the absorbing

[2]The length of stay for patients at geriatric ward can be considered as "regular" due to the ageing process: in general this processes could be well fitted by a Coxian phase type distribution as in [7].

Table 3 Fitted Coxian phase-type distributions for patient length of stay by outcome

Destination	Phases	Parameter estimates
Whole population	4	$\mu_1 = 0.000000, \mu_2 = 0.006868,$
		$\mu_3 = 0.036456, \mu_4 = 0.289186$
		$\lambda_1 = 0.553111, \lambda_2 = 0.543983, \lambda_3 = 0.25273$
Home	6	$\mu_1 = 0.001341, \mu_2 = 0, \mu_3 = 0,$
		$\mu_4 = 0, \mu_5 = 0, \mu_6 = 4.224514$
		$\lambda_1 = 0.302514, \lambda_2 = 0.298179,$
		$\lambda_3 = 0.625320, \lambda_4 = 4.22370, \lambda_5 = 4.224372$
Death	4	$\mu_1 = 0.1669483, \mu_2 = 0.0021845,$
		$\mu_3 = 0.1281496, \mu_4 = 0.2958138$
		$\lambda_1 = 0.1402335, \lambda_2 = 0.2936308, \lambda_3 = 0.1677260$
Transfer	6	$\mu_1 = 976.62, \mu_2 = 0, \mu_3 = 0.002500,$
		$\mu_4 = 0.184324, \mu_5 = 000003, \mu_6 = 0.352500$
		$\lambda_1 = 250570.55, \lambda_2 = 0.371965, \lambda_3 = 0.369465,$
		$\lambda_4 = 0.187698, \lambda_5 = 0.352497$

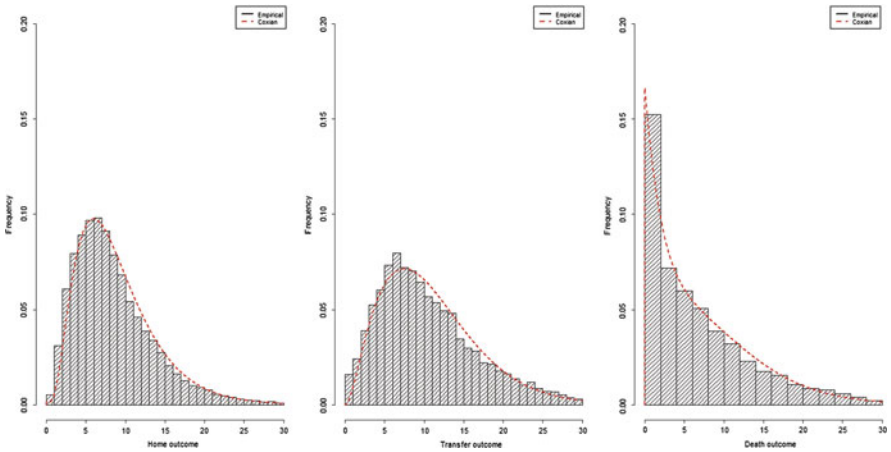

Fig. 3 Length of stay of geriatric patients in the Emilia-Romagna region, 2008–2011

stage (transfer, discharge or death). Patients leave the second state at a very fast rate either through leaving the hospital or moving into another stage of care, the majority of patients continuing through to the third and fourth stages. This suggests that the second phase represents a short acute stay with the remainder of patients all requiring further treatment and or rehabilitation in phase two before leaving the hospital in the later phases.

The Coxian phase-type distribution was then fitted to the three destination groups; home, transfer and death. The results for each of the outcome groups are displayed in Table 3 and Fig. 3. From the graphs it is apparent that the length of stay for the three outcomes is represented well by the distribution, with the parameter

values shown in Table 3. For home and transfer the most suitable representation for the data was the six phase Coxian distribution. For the home group it can be noted that the vast majority of patients went through all the stages of care with only some leaving at the first stage. This could potentially be because those individuals who make it through the first stage (which could be a treatment phase) need a lot of rehabilitation before they can leave for home and so they all progress through to the final sixth phase before leaving the hospital. This outcome group perhaps could collapse down into a two phase distribution considering that all patients after the first state progress to the final stage with the first state possibly representing treatment and the last, rehabilitation. For the transfer group it was found that it too was most suitably represented by the six phase Coxian distribution however by contrast patients leave the hospital from any stage of the process apart from the second state. This suggests that further treatment in a different hospital or indeed a lack of available beds could occur at any stage and so patients are required to be transferred. For the death group however it was found to be most suitably represented by the four phase Coxian distribution, with patients passing away at any stage. This is very different to the other two outcome groups but perhaps there is a logical and reasonable representation for it. This agrees with previous research where the transfer and home outcome groups in general have a length of stay distribution which has more phases than that of the death outcome group. This could be because the transfer patients are having to wait for available beds in their new destination while those returning home may have to wait until not only they are in a suitable condition to return home but also until suitable care and resources are available to them at their own home.

4.2 The Gardiner Approach: Adding the Covariates

Using the approach by Gardiner et al. [8,18] the Emilia-Romagna data set was found to be best represented by a two phase Coxian model which can be further refined slightly by incorporating patient covariates.

It also provides insight into which covariates play a significant part in the patients length of stay. The results can be seen in Table 4. On the basis of the estimates of the 2-phase model, all patients start in state 1 and almost all of them transit from state 1 to state 2. The total probabilities of exiting from state 1 and 2 were 0.675 and 0.325 respectively. Length of stay (LoS) is positively associated with admission into a public hospital, admissions for respiratory disease and admissions with surgery. As expected, the effect of the outcome for home and death on LoS were estimated respectively -0.054 and -0.322, corresponding to an expected decrease in LoS of 0.948 and 0.72 times the LoS of those who transfer. The expected LoS for patients who were admitted as an emergency was 0.95 times the LoS of patients not admitted as an emergency. Moreover the results show that admissions into public hospitals increased the expected LoS by 1.093 times that of private or credited hospitals. Respiratory disease led to an increase in LoS by 1.063 times the LoS of patients with

Table 4 Including covariates into the Coxian phase-type distribution

Parameter	Estimate	Stand. error	t Value	$Pr > \|t\|$
δ_{01}	1.486	0.052	85.61	< .0001
υ_2	0.969	0.017	58.27	< .0001
η_2	0.325	0.002	162.51	< .0001
Out:home	−0.054	0.004	−13.42	< .0001
Out:death	−0.322	0.007	−45.75	< .0001
ER	−0.052	0.004	−12.74	< .0001
Hosp:pub	0.089	0.007	12.60	< .0001
Dis: Respiratory	0.061	0.004	16.35	< .0001
Surgery	0.057	0.004	15.70	< .0001

other kinds of disease. Finally, surgical patients had longer LoS: the expected LoS for surgical patients was 1.059 times the LoS of patients without surgical admission.

Conclusion

The Coxian phase-type distribution is used to represent geriatric patient length of stay in the Emilia-Romagna hospitals for the period of 2008–2011. The results show that the Coxian phase-type distribution suitably represents the data and it also shows a similar pattern of stages of care according to the outcome of the patient on leaving hospital from previous research. The results suggest that patients who are discharged home from hospital will not tend to leave in the first phase but instead will continue through to the final stage with the Coxian phase-type distribution tending to have more phases for this group than for the group of individuals who passed away while in hospital. It is the same for the transfer patients where both of these outcome groups have more phases than the death group.

The paper also considers the incorporation of patient covariates into the Coxian phase-type distribution. This allows the survival time to be modelled in relation to the patient information and identifies which variables have a significant influence on patient length of stay. The incorporation of the covariates in the model results in a Coxian phase-type distribution with fewer phases required to represent the survival distribution. It is therefore reasonable to consider the covariates as providing better insight into the length of stay distribution where their presence reduces the need of a higher phase Coxian. The earlier results of the significance of the destination variable are also confirmed by the results of the covariate model that has destination as a significant covariate in the two phase Coxian model.

References

1. Ausín, M.C., Lopes, H.F.: Bayesian estimation of ruin probabilities with a heterogeneous and heavy-tailed insurance claim-size distribution. Aust. New Zeal. J. Stat. **49**, 415–434 (2007)
2. Ausín, M.C., Wiper, M.P., Lillo, R.E.: Bayesian estimation of finite time ruin probabilities. Appl. Stoch. Model Bus. Ind. **25**, 78–805 (2009)
3. Ausín, M.C., Wiper, M.P., Lillo, R.E.: Bayesian prediction of the transient behaviour and busy period in short- and long-tailed GI/G/1 queueing systems. Comput. Stat. Data Anal. **52**, 1615–1635 (2008)
4. Cesana, G.C.: Il ministero della salute. Note Introduttive Alla Medicina, 2 edn. SEF, Firenze (2005)
5. Faddy, M.J.: Examples of fitting structured phase-type distributions. Appl. Stoch. Model Data Anal. 247–255 (1994)
6. Faddy, M.J.: Phase-type distributions for failure time. Math. Comput. Modell. **22**, 63–70 (1995)
7. Faddy, M.J., McClean S.I.: Markov chain modeling for geriatric patient care. Meth. Inform. Med. 369–373 (2005)
8. Gardiner, J.C.: Modeling heavy-tailed distributions in healthcare utilization by parametric and Bayesian methods. SAS Global Forum (2012)
9. McClean, S.I., Millard, P.: Patterns of length of stay after admission in geriatric medicine: an event history approach. Statistician 263–274 (1993)
10. McClean, S.I., Barton, M., Garg, L., Fullerton, K.: A modeling framework that combines markov models and discrete-event simulation for stroke patient care. ACM Trans. Model. Comput. Simulat. **21**(4), Article No. 25 (2011)
11. Marshall, A.H., McClean, S.I.: Conditional phase-type distributions for modelling patient length of stay in hospital. Int. Trans. Oper. Res. **10**, 565–576 (2003)
12. Marshall, A.H., Zenga, M., Giordano, S.: Modelling students' length of stay at university using coxian phase-type distributions. Int. J. Stat. Probab. 73–89 (2013). Doi:10.5539/ijsp.v2n1p73
13. Millard, P.H.: Geriatric medicine: a new method of measuring bed usage and a theory for planning. MD Thesis, University of London (1988)
14. Millard, P.H.: Throughput in a department of geriatric medicine: a problem of time, space and behaviour. Health Trends **24**, 20–242 (1991)
15. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models, 3rd edn. John Hopkins University Press, Baltimore (1981)
16. Payne, K. Marshall, A.H., Cairns, K.J.: Investigating the efficiency of fitting Coxian phase-type distributions to health care data. IMA. J. Manag. Math. (2011). Doi: 10.1093/imaman/dpr008
17. SAS/ETS 9.22 User's Guide (2010)
18. Tang, X., Luo, Z., Gardiner, J.C.,: Modeling hospital length of stay by Coxian phase-type regression with heterogeneity. Stat. Med. 1502–1516 (2012)

Pathway Composite Variables: A Useful Tool for the Interpretation of Biological Pathways in the Analysis of Gene Expression Data

Daniele Pepe and Mario Grassi

Abstract

Biological pathways represent a useful tool for the identification, in the intricate network of biomolecules, of subnetworks able to explain specific activities in an organism. The advent of high-throughput gene expression technologies allowed to analyze simultaneously the expression of thousands of genes. Pathway analysis is often used to give a meaning to the set of differentially expressed genes. However, classical analyses generate a list of pathways that are over-represented or perturbed (depending on the approach used), but they do not consider, in many cases, the role of the connections between the biomolecules (genes or proteins) in the explanation of the biological phenomena studied. In this note we propose a fine-tuned method, based on Structural Equation Modeling principles, to discover pathway modules eventually able to characterize, in a network perspective, the mechanisms of the pathogenesis of a disease. The procedure relies on the concepts of shortest path, to find the initial modules, and of pathway composite variable, to improve and facilitate the interpretation of the modules proposed. The method was tested on microarray data of frontotemporal lobar degeneration with ubiquitinated inclusions.

Keywords

Gene expression data • Biological pathways • Structural equation modeling (SEM) • Principal component variables (PCVs) • K-core • Co-citation

D. Pepe (✉) • M. Grassi
Department of Brain and Behavioural Sciences, University of Pavia, Via Bassi 21—27100 Pavia, Italy
e-mail: danielepepe84@gmail.com; mario.grassi@unipv.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_22, Published online: 12 November 2014

1 Introduction

Recently a new way to consider pathological phenomena has taken hold. As reported by Barabasi et al. [1] "the phenotypic impact of a defect is not determined solely by the known function of the mutated gene, but also by the functions of components with which the gene and its products interact and of their interaction partners, i.e., by its network context." Therefore, the disease observed would be a consequence of the interdependencies between various perturbed processes that interact in a complex network. A useful concept to analyze microarray data, following the principle of Barabasi, is the one of biological pathways.

A biological pathway could be defined as a set of proteins and other biomolecules that interact with each other to perform a specific activity in an organism. We can have different biological pathways that involve gene regulation, metabolic processes, and signal transduction cascades. With the availability of biological pathway databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [2] or Biocarta (ww.biocarta.com), several resources have been developed to analyze gene expression data. Pathway analysis allows to give a meaning to the list of differential expressed genes (DEGs) found in microarray data in the context of biological networks. Many approaches exist as reported by Khatri et al. [3]. The Signaling Impact Analysis (SPIA) [4] is a Pathway Topology (PT)-based approach that permits also to consider the way in which the genes are connected between them, differently from other classical pathway analysis methods. SPIA allows to individuate biological perturbed pathways but it is not able to propose modules that could explain the phenotype observed. In this note we propose a procedure based on the graph-theoretic implementation of Structural Equation Modeling (SEM) [5], for testing and identifying perturbed pathway models. The selection of the proper model relies on classical network analysis as shortest-path and k-core, in the introduction of the concept of Pathway Composite Variables (PCV) and in use of co-citation analysis. In summary, in confirmatory SEM framework, it is necessary to start from a model. For each pathway a model was achieved finding the shortest paths between the DEGs connected by other microarray genes as in the approach used by Pepe et al. [6], grouping the not-DEGs using the concept of k-core in a unique variable (PCV), and finally, improving the model adding connections that result co-cited in literature.

2 Methods

All analyses below described were performed using samr, SPIA, graphite, igraph, lavaan, CoCiteStats packages of the statistical software R [7].

2.1 Differential Analysis

For the selection of DEGs, we used the Significance Analysis of Microarray (SAM) procedure. This is a statistical technique for finding significant genes in a set of microarray experiments by a set of gene-specific t-tests based on the permutations [8]. The cutoff for the significance is determined by a tuning parameter delta, chosen by the user and based on the false discovery rate (FDR) [9]. Furthermore, it is possible to choose a minimum fold change, i.e. the ratio between the gene expression level of the experimental group against the control group. The method was chosen for its reproducibility and reliability in the detection of microarray-derived lists of differentially expressed genes.

SPIA procedure has been used for the selection of perturbed pathways. It combines the evidence obtained from the classical enrichment analysis with the actual perturbation on a given pathway under a given condition. A global probability value (pG) is calculated for each pathway, incorporating parameters, such as the log fold-change of the DEGs, the statistical significance of the set of pathway genes and the topology of the signaling pathway.

2.2 Generation of Pathway Models by PCVs

A biological pathway can be considered as a directed graph $G = (Y, E)$ where Y is the list of genes or nodes and E is the list of edges that could represent reactions, regulations (activation/inhibition), and signals. Therefore, it is possible to see a pathway as a causal model, where the direction of edges represent the influence of a gene on another. Our approach consists in obtaining a model of the perturbed pathways starting from the DEGs. For model generation we tried to understand how a DEG communicates with another DEG. Therefore, we found the shortest paths between any couple of DEGs in the graph pathway, composed by other microarray genes. We indicated the microarray genes, DEGs, and not-DEGs in the following way: $MG = \{mg_1, mg_2, \ldots, mg_m\}$; $DEG = \{deg_1, deg_2, \ldots, deg_n\}$ and $NDEG = \{ndeg_1, ndeg_2, \ldots, ndeg_{m-n}\}$ where $MG = DEG \vee NDEG$ and $DEG \wedge NDEG = \{\emptyset\}$.

Each shortest path could be represented as a list of nodes $Y_k = (y_i, y_{i+1}, \ldots, y_{j-1}, y_j)$ and a list of the corresponding edges $E_k = (e_{i(i+1)}, \ldots, e_{(j-1)j})$, where $(y_i, y_j) \in DEG$; $(y_{i+1}, \ldots, y_{j-1} \in (DEG \vee NDEG)$; $Y_k \subseteq Y$ and $E_k \subseteq E$.

The shortest paths for each pathway constitute $(k = 1, \ldots, K)$ subgraphs $G_k = (Y_k, E_k)$ of the original pathway, $G = (Y, E)$. To obtain a more significant model we grouped the not-DEGs using the concept of coreness, a measure often used to identify the core proteome [10]. The k-core of a graph is the maximal subgraph in which each vertex has at least degree k, where the degree of a vertex is defined as the number of edges incident to the vertex. In other words, the k-core of a graph is defined as the unique subgraph obtained by keeping the nodes with a degree of k. The coreness of a vertex is k if it belongs to the k-core but not to the (k+1)-core [11].

To evaluate the goodness of the clustering we used the Database for Annotation, Visualization and Integrated Discovery (DAVID) [12, 13], an integrated biological knowledgebase and analytic tool aiming at systematically extracting biological meaning from large gene/protein lists. Submitting the list of genes for each k-core found, we looked for the biological processes in which the genes are involved. This is possible by Gene Ontology (GO), a controlled vocabulary composed of >38,000 precise defined phrases called GO terms that describes the molecular actions of gene products, the biological processes in which those actions occur and the cellular locations where they are present [14]. At each core was attributed the relative biological processes, but it would have been possible to choose also the cellular components or the molecular function. Each core identified, as previously described, was inserted in the module substituting the nodes and the edges relative to the genes that characterize the core.

2.3 SEM Analysis

As reported by Grace et al. [5], one of the main step in SEM framework is to indicate a causal diagram summarizing hypothesized causal connections among variables. In our case the causal models were represented by the pathway models previously defined, $G = (Y, E)$ embedded in structural equations of observed genes Y and covariance structure of unmeasured genes U. This can be written compactly in a matrix form as [15]:

$$Y = BY + U \quad \text{and} \quad \text{COV}(U) = \Psi$$

where B defines the path coefficients of directed edges between observed genes, and Ψ the covariance matrix of bi-directed edges between unmeasured genes.

To insert the cores in a SEM model, we generate the Pathway Composite Variables (PCVs) by Principal Component Analysis (PCA) on the set of genes belonging to each core. If the number of genes in the core set is much larger than the number of samples (p>>n), a sparse PCA (sPCA) is performed [16]. A PCV represents a group of genes (in our case not-DEGs that connect DEGs) in a pathway model obtained on the basis of the k core to which they belong. The biological identification of a PCV is possible looking for the more represented GO term associated to the genes that allowed to define the PCV. The principal component scores of the first principal component (PC1) are considered as values that characterize a PCV. Only PCVs for which the PC1 represents 50 % or more of the total variance are considered. This step allowed to improve and simplify the interpretability of the model generated. The module so obtained was estimated, evaluated and modified in SEM framework.

Briefly, the structural equations specification induces structure on the covariance matrix of the joint distribution of the genes Y as:

$$\Sigma(\theta) = (I - B)^{-1}\Psi(I - B)^{-T}$$

where $\theta = (\beta; \psi)$ is the list of the free parameters in the model of dimension t. The unknown parameters are estimated so that the implied covariance matrix $\Sigma(\theta)$ is close to the observed sample covariance matrix S by using the Maximum Likelihood Estimation (MLE) criterion.

For the evaluation of the fitting we used the Standardized Root-Mean-square Residual (SRMR), a measure based on the differences between observed values (s) and model values (σ) of the covariance matrix:

$$SRMR = \frac{\sum_{j=1}^{p} \sum_{k=1}^{p} (s_{jk} - \sigma_{jk})^2 / s_{jj} s_{kjk}}{p(p+1)/2}$$

SRMR values less than 0.10 are considered an adequate fitting approximation of the model to the data, whereas values <0.05 may be judged as a good fit.

The respecification of the model was based on the inclusion of additional directed or bi-directed connections. All the original edges in the model were considered true as the KEGG database is manually curated by experts. The criteria used for the refinement were based on the combination of modification indexes (MI, an estimate of the decrease in the χ^2-score statistic that would result by freeing each fixed (= 0) parameter in the model), z-tests (=parameter estimate/standard error) of the MLE, combined with co-citation analysis. In synthesis, a path coefficient is added when, it is proposed by MI and it is possible to reveal a connection based on the co-citation analysis. This analysis is based on the concept that two genes, cited in the same papers, are very likely connected to each other. However, one of the principal question about the co-citation analysis is if the co-occurrences in titles and abstracts actually reflect meaningful relationships between genes. We assumed that the answer to this question is "yes", also considering previous articles that treat this topic [17]. Different types of measures can be used to evaluate if two genes are connected [18]. We chose the Jaccard index, defined as the ratio between the articles (extracted from PubMed) where the two genes are co-cited together, divided by the union of the articles where the genes are cited together and singularly. A connection was considered acceptable if the Jaccard index, normalized for the number of genes, resulted greater than 0.5 [19]. The edges are added/deleted until a SRMR < 0.10 is reached. A multiple group analysis was performed on the final model to verify if it differed between the biological groups in the microarray. The test is based on the comparison of the fitted covariance matrix for each group, as reported below:

$$H_0 : \Sigma_1(\theta) = \Sigma_2(\theta) \text{ vs. } H_1 : \Sigma_1(\theta) \neq \Sigma_2(\theta)$$

subjected to $\mu_1 \neq \mu_2$. In the "null" model (H_0), the estimates of the covariances are constrained to be equal across groups; in the "alternative" model (H_1), the estimates covariances are allowed to differ across groups. Statistical significance is determined by comparison of LRT chi-square (χ^2diff) values of fit at given degree of freedom (d.f.diff). If there is significant difference ($p - value < 0.05$) in the chi-squared goodness-of-fit index between two models, it is possible to concluded that the groups differ significantly for one or more specific gene-gene relationships (edges).

2.4 Microarray Data

The data were relative to a microarray experiment that analyzes various brain regions of patients affected by frontotemporal lobar degeneration with ubiquitinated inclusions (FTLD-U) in presence of the mutation in the progranulin gene. Two groups were selected: one affected by FTLD with mutation in the progranulin gene (15 samples) and the other constituted by the control (17 samples). Data are freely available at Gene Expression Omnibus (GEO) [20] database with ID GSE13162.

3 Results

For our analysis we used normalized expression values submitted in the database. In the first step, SAM analysis was performed using a delta value of 1.03 and a minimum fold-change of 2. The number of over-expressed genes was of 207 while the number of down-regulated genes 244. Using this list of DEGs, the SPIA analysis found seven important pathways for the explanation of the role of the progranulin mutation on the FTLD-U, as showed in Table 1. Most of the dysregulated pathways, as the MPAK signalling pathway, the calcium signalling pathway, the gap junction, and the extracellular matrix (ECM)receptor interaction, confirm the analysis of Plotkin et al. [21]. To illustrate our method we chose the pathway of the ECM-receptor interaction. The ECM consists of a complex mixture of structural and functional macromolecules and serves an important role in tissue and organ morphogenesis and in the maintenance of cell and tissue structure and function. Specific interactions between cells and the ECM are mediated by transmembrane molecules, mainly integrins, a family of glycosylated, heterodimeric transmembrane adhesion receptors. In addition, integrins function as mechanoreceptors and provide a force-transmitting physical link between the ECM and the cytoskeleton. The role of ECM in the pathogenesis of dementia is well described in literature [22–24].

To generate the SEM causal model, the ECM-receptor interaction was transformed in a direct graph of 87 nodes and 651 edges. Then a marginalization was

Table 1 Pathways obtained by SPIA on FTLD-U data*

Name pathway	pSize	NDE	pNDE	tA	pPERT	pGFdr	Status
Glutamatergic synapse	77	11	0.000	−6.557	0.064	0.006	Inhibited
GABAergic synapse	60	10	0.000	0.632	0.804	0.017	Activated
Calcium signaling pathway	166	17	0.000	0.072	0.993	0.021	Activated
Amphetamine addiction	55	8	0.001	−2.685	0.457	0.047	Inhibited
Gap junction	85	10	0.001	5.216	0.454	0.047	Activated
MAPK signaling pathway	235	18	0.001	−5.802	0.253	0.047	Inhibited
ECM-receptor interaction	82	7	0.022	6.150	0.015	0.047	Activated

*pSize = number of genes in the pathway; NDE = number of DEGs in the pathway; pNDE = p-value of the enrichment; tA = total perturbation; pPERT = p-value of the perturbation; Status = direction of the perturbation; pGFdr = global p-value corrected by the Fdr

Table 2 Genes belonging to the set of shortest paths between DEGs

Entrez id	Official name	Gene name	Type of gene
22987	Sv2c	synaptic vesicle glycoprotein 2C	DEG
284217	LAMA1	laminin, alpha 1	Not-DEG
3371	TNC	tenascin C	DEG
9899	SV2B	synaptic vesicle glycoprotein 2B	DEG
9900	SV2A	synaptic vesicle glycoprotein 2A	DEG
960	CD44	CD44 molecule	Not-DEG
3909	LAMA3	laminin, alpha 3	Not-DEG
3685	Itgav	integrin, alpha V	Not-DEG
3675	ITGA3	integrin, alpha 3	Not-DEG
3688	Itgb1	integrin, beta 1	Not-DEG
6696	SPP1	secreted phosphoprotein 1	DEG
7058	thbs2	thrombospondin 2	DEG
6382	sdc1	syndecan 1	Not-DEG
1287	COL4A5	collagen, type IV, alpha 5	DEG
3912	LAMB1	laminin, beta 1	Not-DEG

performed, deleting all genes that did not belong to the microarray experiment, obtaining a new graph with 82 nodes and 567 edges. To understand how DEGs were connected between them, a subgraph was extracted by the union of all shortest paths between DEGs. The new graph was composed by 15 nodes and 39 edges. Table 2 contains the information about the 15 genes found. The creation of the PCVs via PCA was obtained by computing the coreness of the not-DEGs. The algorithm returned two cores: (1) one constituted by entrez id genes 3688, 3912, 3909, 3675, 284217, 3685; (2) the other constituted by entrez id genes 6382 and 960. To identify the cores, a search by DAVID on the database GO was performed. Using the information of the biological processes, the two cores were identified as cell adhesion and response to organic substance respectively. The variances explained by the first components for each PCV were of 54 and 93 % respectively. The new PCV model was composed by nine nodes and ten edges. The initial fitting of the SEM was poor (SRMR=0.203). The re-specification of the model was possible using the co-citation analysis, taking in consideration the value, corrected for the number of genes, of the Jaccard index. In this way three edges were added and the fitting indices of the new model resulted adequate (SRMR=0.079). Figure 1 represents the final model.

The two-group analysis of the final pathway was significant (χ^2 diff (df) = 44.7 (17), p - value < 0.001 of H_0: $\Sigma_1 = \Sigma_2$ subjected to $\mu_1 \neq \mu_2$). The analysis of gene-gene connections present in the model, revealed a strong presence of integrins incorporated in the PCV identified as cell adhesion. Functions for integrins and their ECM protein ligands are linked to neurovascular unit development, homeostasis and disease [25]. For example, down-regulation of the integrin protein beta 1 is correlated with elevated degradation of ECM proteins [26]. In addition, the role of

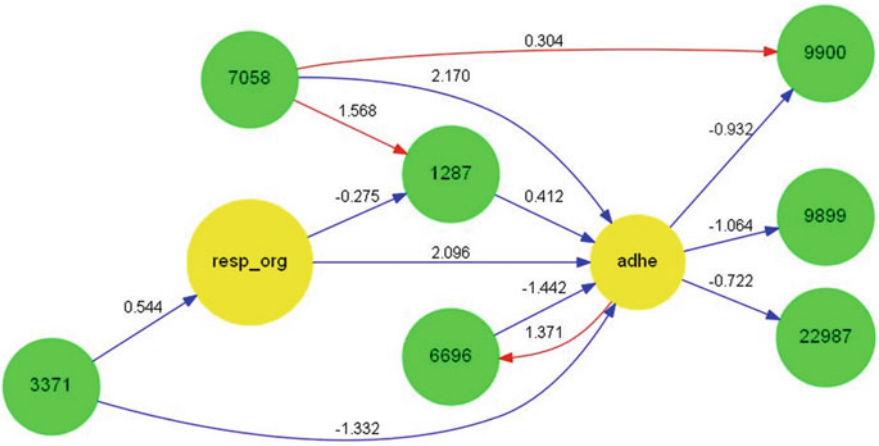

Fig. 1 Final PCV model for ECM pathway. *Green* nodes are DEGs, *yellow* nodes are the PCVs. *Blue* edges are KEGG edges while *red* were added after co-citation analysis. The name of genes are reported using the Entred notation

the proteins synaptic vesicle glycoproteins (SVs) is reported in neurological diseases [27]. Therefore, the model can be considered noteworthy both from a biological and statistical point of view.

4 Discussion

This note illustrates a new pipeline based on the SEM framework for the analysis of perturbed biological pathways. The procedure was applied to a microarray experiment that analyzes the effect of the progranulin gene on patients affected by FTLD-U. Starting from DEGs, a causal model for the ECM-receptor interaction pathway was generated. Then not-DEGs were grouped using the concept of coreness. The identity of each core was unveiled looking for the GO biological processes in which the core-genes are involved. A PCV for each core was created by PCA and then integrated in the model. In particular two PCVs were found: cell adhesion and response to organic substance. At the end of process, the model so obtained, was tested with SEM. The initial fitting was not good. For this reason, it was necessary to respecify the model. MI integrated with co-citation analysis allowed to add three edges, attaining a good fitting. Two-group analysis showed that the model differed significantly between affected and not-affected individuals, The goodness of the procedure was confirmed by the analysis of the genes in the model. For example, the strong presence of integrins, included in the cell adhesion PCV, is an important signal as they are associated to neurovascular unit development, homeostasis and diseases [24]. It is very important also the presence of SV genes, a possible marker for neurological diseases [27]. These results confirm the validity of

the procedure for the selection of the perturbed pathway modules also considering their interpretation from a biological point of view. The approach described above, although powerful, presents some limitations. In fact, the insertion of the PCVs in the pathway model brings to the loss of biological information about the connections between the genes involved in each PCVs. Furthermore, the clustering of not-DEGs using the concept of the k-core does not always lead to an easy biological interpretation of the PCVs generated. However, the procedure could represent a more general approach where the user could use different clustering methods based on statistical and/or topological principles [28]. In addition, the identification of the clusters could be adapted to the needs of the user. We used the GO terms, but it would have been also possible to use other biological evidences as the PIR superfamily [29] or the Disease Ontology (DO) [30] terms. In conclusion, we believe that our approach, based on the PCVs and SEM, represents a powerful and versatile tool for making perturbed models more interpretable and versatile.

References

1. Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nat. Rev. Genet. **12**, 56–68 (2011)
2. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**, 27–30 (2000)
3. Khatri, P., Sirota, M., Butte, A.: Ten years of pathway analysis: current approaches and outstanding challenges. Plos Comput. Biol. **8**, 15–22 (2012)
4. Tarca, A.L., Draghici, S., Khatri, P., et al.: A novel signaling pathway impact analysis for microarray experiments. Bioinformatics **25**, 75–82 (2009)
5. Grace, J.B., Schoolmaster, Jr., D.R., Guntenspergen, G.R., et al.: Guidelines for a graph-theoretic implementation of structural equation modeling. Ecosphere **3**, art73 (2012)
6. Pepe, D., Grassi, M.: Investigating perturbed pathway modules from gene expression data via structural equation models. BMC Bioinform. **15**, 132 (2014). doi:10.1186/1471-2105-15-132
7. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2012). ISBN 3-900051-07
8. Tusher, V.G., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA **98**, 5116–21 (2001)
9. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. **57**, 289–300 (1995)
10. Wuchty, S., Almaas, E.: Peeling the yeast protein network. Proteomics **5**, 444–449 (2005)
11. Haddadi, H., Rio, M., Iannaccone, G., et al. Network topologies: inference, modeling, and generation. Commun. Surv. Tutor. IEEE **10**, 48–69 (2008)
12. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. **4**, 44–57 (2009)
13. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. **37**, 1–13 (2009)
14. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. Nat. Genet. **25**, 25–29 (2000)
15. Shipley, B.: Cause and Correlation in Biology. A User's Guide to Path Analysis, Structural Equations, and Causal Inference. Cambridge University Press, New York (2004)
16. Zou, H., Hastie, T., Tibshurani, R.: Sparse principal component analysis. J. Comput. Graph. Stat. **15**(2), 262–286 (2006)

17. Stapley, B.J., Benoit, G.: Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac. Symp. Biocomput. **5**, 529–540 (2000)
18. Gentleman, R., Scholtens, D., Ding, B., et al.: Case studies using graphs on biological data. In: Bioinformatics and Computational Biology Solutions Using R and Bioconductor, pp. 369–394. Springer, New York (2005)
19. Ding, B., Gentleman, R.: CoCiteStats: Different Test Statistics Based on Co-Citation. R Package Version 1.32.0 (2013)
20. Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**, 207–10 (2002)
21. Chen-Plotkin, A.S., Geser, F., Plotkin, J.B., et al.: Variations in the progranulin gene affect global gene expression in frontotemporal lobar degeneration. Hum. Mol. Genet. **17**, 1349–1362 (2008)
22. Rauch, U.: Extracellular matrix components associated with remodeling processes in brain. Cell. Mol. Life Sci. **61**, 2031–2045 (2004)
23. Fillit, H., Leveugle, B.: Disorders of the extracellular matrix and the pathogenesis of senile dementia of the Alzheimer's type. Lab. Investig. **72**, 249 (1995)
24. Lukes, A., Mun-Bryce, S., Lukes, M., et al.: Extracellular matrix degradation by metalloproteinases and central nervous system diseases. Mol. Neurobiol. **19**, 267–284 (1999)
25. McCarty, J.H.: Integrin-mediated regulation of neurovascular development, physiology and disease. Cell Adh. Migr. **3**, 211–215 (2009)
26. Lee, S.R., Lo, E.H.: Induction of caspase-mediated cell death by matrix metalloproteinases in cerebral endothelial cells after hypoxia-reoxygenation. J. Cereb. Blood Flow Metab. **24**, 720–727 (2004)
27. Lassmann, H., Fischer, P., Jellinger, K.: Synaptic pathology of Alzheimer's disease. Ann. N.Y. Acad. Sci. **695**, 59–64 (1993)
28. Radicchi, F., Castellano, C., Cecconi, F., et al.: Defining and identifying communities in networks. Proc. Natl. Acad. Sci. USA **101**, 2658–2663 (2004)
29. Barker, W.C., Garavelli, J.S., Huang, H., et al.: The protein information resource (PIR). Nucleic Acids Res. **28**, 41–44 (2000)
30. Schriml, L.M., Arze, C., Nadendla, S., et al.: Disease ontology: a backbone for disease semantic integration. Nucleic Acids Res. **40**, D940–D946 (2012)

A Latent Growth Curve Analysis in Banking Customer Satisfaction

Caterina Liberati, Paolo Mariani, and Lucio Masserini

Abstract

Customer satisfaction for banking services is, arguably, a construct that develops and changes over time for a number of different endogenous and exogenous factors (modification of customers contract terms, transparency of banking transactions and financial services, bank charges, customer relationships, changes of market conditions an so on). Measuring change requires a longitudinal perspective: it can be carried out collecting measurements on the same individuals across multiple time points. The aim of this paper is to analyze the dynamic customer satisfaction of particular sub-groups of clients estimating a Latent Growth Curve Model. Results show interesting behavior differences to address strategic management decisions

Keywords

Customer satisfaction • Latent growth • Curve model • Banking services • Dynamic patterns • Latent variables

1 Introduction

Nowadays, every financial institution measures Customer Satisfaction (CS) with a high level of precision, in order to monitor client changing needs. It is known, in fact, that customer concerns and wishes change continuously. That induces businesses

C. Liberati • P. Mariani
Department of Economics Management and Statistics (DEMS), University of Milano-Bicocca, Piazza Ateneo Nuovo 1, Milan, Italy
e-mail: caterina.liberati@unimib.it; paolo.mariani@unimib.it

L. Masserini (✉)
Statistical Observatory, University of Pisa, Lungarno Pacinotti 43, Pisa, Italy
e-mail: l.masserini@adm.unipi.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_12, Published online: 12 November 2014

to monitor effectiveness of their marketing promotions, testing their customer satisfaction by surveys on field. The idea pursued is simple: satisfied customers tend to diffuse a positive image of the bank thereby reinforcing competitive strength. The marketing literature is replete with models on the measurement of customer perceptions of service quality [6, 18]. Studies have also examined the relationship of customer satisfaction and customer loyalty with service quality [1, 6, 7, 22]. Even in the banking sector, research has examined the impact of service quality on customer satisfaction and loyalty [10, 15]. Reference model of those studies are works of [17, 18]; which represents satisfaction as customer response to the perceived discrepancies between pre-consumption expectations and product/service effective performance [16]. According to such framework, marketing researches are based on cross-sectional surveys, because they are less costly to perform respect to the panel interviews, even though the information gained are incomplete. Despite, longitudinal studies on CS effects have found a positive relationship between customer retention, firm revenues, and share-holder value, few studies have been done on panel data: the impact on the share of wallet still remains elusive [5]. Generally speaking, longitudinal satisfaction data is hard to obtain even though a longitudinal view seems to be necessary. An analysis of time-series data, in fact, allows a firm to compare itself with itself over time, and provides useful in-sights about how customer perceptions of changes in service performance affect their global evaluations of service quality [4].

According to such remarks we performed a longitudinal analysis via a Latent Growth Curve Model (LGCM), focusing our attention on the dynamic aspect of CS. Our sample was collected in three different waves (T1, T2 and T3) interviewing 27,000 customers[1] each time. Due to privacy concerns, it was possible to analyze the overall data but it was not possible to obtain longitudinal information on individual customers interviewed (as in panel data). In order to overcome this limit, we performed an a priori segmentation based on employment status, educational qualification, gender and age. Segmentation is the act of defining meaningful sub groups of individuals or objects [20]. At its aim, it is reducing the number of entities being dealt with into a manageable number of groups that are mutually exclusive and share well defined characteristics. These approaches can be split into (1) a priori, the groups are selected from a population in advance based on known characteristics and declared as 'segments' (e.g. socio-demographic characteristics) and (2) post hoc, the empirical investigation through multivariate statistical analysis used to identify segments [8]. We found a coherent solution in terms of estimates and practical descriptions. The easiness of replication makes such an approach a valid alternative for segmentation analysis.

[1] Each wave has little more than an annual basis. Design of the sampling applied yearly, has selected subjects through a simple random sampling among retail clients that in a year have had contact with the bank at least five times, or, have had experience with the contact center and/or the bank's website. Therefore the sample obtained is not a cohort but a pseudo-panel.

2 Latent Growth Curve Model

During the last 30 years, growth curve modeling has become popular in the analysis of longitudinal and panel data [11, 12, 19] for the study of individual change. Growth curve analysis assumes individual growth patterns as represented by curves of the same functional form and with randomly varying parameters for describing differences in trajectories across individuals. The growth curve model can be approached from several perspectives. The Latent Growth Curve Modeling (LGCM) approach under the Structural Equation Modeling (SEM) framework adopts a latent variable view and assumes the existence of continuous underlying or latent trajectories for each individual. Growth trajectories are observed indirectly with the repeated measures [2] and individual differences both in the initial status and in the growth rate are included into the model as latent variables [13]. Latent variable means for the intercept and slope factors describe the mean growth whereas inter-individual differences in the parameters describing the growth curve are modeled as the (co)variances of the intercept and slope factors. Several benefits are associated with the use of LGCM over competing methods, such as the possibility of testing hypotheses about specific trajectories, the incorporation of both time-varying and time-invariant covariates and all the typical advantages of SEM, including the ability to evaluate the adequacy of models using model fit indices, the ability to account for measurement error by using latent repeated measures and the ability to deal effectively with missing data. The general latent growth curve model, for the repeated measure outcome variable y_i for individual $i = 1,\ldots, n$ observed at occasion $t = 1,\ldots, T$, may be expressed in matrix notation in terms of a confirmatory factor model, where the latent factors represent the latent curve components [3]:

$$\mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

Here, $\mathbf{y}_i$ is a $T \times 1$ vector containing the set of T repeated measures of the outcome variable y for individual i, $\boldsymbol{\eta}_i$ is an $m \times 1$ vector of latent factors, $\boldsymbol{\Lambda}$ is a $T \times m$ matrix of factor loadings and $\boldsymbol{\varepsilon}_i$ is $T \times 1$ vector of residuals. Elements of $\boldsymbol{\Lambda}$ are fixed to represent hypothesized trajectories, where each column of loadings represents a specific aspect of change. The conditional vector of latent variables can be expressed in terms of a mean and individual deviations from the means, as follows:

$$\boldsymbol{\eta}_i = \boldsymbol{\mu}_\eta + \boldsymbol{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i \tag{2}$$

where $\boldsymbol{\mu}_\eta$ is an $m \times 1$ vector of factor means, $\mathbf{x}_i$ is a $k \times 1$ vector of explanatory variables for the latent variables, $\boldsymbol{\Gamma}$ is $m \times k$ matrix of regression coefficients between the latent factors and the observed explanatory variables and $\boldsymbol{\zeta}_i$ is an $m \times 1$ vector of residuals. Finally, the variances and covariances of observed variables is contained in a $T \times T$ matrix, $\boldsymbol{\Sigma}$, and can be expressed as follows:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}^{'} + \boldsymbol{\Theta}_\varepsilon \tag{3}$$

where $\boldsymbol{\Psi}$ is an $m \times m$ covariance matrix of the equation errors term, $\boldsymbol{\zeta}_i$, among the latent trajectory factors and $\boldsymbol{\Theta}_\varepsilon$ is $T \times T$ covariance matrix of residuals, $\boldsymbol{\varepsilon}_i$.

3 Results

The managing board of an Italian bank, with a distribution network throughout the country, wanted to analyze its competitive positioning in retail services after a loss in the market share in some regions and a contraction of the average customer lifetime respect to the past. Therefore a survey has been conducted three times, sampling 27,000 retail customers each wave. The questionnaire was framed according to SERVQUAL model, therefore, with five dimensions to analyze perceived quality and expectation of the banking service. All the scores were measured by a Likert scale 1 to 10. A primary descriptive analysis showed a homogeneous distribution across different ages, sex, education and profession segments. This reflects the Italian 'banking population': more than 60 % of the sample is between 26 and 55 years old; the sample is equally distributed between genders and showed a medium low level of education. It is, also, well distributed across the different professional segments employees 24 %, pensioners 22 %, housewives 14 %. The Latent Growth Curve Model is estimated with M*plus* 5.21, using the Maximum Likelihood (ML) method with robust standard errors [14]. The final results are obtained after estimating several competing models. Selection is pursued firstly by considering the null model, which assumes no overall variability in the mean level of satisfaction and no change over time [21], thus by estimating only the intercept, μ_1, and a common disturbance variance (θ_ε). Because the null model is generally used only as a basis for comparison with more complicated models, inter-individual variability in latent growth factors (intercepts and slopes) is further evaluated and finally, customers-level covariates are introduced as far as they help in explaining intercepts and slopes variability. The first model evaluated is an unconditional linear random intercept (η_{1i}) and fixed slope model (η_{2i}), defined specifying the following equations for the vector of latent variables:

$$\eta_{1i} = \mu_1 + \zeta_{1i} \tag{4}$$

$$\eta_{2i} = \mu_2 \tag{5}$$

Here, μ_1 is the mean level of satisfaction at Time 1,ζ_{1i} represents individuals' deviations from the mean intercept whereas μ_2 is the constant mean slope. This model is based on the reasonable idea that customers have different level of satisfaction at Time 1 but the same growth rate. The significant intercept variance ($\Psi_{11} = 0.075$; $P < 0.001$) indicates the presence of intra-individual variability in customer satisfaction at the initial status, thus confirming our hypothesis.

Because the random intercept model does not provide an adequate fit ($\chi^2 = 354.190$; $df = 4$; $P < 0.001$), variability in the linear slope factor is investigated, in order to evaluate the hypothesis that also the rate of change in satisfaction varies

across customers. This model differs from the previous one only for specifying the slope variance parameter to be random, by introducing the residuals, ζ_{2i}, while the equation that defines the random intercept remains the same (thus not shown):

$$\eta_{2i} = \mu_2 + \zeta_{2i} \tag{6}$$

The results show a significant value both for the intercept ($\Psi_{11} = 0.107$; $P < 0.001$) and for the slope variance ($\Psi_{22} = 0.031$; $P < 0.001$), resulting in a better model fit ($\chi^2 = 235.6194$; $df = 1$; $P < 0.001$) so the hypothesis that customers satisfaction differs in both baseline level and growth rate seems to be more realistic. For this model, the estimated covariance between the intercept and slope growth factors is -0.027 (i.e., correlation of -0.461), indicating that the initial level of customer satisfaction is highly and negatively correlated with the rate of change. For evaluating whether the linear growth rate is adequate for these data, the model is further modified allowing for a more flexible estimation of the growth trajectory over time by setting the third factor loading free. Freely estimating the time scores allows the shape of the growth trajectory to be determined by data. This change produces a better model fit ($\chi^2 = 126.462$; $df = 1$; $P < 0.001$) and shows that the growth trajectory is not exactly linear. For this model the variances of the random intercept ($\Psi_{11} = 0.553$; $P < 0.001$) and slope ($\Psi_{22} = 0.483$; $P < 0.001$) are still significant whereas covariance of the intercept and slope growth factors is estimated to be -0.473 (i.e., correlation of -0.916).

Inter-individual differences in the growth curve factors (intercepts and slopes) are accounted for after introducing customers-level predictors variables into the random intercept and slope model with freely estimated time scores, with the following form for the vector of latent variables:

$$\eta_{1i} = \mu_1 + \boldsymbol{\gamma}_1 \mathbf{x}_i + \zeta_{1i} \tag{7}$$

$$\eta_{2i} = \mu_2 + \boldsymbol{\gamma}_2 \mathbf{x}_i + \zeta_{2i} \tag{8}$$

Here, $\mathbf{x}_i$ is the common vector of customers-level explanatory variables whereas $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are the vectors of the associated regression coefficients for the random intercept and slope equations, respectively. The mean level of satisfaction at Time 1 is $\mu_1 = 7.347$ whereas the slope is positive ($\mu_2 = 0.062$; $P < 0.001$) and shows a moderate increase in customer satisfaction over time. The covariance of the intercept and slope growth factors is -0.430 (i.e., correlation of -0.981), indicating that at Time 1 customers more satisfied than average tend to have a lower growth rate. This seems to reveal that customers more satisfied have a lower or even a negative growth rate in satisfaction than less satisfied customers. This is a typical tendency that happens 'if marketers raise expectations too high, the buyer is likely to be disappointed' [9]. The trajectories estimated from the conditional random effects model are graphically represented in Fig. 1 for the sub-groups of clients (Employment position, Education, Age groups and Gender).

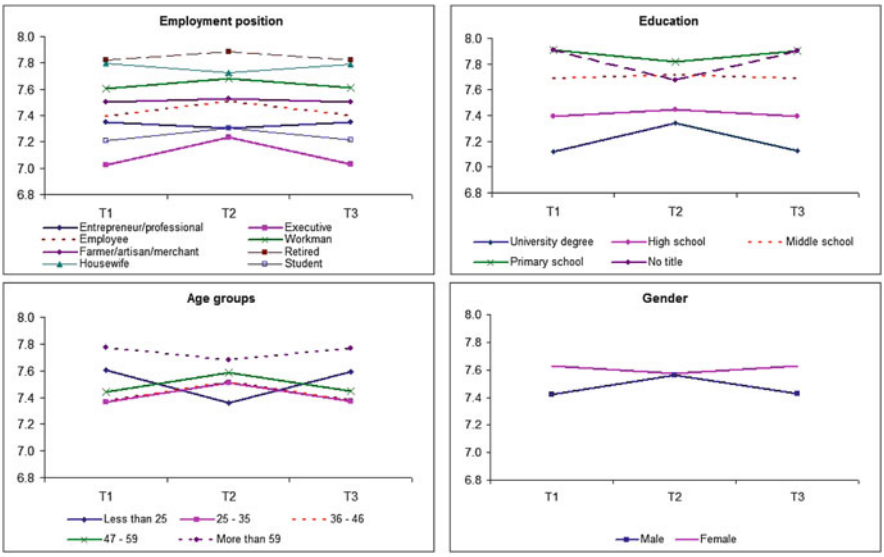

Fig. 1 Estimated trajectories from the conditional random effects model

Differences in baseline customer satisfaction are observed for some Employment position, Education, Age groups and for Gender. More specifically, about Employment position, the baseline satisfaction is significantly higher for Workman (+0.116), Retired (+0.220) and Housewives (+0.195) but lower for Entrepreneur/Professional (−0.109), Executives (−0.116) and Students (−0.282), compared to the Employee chosen as the reference category. As regard Education, the baseline satisfaction is significantly higher for Middle school (+0.213), Primary school (+0.465) and No title (+0.369) but lower for University degree (−0.243), compared to High school chosen as the reference category. Moreover, about Age groups, the baseline satisfaction is significantly higher for customers with age Less than 25 years (+0.184), 25–35 years (+0.029), 47–59 years (+0.032) and, above all, for customers with More than 59 years (+0.246), compared to 36–46 years chosen as the reference category. Finally, Males (−0.141) show a lower level of satisfaction compared to Females. As a consequence, the residual intercept variance not accounted for by the predictor variables is reduced but it still remains significant ($\Psi_{11} = 0.425$; $P < 0.001$). Also, differences in random slopes are observed in growth rate for some Employment position, Education, Age groups and for Gender. About Employment position, a higher rate of change is observed for Retired (+0.123) and Students (+0.178) whereas it is lower for Entrepreneur/Professional (−0.086), compared to the Employee chosen as the reference category. As regard Education, a positive effect on the slope is observed only for University degree (+0.155) whereas a negative effect is observed for Primary school (−0.207) and No title (−0.364), compared to High school chosen as the reference category. As regard Education, a positive effect on the slope is observed only for University

degree (+0.155) whereas a negative effect is observed for Primary school (−0.207) and No title (−0.364), compared to High school chosen as the reference category. Moreover, about Age groups, a negative effect is observed for customers with age Less than 25 years (−0.337) and with More than 59 years (−0.228), compared to 36–46 years chosen as the reference category. Finally, Males (+0.179) have a higher rate of change compared to Females. As a result, the residual slope variance not accounted for by the predictor variables is reduced but it still remains significant ($\Psi_{22} = 0.451$; $P < 0.001$). The final model provides an adequate overall fit ($\chi^2 = 951.775$; $df = 17$; $P < 0.001$), is characterized by low average residuals (RMSEA = 0.047; SMR = 0.012) and by a satisfactory incremental fit index (CLI = 0.920).

Conclusions

The dynamic analysis allowed us to monitor the satisfaction and expectations over time, illustrate the effects of some medium-term interventions, evaluate any changes to the strategy. The dynamic analyses are fundamental tool for planning, to assign the right targets at different branches thanks to the fact of identifying the strengths and weaknesses in service provision by the performance of the perceptions and evaluations over time. In our case the bank's management, conscious of the loss of competitiveness in some areas and on some types of customers (employees), tried to find an economic solution for the bank in terms of investments. The results show that, although the analyzed segments react differently to stimuli to which they are subject, there is a clear sign of the changing needs which, if not recognized by the bank, could make it hard the future growth or even make it probable a decline in economic performance.

References

1. Andreassen, T., Lindestad, B.: Customer loyalty and complex services: the impact of corporate image on quality, customer satisfaction and loyalty for customers with varying degrees of service expertise. Int. J. Serv. Ind. Manage. **9**(1), 7–23 (1998)
2. Bollen, K.: Latent variables in psychology and the social sciences. Annu. Rev. Psychol. **53**, 605–34 (2002)
3. Bollen, K.A., Curran, P.J.: Latent Curve Models. Wiley, New York (2006)
4. Bolton, R.N., Drew, J.H.: A longitudinal analysis of the impact of service changes on customer attitudes. J. Mark. **55**(1), 1–10 (1991)
5. Cooil, B., Keiningham, T., Aksoy, L., Hsu, M.: A longitudinal analysis of customer satisfaction and share of wallet: investigating the moderating effect of customer characteristics. J. Mark. **71**, 67–83 (2007)
6. Cronin, J., Taylor, S.: Measuring service quality: a reexamination and extension. J. Mark. **56**(3), 55–68 (1992)

7. Dabholkar, P., Shepherd, C.D., Thorpe, D.I.: A comprehensive framework for service quality: an investigation of critical conceptual and measurement issues through a longitudinal study. J. Retail. **76**(2), 67–83 (2000)
8. Green, P., Krieger, A.: Alternative approaches to cluster based market segmentation. J. Mark. Res. Soc. **37**(3), 221–239 (1995)
9. Kotler, P., Keller, K.: Marketing Management. Prentice Hall Upper Saddle Raver N.J. (2012)
10. Krepapa, A., Berthon, P., Webb, D., Pitt, L.: Mind the gap: an analysis of service provider versus customer perceptions of market orientation and the impact on 187 satisfaction. Eur. J. Mark. **37**(2), 197–218 (2003)
11. McArdle, J.J.: Latent variable growth within behavior genetic models. Behav. Genet. **16**, 163–200 (1986)
12. Meredith, W.M., Tisak, J.: Latent curve analysis. Psychometrika **55**, 107–122 (1990)
13. Muthén, B., Khoo, S.T.: Longitudinal studies of achievement growth using latent variable modeling. Learn. Individ. Differ. **10**, 73–101 (1998)
14. Muthén, L., Muthén, B.: M*plus* User's Guide, 6th edn. Muthén and Muthén, Los Angeles. http://statmodel.com/ugexcerpts.shtlml (1998–2012)
15. Ndubisi, N.O., Wah, C.K.: Factorial and discriminant analyses of the underpinnings of relationship marketing and customer satisfaction. Int. J. Bank Mark. **23**(7), 542–557 (2005)
16. Oliver, R., Winer, R.: A framework for the formation and structure of consumer expectations: review and propositions. J. Econ. Psychol. **8**, 469–499 (1987)
17. Parasuraman, A., Zeithaml, V.A., Berry, L.L.: A conceptual model of service quality and its implications for future research. J. Mark. **49**(4), 41–50 (1985)
18. Parasuraman, A., Zeithaml, V.A., Berry, L.L.: Servqual: a multiple-item scale for measuring consumer perceptions of service quality. J. Retailing **64**(1), 12–40 (1988)
19. Singer, J.D., Willett, J.B.: Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford University Press, New York (2003)
20. Wedel, M., Kamakura, W.: Market Segmentation: Conceptual and Methodological Foundations. Kluwer Academic, Dordrecht (1998)
21. Widamann, K.F., Thompson, J.S.: On specifying the null model for incremental fit indices in structural equation modeling. Psychol. Methods **8**(1), 16–37 (2003)
22. Zeithaml, V., Berry, L., Parasuraman, A.: The behavioral consequences of service quality. J. Mark. **60**(2), 31–46 (1996)

Non-Metric PLS Path Modeling: Integration into the Labour Market of Sapienza Graduates

Francesca Petrarca

Abstract

Non-Metric Partial Least Squares Path Modeling is a recent methodology based on the concept of Optimal Scaling applied to PLS Path Modeling algorithms. We adopted Non-Metric PLS Path Modeling to analyse a large administrative dataset containing nominal and ordinal variables using a specialized R package now available. We suggest a model in order to perform a preliminary quantitative study of the job success of Sapienza University of Rome graduates in terms of quality of work.

Keywords

Optimal scaling • PLS path modeling • Structural equation models • Categorical variables • Administrative archive • Economics education research

JEL classifications: C100, C390, A200.

1 Introduction

This paper has the following purposes:

1. to give a short presentation of the UNI.CO archive [2], which is the up-to-date complete administrative database of the integration into the Italian labour market of Sapienza graduates;
2. to study indicators of job success and to estimate their relationship with educational and job curricula.

F. Petrarca (✉)
Department of Economics, University of Roma Tre, Via Silvio D'Amico, 77, 00145, Rome, Italy
e-mail: francesca.petrarca@uniroma3.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_16, Published online: 12 November 2014

3. to model job success as a latent variable in PLS-PM framework;
4. to assess the effectiveness of Non-Metric approach [13] to Partial Least Square Path Modeling (PLS-PM) in the analysis of variables observed on different measurement scales.

The general framework is the study of the subordinate and para-subordinate employment offered to the Sapienza graduates by the Italian labour market. The aim is to extract from the information contained in the database, indicators that have an impact on the job position after graduating.

The goal is to define and measure the possibility of getting a good job position, i.e., satisfactory, well paid, stable over time, with possibility of improvements in career, consistent with university curriculum.

This study is based on the data of the UNI.CO archive[1] which contains the joint integration of the Sapienza graduates' archive and the Italian Ministry of Labour archive (known as *Compulsory Communication* (CO)). The integration of the two archives has produced a remarkable improvement in the quality of the information contained in the dataset with contributions entirely additional in respect of those provided singularly by each of them.

Measuring job qualifications is not an easy task, both in absolute or in relative terms [6]. In literature many indicators have been studied (e.g. index of job desirability [8], job quality index [12]).

In this paper we are interested in studying two new composite indicators [1] that are related to the possibility of the success in terms of Sapienza graduates best employment status.

In the International Standard Classification of Occupations (ISCO) a highly qualified position is identified with ISCO1 (managers) and ISCO2 (intellectual and scientific professions). Our indicators quantify the concept of job success using the definition of *optimal* and *quasi-optimal contract* based on the ISCO classification of job quality and on a minimum continuative duration of the job. The two indicators that we want to study are:

- *An optimal contract*: a contract that offers a permanent and highly qualified position (by ISCO Classification) with an actual duration of at least 8 months.

[1] The UNI.CO archive has been generated by an experimentation started in 2012 with the aim of establishing the integration of administrative archives. The results of the preliminary analysis of this new archive can be found in a first report (to be published) based on the Compulsory Communications archive for the study of labour demand for Sapienza graduates edited by the UNI.CO workgroup under the supervision of Giorgio Alleva. This group is composed of researchers coming from Sapienza University of Rome, Italian Ministry of labour and Italia Lavoro. For the Sapienza University: Pietro Lucisano, Carlo Magni, Silvia Massimi, Francesca Petrarca, Alessandro Sanzo, Bruno Sciarretta and Eleonora Renda. For the Italian Ministry of labour: Daniele Lunetta and Maurizio Sorcioni and for Italia Lavoro Giuseppe De Blasio. The workgroup was supported by a Scientific and Technical Committee of the Sapienza University composed of: Giorgio Alleva, Tiziana Catarci, Rosalba Natale and Cristiano Violani.

- *A quasi-Optimal contract*: a contract that offers a highly qualified position (by ISCO Classification) with an actual duration of at least 8 months.

The threshold of at least 8 months comes from D.lgs.181/2000 that considers in a status of unemployment workers with a job contract of less than 8 months.

The concept of a good job is rather theoretical and it needs a quantification in order to be inferred from data.

The class of the PLS methods [4, 5] seems to be the most suitable methodology to tackle this kind of problems because their capability to:

– quantify the latent variables (LVs) representing unobservable constructs;
– provide an estimate of the LVs for each observation;
– work without distributional hypotheses.

The last point is important because in the social sciences it is often the case that the distributions of the variables are asymmetric and very far from the Gaussian distribution.

Our suggestion is to measure the concept of good job defining it as a latent variable in PLS-PM framework. Our analysis is based on a dataset of variables which are observed on different measurement scales (numerical, ordinal and nominal). An interesting possibility to address this issue has been recently offered by a new procedure called Non-Metric Partial Least Squares (NM-PLS) [13] which is based on the implementation of the optimal scaling method applied to PLS algorithms. NM-PLS extends the applicability of PLS methods to data measured on different measurement scales, as well as to variables linked by non-linear relationships. A distinctive feature of these algorithms is that they provide a new metric both to non-metric and to metric variables. In this paper we adopted the term non-metric data to refer to ordinal and nominal variables.

The structure of this paper is as follows. In Sect. 2 we present a brief description of PLS-PM, the main feature of the NM-PLS and the assessment procedure adopted by this method. In Sect. 3 we discuss the dataset adopted and the model suggested. In Sect. refrisultati we discuss the results taking into account the assessment of the model and the quantification procedure. Section "Conclusions" draws conclusions.

2 PLS-PM

The Partial Least Square can be viewed as a set of methods for analysing multiple relationships between various blocks of variables. In particular the most common application of PLS-PM is the calculation of indices to quantify some key concepts or constructs called *latent variables (LVs)* that cannot be measured directly. One can analyse these concepts combining and summarizing a set of information that in some way reflect the meaning of the concept. The latent variables are indirectly measured by means of variables which can be observed/measured called *manifest variables (MVs)*. The manifest variables are divided in blocks which reflect to some extent the latent construct they are associated with.

The PLS methods are part of Structural Equation Models (SEM) [3, 10] that include a number of statistical methodologies meant to estimate a network of causal relationships, based on a theoretical model, linking two or more latent concepts, each measured by means of a number of observable indicators.

PLS-PM estimates the network of linear relations among the MVs and their own LVs, and among the LVs inside the model, through a system of inter-dependent equations based on simple and multiple regressions. The corresponding conceptual model can be represented by path diagrams where the LVs are represented by circles, the MVs by rectangles and the dependence relationships among the variables by arrows.

The difference between PLS-PM and SEM is that the first has been introduced as a component-based estimation procedure [16] and the second as confirmatory approach based on the estimation of the covariance matrix [9].

The PLS-PM is considered as a *soft-modeling* approach because it does not require strong assumptions with respect to the distributions, the sample size and the measurement scale. So the inferential approach is based on resampling technique that allow to obtain empirical distributions of the parameters. In the PLS-PM the outer weights, linking each MV to corresponding LV, are estimated by an iterative algorithm in which the latent variable scores are obtained through the alternation of the outer and inner estimations of the LVs. The PLS-PM consists of two sub-models:

- the structural model (or inner model) where the relationships among latent variables are established;
- the measurement model (or outer model) where the relationships between each latent variable and its block of manifest variables are established.

No formal proof of convergence of the general algorithm has been provided until now even though in some cases the PLS-PM loop is proven to converge monotonically, and the convergence is always reached in practice, for details on the convergence of the procedure refer to [7, 11].

The only two hypotheses underlying PLS models are:

- Each variable is measured on a interval (or ratio) scale;
- Relations between variables and latent constructs are linear and, consequently, monotone.

Therefore, standard PLS methods cannot handle data which are measured on a scale which does not have metric properties, nor non-linear relationships.

To overcome this problem a recent technique called Non-Metric Partial Least Squares (NM-PLS) algorithm has been set up, [13]. It consists in a new class of PLS algorithms that allow the PLS iteration to work as an optimal scaling algorithms, calculating iteratively both scaling and model parameters.

2.1 NM-PLSPM

The Non-Metric PLS Methods are so called thanks to their capability to provide optimally scaled data ($\hat{\mathbf{x}}$) with a new metric structure, which does not depend on the metric properties of the raw data ($\mathbf{x}^*$). In other words, NM-PLS methods yield a metric to non-metric data, and a new metric to metric data, linearizing the relations between variables and latent constructs, as required by the hypotheses of standard PLS models, [13].

The NM-PLS algorithms optimize criteria under two sets of parameters: the model parameters and the scaling parameters constrained to the restrictions due to the scaling level chosen for each raw variable $\mathbf{x}^*$. In the NM-PLS framework the quantifications are not determined by an external criterion but are obtained by the optimal quantifications method with respect to a latent construct called *Latent Criterion (LC)* which is represented by an unknown vector (centered by construction), for which we use the generic symbol $\boldsymbol{\gamma}_{x^*}$. For the NM-PLS, three levels of scaling are adopted according to measurement scale of the variables: nominal, ordinal and polynomial (or functional). A scaling (numeric) value [13] is assigned to each of the K categories (or distinct values) ϕ_k ($k = 1, \ldots, K$) of $\mathbf{x}^*$, such that:

- it is coherent with the chosen scaling level;
- it optimizes the model criterion.

In this way, each raw variable $\mathbf{x}^*$ is transformed as $\hat{\mathbf{x}} \propto \tilde{\mathbf{X}}\boldsymbol{\phi}$ where $\boldsymbol{\phi}' = (\phi_1, \ldots, \phi_K)$ is the vector of optimal scaling parameters. The matrices $\tilde{\mathbf{X}}$ are the indicator matrices of the different categories of variables and they define a space in which the constraints imposed by the scaling level are respected. For example, at nominal scale level grouping property is preserved while ordinal scale level preserves grouping and order properties. The symbol $\propto$ means that the left side of the equation corresponds to the right side normalized to unitary variance. The raw data $\mathbf{x}^*$ are transformed by different real functions (scaling functions) $Q(\mathbf{x}^*\boldsymbol{\phi}, \gamma_{x^*})$, one for each scaling level, which generate the optimal scaled value $\hat{\mathbf{x}}$ for each observations. The scaling functions Q optimize the criterion

$$\arg\max_{\phi} \operatorname{cor}^2 \left(\tilde{\mathbf{X}}\phi, \gamma_{x^*}\right)$$

under the constraints chosen for the $\mathbf{x}^*$.

The resulting scaling values for the different $\mathbf{x}^*$ are the least square regression coefficients of $\mathbf{X}$ on $\boldsymbol{\gamma}$ which correspond to the average of $\gamma_{\mathbf{x}^*}$ conditioned to $\mathbf{x}^*$ categories. The geometric representation of the scaled variable $\hat{\mathbf{x}}$, normalized to unitary variance, can be obtained projecting $\gamma_{\mathbf{x}^*}$ on the space defined by the columns of $\tilde{\mathbf{X}}$.

2.2 Assessment of the Model

In the PLS-PM frame, due to the fact that the model does not require distributional assumptions, the estimates of the parameter variability are obtained empirically by means of a bootstrap procedure. The validation of the quality of the model can also be studied by the evaluation of a few indicators that we briefly discuss in the following [4, 14].

For the measurement model the loadings represent the correlations between a latent variable and its indicators whereas the communalities are the squared correlations. They represent the amount of variability explained by a latent variable (e.g. a loading greater than 0.7 means that more than $0.7^2 \approx 50\,\%$ of the variability in an indicator is explained by its latent variable). Therefore a value around 0.7 or more is usually considered good for the loadings. The average communality ($Av.C$) represents how much of the block variability is reproducible by the latent variable and the average variance extracted (AVE) represents the amount of variance that a latent variable captures from its manifest variables in relation to the amount of variance due to measurement errors. A good value of AVE index is at least 0.50 which means that 50 % or more of the variance is accounted for.

For the structural model, the goodness of fit indexes taken into account are: the determination coefficients (R^2), the redundancy index and the average redundancy ($Av.R$). The R^2 represents the amount of variance in the endogenous latent variable explained by its independent latent variables. The redundancy index represents the amount of variance in an endogenous block of MVs explained by its independent latent variables. High redundancy means high ability to predict. The average redundancy represents the percentage of the variance in the endogenous block that is predicted from the independent LVs associated to the endogenous LV. This index and the R^2 index are available only for the endogenous construct.

An index that takes into account the model performance in both the measurement and structural model and thus provides a single measure for the overall prediction performance is the GoF that asses the goodness of fit of the whole model. GoF is calculated as the geometric mean of the average communality and the average R^2 value.

3 Dataset and Model

In this paper we take into account a sub-set of the UNI.CO archive: we consider only the master degree graduates of the Sapienza University who belong to the engineering disciplinary sector. Moreover we consider only graduates that subscribed more than one contract during the three years after graduation (458 statistical units).

In this preliminary study we propose a model in which the Job Success depends on the Educational and Job curricula. The set of manifest variables for each of the three latent variables representing Job Success, Educational Curriculum (Edu. Curr.) and Job Curriculum (Job Curr.) are described in Table 1. In the Job Success block are included as manifest variables only the two composite indicators: Optimal

and Quasi-Optimal. In our model all the manifest variables are treated as reflective i.e., the LVs are to be considered as the cause of the MVs belonging to its own block. We performed a Non-Metric PLSPM analysis on the model by using the option centroid for the inner weight estimation (this choice only considers the sign of the correlations between a LV and its adjacent LVs). As shown in Fig. 1 our model relates Job Success with Edu. Curr. and Job Curr. and also it analyses the relationship between Educational and Job curricula.

Table 1 Set of manifest variables for each latent variable

LVs	MVs	Description	Scale
Edu. Curr.	Age	Class of age at university graduation	Numerical
	Final grade	Final university grade	Numerical
	Average grade	Average graduation grade	Numerical
	Isee	Indicator of economic equivalent situation: it measures the economic status of the families	Ordinal (5)
Job Curr.	N_cn	Class of number of job relationships	Ordinal (13)
	gg_work	Class of number of worked days	Ordinal (7)
	gg_isco12	Class of number of worked days with high professional position	Ordinal (7)
	gg_al243	Class of number of worked days with an actual duration of the contract of at least 8 months	Ordinal (7)
	gg_CTI	Class of number of worked days with a permanent contract	Ordinal (7)
Job success	Optimal	The graduate has got optimal contract	Nominal
	Quasi-optimal	The graduate has got a quasi-optimal contract	Nominal

In brackets the number of levels for each ordinal variable are reported

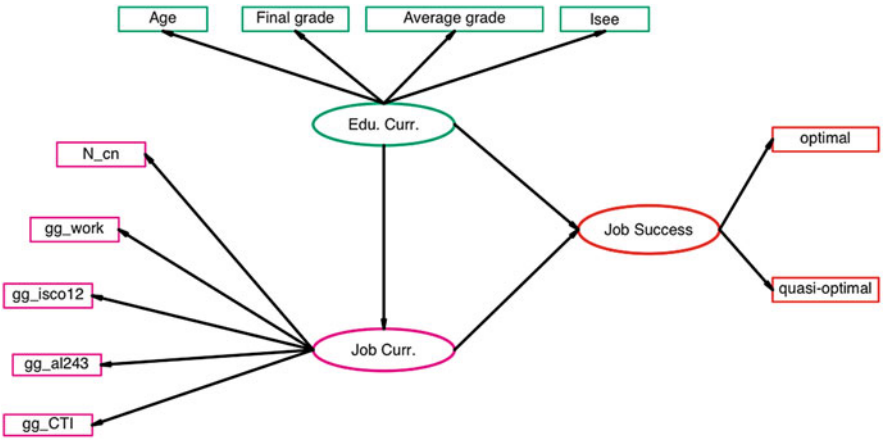

Fig. 1 Path diagram depicting our model

4 Discussion of the Results

We performed a NM-PLSPM analysis on the model described in Sect. 3 using a code written in the R language by Russolillo and described in [13], details about the PLS-PM in R are given in [14, 15]. A new improved version of the PLS-PM R package, containing the non metric extension, will be available shortly when the present test phase will be completed.

The iterative algorithm of the Partial Least Square Path Modeling separately estimates the several blocks of the measurement model and then, in a second step, estimates the structural model coefficients. In the Non-Metric PLSPM the standard PLS-PM procedure is combined with optimal scaling methods, during the cycles of iteration the model and the scaling parameters are alternately optimized in a modifies PLS-PM loop where the quantification phase is added.

In our case the convergence of the algorithm has been achieved after only nine cycles.

In Fig. 2 we report, for all the variables, the plots of the raw values versus the scaling values obtained at the end of the convergence of the iterative procedure. These plots show that all the non-metric manifest variables are properly quantified using monotone transformations of the quantitative MVs.

As matter of comparison, we have performed the basic analysis by PLS-PM (without optimal scaling) replacing all the categorical variables with dummy variables, produces in many cases a non monotonic relationship between the dummy and the LV score with a non clear interpretation of the results. In the following we report the results coming from the application of the NM-PLSPM to our model.

We start now to examine the results of the outer model. The values of the parameters of the outer model with corresponding 95 % confidence intervals built by means of 1,000 bootstrap samples are reported in Table 2. In the block of Edu. Curr. we have for all the MVs high loadings with the exception of the manifest variable Isee (0.29). Thus we could consider removing this variable from the model. Moreover the empirical validation of the model shows that this value is not significant. The block of Edu. Curr. is positively affected by all the its MVs with the exception of Age that is negatively correlated. This is a trivial fact because the older the graduates the less is the study success. In the block of Job Curriculum we find a similar situation to the one found in the previous block. Only for N_cn we have a very small values of the loadings (0.11). Also in this case the bootstrap procedure indicates a non significant value. We have checked that removing Isee and N_cn from the model, the GoF increases from 0.42 to 0.46. All the MVs in this block are positively correlated with the own LV with the exception of N_cn that is negative. In the block of Job Success we have high loadings for all the MVs.

The Optimal and Quasi Optimal indicators are discriminant to the construction of the Job Success block, see Table 2. In fact, the weights of the MVs quantified at nominal scaling, which reflect the variability of the corresponding LV explained by the categories of the MVs, have high values particularly in the case of the Quasi Optimal indicator.

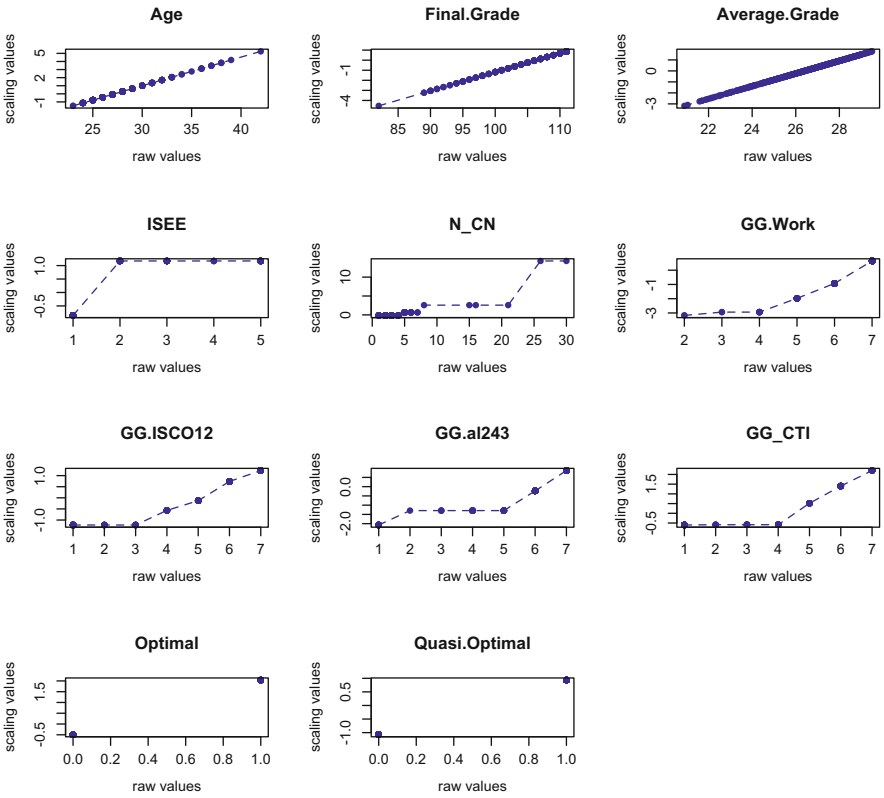

Fig. 2 Values of the original variable plotted versus corresponding optimal scaling values

Results of the structural model with corresponding 95 % confidence intervals built by means of 1,000 bootstrap samples are reported in ctab:3. The path coefficient from Edu. Curr. to Job Curr. is moderately small (0.30) indicating a feeble influence of educational curriculum on job experiences. In the case of the regression of Job Success in respect of Edu. Curr. and Job Curr. we see that, while the Job Curr. influences the Job Success very much (0.75), the Edu. Curr. has a small coupling with Job Success and a negative sign (−0.07). The bootstrap intervals for the path coefficient of Edu. Curr. to Job Success contain the value zero, so this coefficient is not significant a 5 % confidence level, see Table 3. It is also interesting that the indirect effect of Edu. Curr. to Job Success i. e. the path: Edu. Curr.−Job Curr.−Job Success, gives a positive contribution of 0.17 which is not negligible.

The results of this regression suggest to analyse a simpler inner model in which Edu. Curr. is linked with Job Curr. and Job Curr. with Job Success. This path follows the natural temporal sequence from Edu. Curr. to Job Curr. and then to Job Success of a standard student. We have checked this model and the results are substantially unchanged.

Table 2 Main results of the measurement (outer) model: the weights and loadings (λ) are shown

LVs	MVs	Weights	λ	Std. Error	perc.025	perc.975
Edu. Curr.						
	Age	−0.10	−0.80	0.12	−0.85	−0.70
	Final grade	0.11	0.93	0.13	0.88	0.96
	Average grade	0.12	0.94	0.13	0.89	0.96
	ISEE	0.04	0.29	0.16	−0.10	0.47
Job Curr.						
	N_CN	−0.06	−0.11	0.12	−0.28	0.14
	GG Work	0.23	0.78	0.03	0.70	0.82
	GG ISCO12	0.42	0.74	0.03	0.68	0.78
	GG al243	0.26	0.78	0.03	0.70	0.82
	GG_CTI	0.25	0.57	0.05	0.45	0.66
Job success						
	Optimal	0.27	0.82	0.03	0.76	0.86
	Quasi optimal	0.33	0.89	0.02	0.85	0.92

For the loadings the corresponding 95 % confidence intervals built by means of 1,000 bootstrap samples are reported

Table 3 Results of the structural (or inner) model with corresponding 95 % confidence intervals built by means of 1,000 bootstrap samples

Paths	R^2	β	Std. Error	perc.025	perc.975
Edu. Curr. → Job Curr.	0.09	0.30	0.04	0.21	0.37
Edu. Curr.→ Job success	0.54	−0.07	0.03	−1.08	0.02
Job Curr. → Job success		0.75	0.02	0.72	0.79

The R^2 and the path coefficients (β) are shown

In Table 3 we also reported the R^2 values of the endogenous latent variables for each regression in the structural model. We have $R^2 = 0.09$ for the regression where the endogenous variable is Job Curr. and a higher value $R^2 = 0.54$ in the case of the endogenous variable Job Success. The value 0.09 for the first R^2 is rather low but it is confirmed by the bootstrap procedure as well as the corresponding path coefficient. Moreover, it should be taken into account that high values of R^2 are not expected because our endogenous manifest variables (Optimal and Quasi Optimal) in the block of Job Success are binary and they are analysed together with nominal, ordinal and numerical variables. The values of the main goodness indices obtained from our model are reported in Table 4. The average redundancy for Job Success indicates that Edu. and Job Curricula predict 40 % of the variability Job Success indicators whereas the average redundancy for Job Curriculum indicates that Edu. Curriculum predicts lower value of 3 % of the variability of Job Curriculum. The AVE index shows good values for all our constructs except for Job Curriculum. Finally, we obtained that the whole prediction power of the model is GoF = 0.42.

Table 4 Results of the main indices for the evaluation of the model

LVs	Type	Av.C	Av.R	AVE
Edu. Curr.	Exogenous	0.62		0.62
Job Curr.	Endogenous	0.42	0.03	0.42
Job success	Endogenous	0.73	0.40	0.73
GoF	0.42			

Average Communalities (Av.C), Average Redundancy (Av.R), AVE and GoF are shown

Conclusions

We have presented one of the first statistical analysis based on the data of the UNI.CO archive which is the more complete administrative archive of the Sapienza graduates available to date.

In this study, the NM-PLS have demonstrated a great adaptability to handle a large dataset with numerical, nominal and ordinal variables therefore confirming that the NM-PLS approach makes the PLS methodology even more flexible. The manifest variables are properly quantified by the optimal scaling technique that is adopted in this new procedure and it is implemented in the new R package.

The model studied in this paper to the aim of analysing the job success, taking into account the fact that the database is large and that it contains non-metric variables, gives a satisfactory representation of the data variability.

The high values of the measurement model have confirmed that the Optimal and Quasi Optimal indicators are discriminant to the construction of the Job Success block. We have seen that two variables (Isee and C_cn) can be removed without reducing the capacity of the model to explain the variance and also the structural inner model can be reduced to a model with a simpler structure where the path among the LVs becomes Edu. Curr., Job Curr. and Job Success. We have found that the age of graduation influences negatively the final job success, therefore the early conclusion of the scholastic career positively affects the success in the labour market. The overall frame that arises from this study is that the scholastic path of Sapienza engineering graduates does not seem to have a great direct influence to the aim of getting a satisfactory job.

However the model adopted in this preliminary work, that it has been chosen for its simplicity, probably does not contain the needed flexibility to explain the large quantity of information contained in UNI.CO archive. A model with more complex structure capable to recognize new composite indicators of the job success is under study.

Acknowledgements I would like to warmly thank Prof. G. Alleva, J. Mortera, G. Russolillo, and S. Terzi for their invaluable help during my Ph.D. studies and the writing up of this paper; I would like to thank Prof. G. Saporta for the warm hospitality at the CNAM in Paris. I would like to thank Dr. G. Sanchez for discussions about the new R package.

References

1. Alleva, G., Petrarca, F.: New integration indicators for Sapienza graduates in the employee and para-subordinate labour market designed for the UNI.CO archive, (working paper)
2. Alleva, G., Petrarca, F., Renda, E., Lucisano, P., Magni, C.: Potenzialità della matrice UNI.CO per lo studio delle caratteristiche della domanda di lavoro dei laureati della Sapienza: primi risultati e possibili sviluppi. Workshop on"Monitoring of the dynamics of the professional graduates". Italia Lavoro (2012)
3. Bollen, K.A.: Structural Equations with Latent Variables. Wiley, New York (1989)
4. Esposito Vinzi, V., Chin, W., Henseler, J., Wang, H., (eds.): Handbook of Partial Least Squares (PLS): Concepts, Methods and Applications. Springer, Berlin/Heidelberg/New York (2010)
5. Esposito Vinzi, V., Russolillo, G.: Partial least squares algorithms and methods. WIREs Comput. Stat. **5**, 1–19 (2013). doi:10.1002/wics.1239
6. Fabbris, L. (ed.): Indicators of Higher Education Effectiveness. McGraw-Hill, Milano (2012)
7. Hanafi, M.: PLS path modelling: Computation of latent variables with the estimation mode B. Comput. Stat. **22**(2), 275–292 (2007)
8. Jencks, C., Perman, L., Rainwater, L.: What is a good job? A new measure of labor-market success. Am. J. Sociol. **93**(6), 1322–1357 (1988)
9. Jöreskog, K.: A General method for analysis of covariance structure. Biometrika **57**, 239–251 (1970)
10. Kaplan, D.: Structural Equation Modeling: Foundations and Extensions. Sage, Thousands Oaks/California (2000)
11. Krämer, N.: Analysis of high-dimensional data with partial least squares and boosting. Ph.d. thesis, Technischen Universität Berlin, Berlin (2007)
12. Leschke, J., Watt, A.: Job Quality in Europe. WP 2008-07, ETUI-REHUS, Brussels (2008)
13. Russolillo, G.: Non-metric partial least squares. Electronic J. Stat. **6**, 1641–1669 (2012)
14. Sanchez, G.: PLS Path modeling with R. http://www.gastonsanchez.com/PLS_Path_Modeling_with_R.pdf, http://www.gastonsanchez.com/pathdiagram.
15. Sanchez, G., Trinchera, L.: R package of PLSPM. http://cran.r-project.org/web/packages/plspm/index.html
16. Tenenhaus, M.: Component-based structural equation modelling. Total Q. Manag. Bus. Excell. **19**(7), 871–886 (2008)

Single-Indicator SEM with Measurement Error: Case of Klein I Model

Adam Sagan and Barbara Pawełek

Abstract

Structural models with latent variables are one of the dominant analytical approaches in social sciences. They constitute a combination of two types of models: confirmatory factor analysis and regression analysis. The aim of this study is to respecify Klein I model by taking into account the econometric and psychometric perspective in the construction of the structural model with latent variables. It involves the introduction latent variables and inclusion of measurement errors into the model. The authors used the SEM approach to estimate an econometric model with measurement errors, identities and constraints imposed on model parameters.

Keywords

Klein I model • Measurement error • Single-indicator latent variable

1 Introduction

The evaluation of measurement errors is inevitably related to the modelling of economic phenomena. It combines two parallel research traditions. The first is related to the econometric modelling of complex economic relationships, their causal interpretation, and the stability and equilibrium of economic systems. The second tradition is derived from the psychometric analysis with latent variables representing theoretical constructs and traits, where reliability assessment and measurement error is crucial for development of attitude scales.

A. Sagan (✉) • B. Pawełek
Cracow University of Economics, Rakowicka 27, 31-510 Cracow, Poland
e-mail: sagana@uek.krakow.pl; pawelekb@uek.krakow.pl

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_25, Published online: 12 November 2014

Combining these two traditions is associated on the one hand with a stronger focus on the problem of exogeneity and causality in psychometric modelling, on the other—with the necessity to take into account the measurement errors in economic and econometric models. Treatment economic constructs, as error—free manifest variables without explicit measurement errors specification, may lead to negatively biased regression coefficients and broadening the confidence intervals for the parameters.

The aim of the paper is to: (1) reformulate the Klein I model of US economy as model with latent variables with single indicators, and (2) estimate the (hypothetical) measurement errors of single-indicator economic latent variables.

It provides an opportunity to integrate psychometric and econometric tradition in estimation of economic models using standard SEM framework and software. Additionally, the respecification is aimed at using the SEM approach in estimating econometric models with measurement errors, identities and restrictions imposed on model parameters.

The structure of the paper is as follows: in the first part, the research traditions in use of latent variables in economic models are presented, linking the evolution of econometric and psychometric approaches in this field. The second part presents the problems of estimation and identification of latent variable models with unit loading indicators. The third part is devoted to the presentation of Klein I model on the basis of contemporary methods of its estimation. The fourth part presents the simulation study of estimation of measurement error of single-indicator latent variables and reformulation of Klein I model as a model with single-indicator latent variables with measurement error. In the simulation, three methods are used for estimation of error variances in measurement models: modification indices, two-step approach and AIC/BIC-based specification search. The last part presents a short discussion and suggestion for further research.

2 Latent Variables in Economic Models

In the econometric literature the term latent variable has various connotations. Historically it was introduced by Koopmans [16] referring to stochastic disturbances in simultaneous equation model. Kmenta [15] distinguishes three main classes of latent variables in econometric models: (1) variables for which exact measurements do not exist and are represented by error-contaminated substitutes, (2) unobservable variables that can be represented by proxies, and (3) variables that are intrinsically not measurable. From this point of view we can distinguish five types of unobservables in econometric modelling that can cause three sources of errors in the models: (1) disturbances (errors in equation) that represent latent causes that give rise of unexplained variance of endogenous variables, (2) latent response variables (errors in measurement) that are continuous perfect substitute for imperfect binary or ordinal indicators, (3) errors of measurement (errors in variables) that reflect unexplained variance of indicator by true latent variable that are defined as common variance of its indicators, (4) common factor latent

variable that reveals common variance of reflective indicators in true-score model of measurement, and (5) unobserved heterogeneity due to omitting contextual or group-level factors in the model (errors in population).

In econometric tradition, the first two latent variables are thoroughly discussed in the literature [1, 15, 24]. Disturbance term as latent variable is very important for endogeneity analysis and testing the assumption of uncorrelatedness of instrumental variables (as proxies) with disturbance term and therefore IV does not explain variance in the 2SLS residuals [3]. Latent response variables are key concepts in econometric discrete choice probit models in preference modelling, in the framework of random utility models [19]. Unobserved heterogeneity is also taken into account in economic behavior and microeconometric models of panel data [6]. The last two classes of latent variables were paid less attention in economic literature. Economic variables were regarded as error-free (without measurement error term) and as manifest variable (not intrinsically latent). Seminal works of Zellner [25] and Goldberger [7] introduced the concept of a latent variable in economic literature and opened space for integration psychometric and econometric perspectives in SEM.

Latent variable (η) with reflective indicators is identified as a common variance factor and consists of measurement building block of SEM as confirmatory factor analysis (CFA). Identification of CFA model is based on three-indicator rule that says that model is identified if latent variable is measured by at least three of its manifest indicators (assuming no correlation between latent variables). This model is sometimes defined as a multiple effect indicators model or a model with effect indicators (in the cause-effect relationship indicators are dependent (effect) variable that is caused by the latent variable).

In economic models latent variables with formative indicators are even more popular. However, the identification of a latent variable with formative indicators (composite latent variable) is more problematic. Generally, identification of a formative latent variable is based on 2+ rule that involves for identification another latent variable with reflective indicators or the same latent variable with reflective indicators (as in MIMIC rule).

SEM model combines two types of models: confirmatory factor analysis (measurement part) and regression/path analysis (structural part). It links psychometric tradition with econometric. The use of structural models in psychometrics and social studies involves mainly problems of estimation of random and systematic measurement errors of theoretical constructs and the associated issues of reliability and validity of scales. Confirmatory factor analysis allows for the construction of a composite reliability model and the use of numerous indicators of the so-called greatest lower bound (GLB) of accuracy in the analysis of the reliability of measurement. The structural part of the model related to analysis of the relationship between theoretical constructs (these models are also known as causal models) draws most on the achievements of econometrics [10]. Especially causal claims seem to be a controversial issue in economic application of SEM that are echoed in the framework of simultaneous equation models. The causality inference is rooted in many methodological traditions and approaches as endogeneity testing using

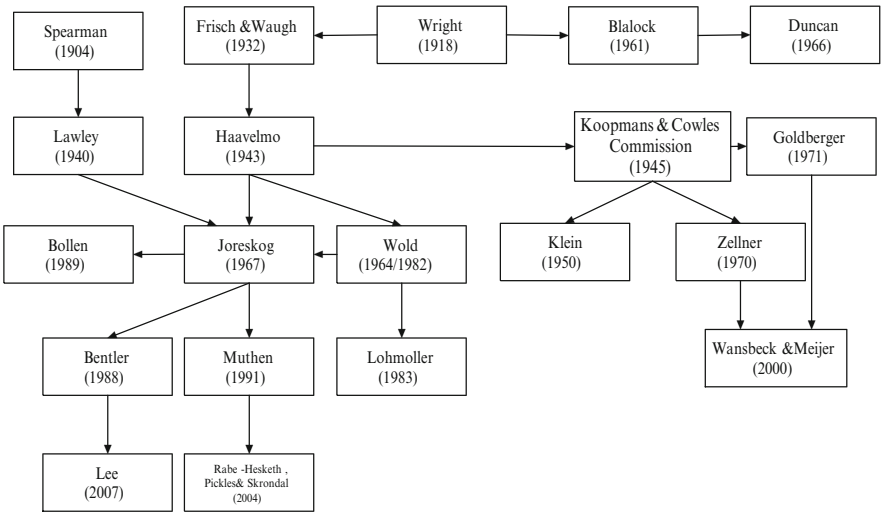

Fig. 1 Traditions in SEM modelling

2SLS and IV estimation along with Sargant and Bassman test [3], assessment of relevance of and control for background conditions, direct and indirect effects in moderational-mediational framework [4], building SEM models in the potential-outcome framework (counterfactual analysis) in latent difference model in quasi-experimental and true-experimental setting [23]. It deals with theoretical model specification problems [22] resulting from the models substantive error (error of approximation) or a statistical error (error of estimation) and understanding of the role of the relationship between the theoretical data-explaining model and the stochastic process generating data (DGP) and the assumptions associated with it (relating to distribution type, independence and heterogeneity) [11].

Identification of causal relationships in non-experimental research in these models is carried out on the basis of the principle of common cause and the causal Markov condition [21] by testing the effects of mediation and the conditional independence of variables and input of instrumental variables into the model [2]. In SEM, analysis of the potential (counterfactual) effects in experimental approaches is made usually by means of latent difference models [23]. Figure 1 displays the traditions in integrations of psychometric SEM (structural equation modelling) and econometric SEM (simultaneous equation modeling) in development of integrated SEM models.

3 Single-Indicator Latent Variables with Measurement Error

The special type of a latent variable in economic modelling is a latent variable with an error term measured by one indicator (single-indicator latent variable). The main problem in the use of single indicator measurement models is connected with the

estimation of measurement error variance. Single indicator models (with a reflective indicator) constitute a special type of models with latent variables. This type of model is less common in psychological or marketing research where multi-item scales are commonly used, but are popular in sociology and economics.

The introduction of single-indicator latent variables may be due to the following reasons: (1) the assumed lack of measurement error, (2) the assumption that the formative indicator fully determines the measured phenomenon (taking account of the unexplained variance of the latent variable) or it reflects it without measurement error, (3) parcelling of the indicators in hierarchical second-order SEM, (4) the use of summated rating scale, (5) a wish to obtain Thurstonian simple structure, (6) elimination of possible distractions from measurement part of model, (7) selecting the "single best indicator" of the latent variable.

Introducing the single-indicator construct into structural equations and estimation of the so called partially latent structural regression model [14] may result in several solutions. First, the attempt to estimate such a model with error measurement variance as a free parameter, which causes the problem with model identification. The second approach is to set the error variance as a fixed parameter based on prior knowledge or estimates. Third solution suggests making a wider range of simulations and estimation of the alternative models for testing the impact of the assumptions about measurement errors for the obtained solutions.

A single indicator measurement model assuming the lack of measurement error is identified using the ULI (unit loading indicator) approach. It means that factor loading is set as one and the error variance is zero.

Error variance estimation for single-indicator latent variables arises from the basic equation of the measurement model.

$$\sigma_x = \lambda^2 \times \theta + \delta_x, \tag{1}$$

where σ_x—indicator variance, λ—factor loading, θ—latent variable variance, δ_x—error variance. Using equation (1) one can point out that error variance is a function of indicator variance and measurement reliability:

$$\delta_x = (1 - \alpha) \times \sigma_x, \tag{2}$$

Reliability coefficient of latent variable has the following formula:

$$\alpha = \frac{\lambda^2 \times \theta}{\sigma_x}. \tag{3}$$

The structural equation model containing measurement errors for single indicator is shown in Fig. 2.

Latent variables representing measurement error d1 and e1 are also called errors-in-variables (EIV), and the variable z1 constituting the latent variable representing the disturbance associated with unexplained variance of the dependent variable is also known as an error-in-equations (EIE). Failure to take account of the

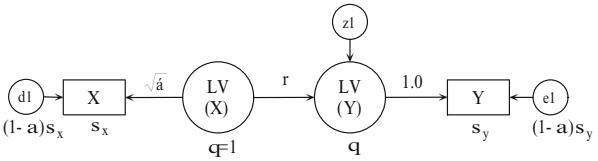

Fig. 2 Single-indicator latent variable with measurement error

measurement error in the model leads to getting biased regression parameters, in which the "true" values of the regression or correlation coefficients between latent variables are attenuated by the unreliability of the indicators. The bias of regression parameters results from a correlation of the random component with the independent variable. In addition, failure to account for the measurement error causes underestimation of the coefficient of determination and reduction in models explanatory power [16, p. 17].

4 Klein I Model with Latent Variables

An attempt to apply the SEM methodology to economic research began from the consideration of selected classical econometric models for macroeconomic phenomena. The model of the U.S. economy constructed by Klein in the forties (so-called Klein I model) was taken into account. Klein I model is often cited in works from the field of econometrics as an example of the simultaneous equations system. The analysis presented in this paper assumes the following formula of Klein I model:

$$C = \alpha_1 + \alpha_2 P + \alpha_3 P_{-1} + \alpha_4 W + \epsilon_1, \tag{4}$$

$$I = \alpha_5 + \alpha_6 P + \alpha_7 P_{-1} + \alpha_8 K_{-1} + \epsilon_2, \tag{5}$$

$$W_p = \alpha_9 + \alpha_{10} E + \alpha_{11} E_{-1} + \alpha_{12} t + \epsilon_3, \tag{6}$$

$$Y = C + I + G - T, \tag{7}$$

$$P = Y - W, \tag{8}$$

$$K = K_{-1} + I, \tag{9}$$

$$W = W_p + W_g, \tag{10}$$

$$E = Y + T - W_g, \tag{11}$$

where C—consumption, I—investments, W—total wage bill, W_p—private wage bill, W_g—government wage bill, P—profits, K—capital, E—private sector output,

Y—income, T—indirect taxes, t—time variable, G—government spending, ϵ_1, ϵ_2, ϵ_3—random components.

Thus formulated model (4)–(11) consists of three stochastic equations (4)–(6) and five identity equations (7)–(11). The variables include eight endogenous variables (C, I, W_p, Y, P, K, W, E) and seven pre-determined ones (Wg, G, T, t, P_{-1}, K_{-1}, E_{-1}). The Klein's model assumes the manifest variables only (without measurement errors), and it has been estimated by means of: full information maximum likelihood method (FIML), limited information maximum likelihood method (LIML), three-stage least squares method (3SLS) and two-stage least squares method (2SLS).

In order to estimate the values of variables C, I, G, P, Wp and Wg, Klein [13, pp. 135–141] used 31 macroeconomic variables. For example the consumption formula consists of following composites: [13, pp. 135–136]:

$$C = \frac{C_1 + C_2}{C_3}, \tag{12}$$

where C_1—consumer expenditures on goods and services (in billions of 1934 dollars), C_2—individuals' net imputed rent (in billions of 1934 dollars), C_3—average consumers' outlay.

Statistical data used by Klein to calculate the consumption level (one of the endogenous variables in Klein I model) in the United States in 1920–1941 were derived from source works [20, p. 873], [18, p. 735], [17, p. 145]. After performing calculations on the original data, more accurate estimations of consumption than those which are included in the work of Klein [13] were obtained. The results of calculations performed on the basis of original data differ by $+/-0.1$ for years 1927, 1929, 1932, 1933 and 1935 compared to data in Klein work [13, p. 135] subsequently used by many authors in their works (cf. e.g. [5, p. 10], [8, pp. 563–564], [9, p. 950]). After converting the original data as rounded by the Klein one obtains the same values for variable C as those published by Klein in 1950.

The differences noted between the considered sets of consumption values were one of the reasons for undertaking research, some results of which are presented in this paper. In our opinion these differences justify reviewing Klein I model I in the context of a measurement error.

5 Klein I Model Estimation with Observed Variables

Specification of Klein I model [13] in the area of structural modelling involves FIML (Full Information Maximum Likelihood) or FILGRV (Full Information Generalized Least Residual Variance) methods of estimation. The use of identity equations in the model results in the lack of identification of the model caused by the failure to satisfy the rank rule for the observed and implied covariance matrix.

In order to identify the structural model, observable endogenous variables in the model are treated as latent variables, of which only dependent latent variables asso-

Table 1 Behavioral equations estimation using SEM and GRETL packages

Parameters	ML Mplus	GLS-ML Statistica	LIML GRETL	FIML GRETL	3SLS GRETL
Int	16.68	–	17.148	18.343	16.441
P	0.162	0.370	−0.223	−0.232	0.125
P-1	0.221	0.330	0.396	0.386	0.163
W	0.725	0.870	0.823	0.802	0.790
Int	26.17	–	22.591	27.264	28.178
P	0.090	0.370	0.075	−0.801	−0.013
P-1	0.750	0.710	0.680	1.052	0.756
K-1	−0.176	0.168	−0.168	0.148	−0.195
Int	2.930	–	1.526	5.794	1.797
E	0.320	0.140	0.434	0.234	0.400
E-1	0.239	0.236	0.151	0.285	0.181
t	0.179	0.182	0.132	0.235	0.150

Own research based on estimation in Mplus 7.0, Statistica 10, and GRETL
Int intercept, *P* profits, *P-1* profits in t-1, *W* wages, *K-1* capital in t-1, *E* private sector output, *E-1* private sector output in t-1, *t* time

ciated with consumption (C), investment (I) and wages (Wp) are measured by unit loading indicator (ULI). In addition, due to the larger number of dependent latent variables with respect to manifest variables, starting values for the disturbances parameters were set by the researchers. Determining the initial values causes the implied covariance matrix to be positive definite in the first iteration steps in the Newton-Raphson algorithm. However, its use is associated with biased χ^2 statistics and standard errors (normality assumption of indicator distribution and small sample size). Additionally, the assumption of independence of observation in Klein I model is violated [12, pp. 164–170].

Table 1 presents the results of estimation of behavioral equations with the observed variables only. In order to compare the estimations, both SEM programs (Statistica and MPLUS) and econometric package GRETL were used. The parameters depicted in Table 1 shows that there exist minor differences with respect to strength of the parameters when statistical packages and methods of estimation are compared. The direction and significance of relationship is comparable across both SEM packages and GRETL solution.

The absence of error variances in original models (where only manifest variables are used), disabled the evaluation of measurement errors in the model. Therefore, Authors propose a simulation approach in the assessment of measurement error of single-indicator latent variables that reflect economic categories in the Klein I model.

6 Simulation of Measurement Errors for Single Indicators of Klein I Model

Estimation of the measurement error in Klein I model with single indicators of latent variables was carried out on the basis of three methods: (1) modification indices (MI) and expected parameter change (EPC), (2) two-step method (T-S), and (3) specification search (SS) minimizing the AIC and BIC information criteria. The results of comparative analysis are depicted in Table 2.

Table 2 shows the estimations of the variance of the measurement errors for the simulations (based on modification indices, in the two-step approach and for exploratory search for the best model specification). The MI and EPC suggest adding the error covariance between Wp and T indicators to diminish χ^2 statistic and improve the overall goodness of fit of the model. The negative sign of covariance suggests that it may be due to the omitted variable that has an opposite influence on both indicators.

In the two-step approach during the first step, a model with zero error variance for indicators (which is identical to the classical structural model without latent variables) was estimated, while during the second step, error variances were estimated for fixed parameters of the structural model (with constraints arising from the previously estimated model). This approach shows that constructs of private sector output (E-1) and capital (K-1) are measured with a significant amount of error. The third approach involved exploratory specification search for error variances for the indicators that minimize AIC and BIC information criteria. Figure 3 shows the result of AIC and BIC-based simulation of errors for E-1 indicator.

Due to the fact that in single indicator CFA the error variance is a function of factor loading, the simulation looked for such combinations of level of error terms

Table 2 Measurement errors of estimates

Variables	MI	T-S	SS
C	–	0.042	0.000
I	–	0.000	0.000
Wp	−0.539	0.000	0.000
P-1	–	0.000	0.000
K-1	–	0.237	0.200
E-1	–	0.591	0.600
Wg	–	0.000	0.000
T	−0.539	0.000	0.000
G	–	0.039	0.000

Own research based on estimation in Mplus 7.0

C consumption, *I* investment, *Wp* private wage bill, *P-1* profits in t-1, *K-1* capital in t-1, *E-1* private sector output in t-1, *Wg* government wage bill, *T* indirect taxes, *G* government spending

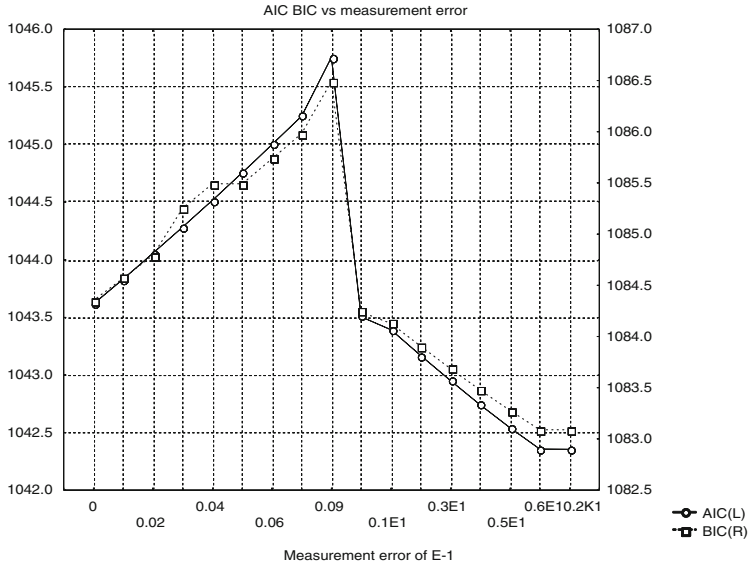

Fig. 3 AIC/BIC-based specification search

and loadings for the factor E-1 that minimize the overall AIC/BIC criterion in the model. Both criteria converge on the minimum with the level of error variances equal 0.6. The similar approach (not presented) was used for the specification search of measurement error for capital factor (K-1). The simulation confirms the result obtained by the two-step approach.

Comparing the models log-likelihoods it should be noted that the introduction of measurement errors for the lagged endogenous variables LK-1 and LE-1 (which turned out to be statistically significant) allowed an (only small) improvement in the fit of Klein I model (compared to the model with zero error variances) of 1 % (χ^2 statistic for "error-free" model was 215.785, compared to 214.523 in the model with measurement errors minimizing AIC).

On the basis of the simulated estimates, the respecified Klein I model with measurement errors was developed.

7 Application of Latent Variables in Klein I Model

Figure 4 depicted the maximum likelihood estimation of the single-indicator Klein I model with measurement errors within the framework of SEM graphical presentations.

Factor loadings and regression coefficients are marked as ($\leftarrow$) and identity equations as ($\twoheadrightarrow$). For example: $LY \equiv LG + LI - LT + LC$. Dotted arrows ($\dashrightarrow$)

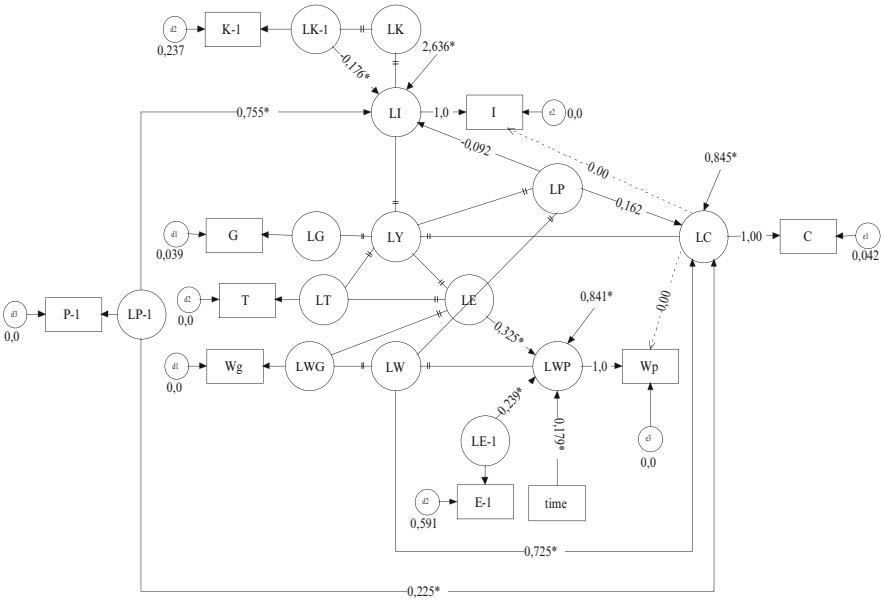

Fig. 4 Klein I model with latent variables and measurement errors. (*) Significant parameters ($p < 0.05$)

indicate the dummy paths introduced for identification purposes. The parameters of structural equations are similar to the original Klein model.

The large amount of measurement error for E-1 economic construct (private sector output) and relatively smaller for K-1 (capital) show potential measurement unreliability.

Concluding Remarks

The integration of psychometric and econometric traditions in modelling with latent variables is a promising way of development economic and social models with single indicators. The first, places special emphasis on the interpretation of economic constructs in terms of latent variables (measured by a single indicator or multiple indicators) with an explicitly stated measurement error. The second, is related to the methodologically correct specification of causal structural dependencies between constructs.

The results show that even in the relatively simple Klein I model the economic aggregates are contaminated with error terms and isolation of error may improve the model interpretation and obtain unbiased structural parameters and their standard errors.

It provides the way for further research on respecification of economic macro-models, which, without taking into account the quality of construct measurement, may mask insignificant relationships between variables.

References

1. Aigner, D.J., Hsiao, C., Kapteyn, A., Wansbeck T.: Latent Variable Models in Econometrics. Elsevier, Amsterdam (1984)
2. Angrist, J.D., Imbens, G.W., Rubin, D.B.: Identification of causal effects using instrumental variables (with discussion). J. Am. Stat. Assoc. **91**, 444–472 (1996)
3. Antonakis, J., Bendahan, S., Jacquart, P., Lalive, R.: On making causal claims: a review and recommendations. Lead. Quart. **21(6)**, 1086–1120 (2010)
4. Baron, R.M., Kenny, D.A.: The moderator mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J. Personal. Soc. Psychol. **51**, 1173–1182 (1986)
5. Bianchi, C., Calzolari, G., Corsi, P.: Alternative Estimates of the Klein I Model. IBM Scientific Center, Pisa (1981)
6. Cherry, T.L.: Unobserved heterogeneity bias when estimating the economic model of crime. Appl. Econ. Lett. **6(11)**, 753–757 (1999)
7. Goldberger, A.S.: Maximum-likelihood estimation of regressions containing unobservable independent variables. Inf. Econ. Rev. **13**, l–15 (1972)
8. Goldberger, A.S., Nagar, A.L., Odeh, H.S.: The covariance matrices of reduced-form coefficients and of forecasts for a structural econometric model. Econometrica **29**(4), 1–7 (1961)
9. Greene, W.H.: Econometric Analysis. Prentice Hall/Pearson Education Inc., Upper Saddle River (2003)
10. Heckman, J.J.: Causal parameters and policy analysis in economics: a twentieth century retrospective. Q. J. Econ. **115(1)**, 45–97 (2000)
11. Johansen, S.: Confronting the economic model with the data. In: Colander, D. (ed.) Post Walrasian Macroeconomics: Beyond the DSGE Model. Cambridge University Press, Cambridge (2007)
12. Joreskog, K., Sorbom, D.: LISREL 8: Users Reference Guide. SSI Scientific Software Inc., Lincolnwood (2001)
13. Klein, L.R.: Economic Fluctuations in the United States 1921–1941, Cowles Commission for Research in Economics, Monograph No. 11. Wiley/Chapman Hall, New York/London (1950)
14. Kline, R.: Principles and Practice of Structural Equation Modeling. Guilford Press, New York (2011)
15. Kmenta, J.: Latent variables in econometrics, Stat. Neerl. **45**(2), 73–88 (1991)
16. Koopmans, T.: Statistical estimation of simultaneous economic relations, J. Am. Stat. Assoc. **40**, 488–566 (1945)
17. Kuznets, S., Epstein, L., Jenks, E.: National income, aggregate payments, and consumers' outlay. In: Kuznets, S., Epstein, L., Jenks, E. (eds.) National Income and Its Composition, 1919–1938, vol. I. National Bureau of Economic Research (NBER) (1941). http://www.nber.org/chapters/c5540. Accessed 6 Feb 2012
18. Kuznets, S., Epstein, L., Jenks, E.: Basic data, sources and methods: finance. In: Kuznets, S., Epstein, L., Jenks, E. (eds.) National Income and Its Composition, 1919–1938, vol. II. National Bureau of Economic Research (NBER) (1946). http://www.nber.org/chapters/c5557. Accessed 6 Feb 2012
19. McFadden, D.L.: Econometric analysis of qualitative response models. In: Griliches, Z., Intriligator, M.D. (eds.) Handbook of Econometrics, pp. 1395–1457. Elsevier, Amsterdam (1984)
20. Painter, M.S.: Estimates of Gross National Product, 1919–1928 (On the basis of the Department of Commerce concept), Board of Governors of the Federal Reserve System (U.S.) Washington, Federal Reserve Bulletin, September 1945, vol. 31, No. 9 (2012). FRASER, http://fraser.stlouisfed.org/publication/pid=62. Accessed 6 Feb 2012
21. Pearl, J.: Causality, Cambridge University Press, Cambridge (2000)
22. Spanos, A.: Statistical Foundations of Econometric Modelling. Cambridge University Press, Cambridge (1986)

23. Steyer, R.: Analyzing Individual and Average Causal Effects via Structural Equation Models. Methodology **1**(1), 39–54 (2005)
24. Wansbeck, T., Meijer, E.: Measurement Error and Latent Variables in Econometrics. Nord Holland, Amsterdam (2000)
25. Zellner, A.: Estimation of regression relationships containing unobservable independent variables. Int. Econ. Rev. **11**, 441–454 (1970)

Investigating Stock Market Behavior Using a Multivariate Markov-Switching Approach

Giuseppe Cavaliere, Michele Costa, and Luca De Angelis

Abstract

By stressing the latent nature of expected return and risk, we develop a two-step procedure for obtaining new insights about the properties of financial returns. The first step consists in achieving a time-invariant classification of stocks into homogenous groups under the risk-return profile, thus providing innovative measures of expected return and risk. In the second step, we investigate the dynamic behavior of the stocks belonging to each group by using multivariate Markov-switching models. We find evidence of different dynamic features across groups of stocks and common dynamic properties within groups which can be exploited for both interpretative and predictive purposes.

Keywords

Latent variables • Markov-switching • Mixture models • Multivariate analysis • Risk-return profile

1 Introduction

Latent variables have been extensively used in both theoretical and empirical research and cover a wide range of academic and operational fields. However, despite the relevant progresses made in the last years, the usefulness of latent variables in financial studies is still largely unexplored.

By taking into account the unobservable (i.e., latent) nature of expected return and risk, in this paper we propose a two-step method for the analysis of stock

G. Cavaliere • M. Costa • L. De Angelis (✉)
Department of Statistical Sciences, University of Bologna, Via Belle Arti, 41, 40126 Bologna, Italy.
e-mail: giuseppe.cavaliere@unibo.it; michele.costa@unibo.it; l.deangelis@unibo.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_3, Published online: 28 October 2014

market behavior. Specifically, the first step consists in achieving a time-invariant classification of the stocks into homogenous groups under the (latent) risk-return profile. This is done by employing mixture models [12], which allow the classification of multivariate data following a model-based approach and the determination of M groups using the information provided by a set of observable indicators. One important feature of this step is that, by means of statistical methods, we are able to determine the number of mixture components, M, i.e., to define the number of groups of stocks characterized by different financial features [2].

The second step of our procedure aims at investigating the dynamic behavior of the stocks belonging to each group. Latent variables play a relevant role also in the analysis of time series dynamics. In particular, Markov-switching (MS) models, which are characterized by a (dynamic) latent variable component where an unobserved Markov process drives the data generating process, are particularly suitable for describing correlated data that exhibit distinct dynamic patterns during different time periods. Since the seminal work of Hamilton [9] many contributions and extensions have been developed for analyzing financial variables. Among many others, Rydén et al. [13], Haas et al. [8], Guidolin and Timmermann [7], Gallo and Otranto [5], Al-Anaswah and Wilfling [1], and De Angelis and Paas [3] show that MS models are useful for investigating regime switching of returns and volatilities in stock markets and are able to capture many stylized facts of return series. In particular, the MS representation allows us to (1) endogenously detect the various stock market regimes, represented by the K (latent) states of the unobservable Markov process, using model selection methods, (2) interpret the different regimes on the basis of the switching parameters, and (3) obtain the probabilities of switching from one regime to another. These achievements may provide a valuable help in the development of an early warning predictive system.

2 A Two-Step Procedure for Investigating Stock Market Behavior

In this section, we briefly outline our two-step procedure for (1) obtaining the different groups of stocks, i.e., portfolios, and (2) investigating their dynamic behavior using latent variable models. In the first step we resort to mixture models in order to make inference about the latent variable of interest and the subsequent classification of the stocks into a certain number of groups, M. Once obtaining the M groups, in the second step we focus on the analysis of the dynamics of the stock's return series within each of the M groups.

2.1 Definition of the Groups

Consider a set of N stocks and, for each stock, consider a set of Q indicators $z_{k,h}$, for $k = 1, \ldots, Q$ and $h = 1, \ldots, N$ based on the unconditional distribution of the returns. In order to obtain the stock's classification, we use a mixture model where

the probability (density) distribution of stock return h is given by:

$$f(z_h) = \sum_{y=1}^{M} \pi_y \prod_{k=1}^{Q} f(z_{k,h}|y) \tag{1}$$

where $y = 1, \ldots, M$ denotes the time-constant latent variable which is characterized by M mixture components and is assumed to explain all the relationships among the indicators. That is, the observed variables, z_k, are assumed to be independent conditionally on the groups. For each group, the term π_y denotes the (prior) probability of belonging to a given group, where $\sum_{y=1}^{M} \pi_y = 1$. The conditional distributions, $f(z_{k,h}|y)$, for $k = 1, \ldots, Q$, are assumed to be Gaussian with (conditional) mean $\mu_y(z_k)$ and variance $\sigma_y^2(z_k)$.[1] The parameters are estimated by maximizing the associated log-likelihood function using the iterative procedure of the EM algorithm. The stock's classification is achieved using the Bayes' theorem, thus according to the modal rule: $\arg\max_{y=1,\ldots,M} h(y|z_h)$, where $h(y|z_h)$ denotes the posterior probabilities $h(y|z_h) = \pi_y f(z_h|y)/f(z_h)$.

Model selection is a well-known open issue in mixture modeling since there is no commonly accepted indicator for choosing the number of mixture components. For instance, information criteria, such as the Bayesian information criterion (BIC), are shown to consistently select the number of components, but they tend to underestimate M in small samples. Hence, we decide to rely on the Akaike information criterion (AIC) which, besides being a widespread and easy to compute procedure, also has the known tendency to never underestimate M. Within our two-step approach, dealing with more groups considerably eases the analysis at both interpretative and computational levels. In particular, selecting more groups tends to reduce the misclassification error and facilitates the analysis of the dynamic behavior at the second step since low-dimensional multivariate MS models are easier to estimate and their results are easier to interpret.

2.2 Analysis of the Dynamic Behavior

For $t = 1, \ldots, T$, we consider the m_y-dimensional vector of returns r_t, where m_y is the number of stocks classified into a specific group y, $y = 1, \ldots, M$. Furthermore, we consider a vector of explanatory variables x_t, which can be divided into two sub-vectors $x_t = [x'_{1,t}, x'_{2,t}]'$, where $x_{1,t}$ and $x_{2,t}$ are the q_1 switching and q_2 non-switching exogenous regressors, respectively.

Thus, a general MS model specification for the vector of returns r_t belonging to group y is given by:

[1] The normality assumption in mixture models is well established since the Gaussian distribution can be used as a cluster shape prototype, given that return distributions can be closely approximated by a Gaussian mixture.

$$r_t = \mu_{S_t} + \sum_{i=1}^{q_1} \beta_{i,S_t} x_{1,i,t} + \sum_{j=1}^{q_2} \beta_j x_{2,j,t} + \varepsilon_t \tag{2}$$

where $S_t = 1, \ldots, K$ denotes the discrete latent process governed by a Markov chain with K regimes and μ_{S_t} is the vector of m_y switching intercepts. In the following, the innovations ε_t are assumed to be Gaussian distributed with zero mean and switching covariance matrix Σ_{S_t}, i.e., $\varepsilon_t \sim N(0, \Sigma_{S_t})$. Nevertheless, the innovations may have a more general probability density function. For instance, we may assume that ε_t is either distributed as a Student-t with ν_{S_t} degrees of freedom, i.e., $\varepsilon_t \sim t(0, \Sigma_{S_t}, \nu_{S_t})$, or a Generalized Error Distribution with κ_{S_t} degrees of freedom, i.e., $\varepsilon_t \sim GED(0, \Sigma_{S_t}, \kappa_{S_t})$.

The specification in (2) allows for different MS models. For instance, when x_t includes only the lagged values of r_t, we obtain the MS VAR model by Krolzig [11]; or, when $x_t = \varnothing$, the MS model simplifies to the hidden Markov model (see, e.g., [13]).

3 Empirical Analysis

We apply the two-step procedure outlined in Sect. 2 to analyze the dynamic behavior of the $N = 30$ stocks included in the Dow Jones (DJIA) index at the end of 2012. We analyze the monthly return series from March 1993 to December 2012. The results are obtained using the Latent GOLD 4.5 Syntax module [14].

3.1 Definition of the Groups

In the first step of the procedure, we consider $Q = 6$ indicators, z_k, based on the unconditional distribution of the returns. We proxy the expected return-risk profile by using the following six indicators: mean return, mean return during crises, standard deviation, standard deviation during crises, 90th percentile, and 5th percentile. The crises are endogenously detected using two conditions: (1) the return of the DJIA index is below its 5th percentile, (2) the standard deviation of the return distribution of the DJIA index computed on the period is greater than 1.5 times the standard deviation of the DJIA index return series. The detected crisis periods are the following: June–September/2001, April/2002–February/2003, May/2008–February/2009, May–August/2010, and May–September/2011.

According to the AIC, a Gaussian mixture model with $M = 6$ components is detected, thus indicating the presence of six groups of stocks (i.e., portfolios). The interpretation of the (latent) risk-return profile of each group can be obtained on the basis of the estimated conditional means for each indicator. A summary of the results and the corresponding risk-return interpretations of the six groups are reported in Table 1. In particular, as reported below Table 1, we are able to detect the portfolio with the lowest level of risk (group 4), the portfolio characterized by the lowest

Table 1 Results for the Gaussian mixture model with $M = 6$ components and risk-return interpretation of the groups (Step I)

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Group size	0.070	0.231	0.231	0.231	0.134	0.102
Indicators						
Mean	0.022	0.591	0.638	0.759	1.017	1.500
StdDev	11.296	6.877	8.201	5.827	10.990	9.154
VPerc	−15.704	−10.801	−12.947	−8.617	−17.284	−12.843
ICPerc	11.532	8.518	10.020	7.533	14.098	12.290
Mean_{Crisis}	−10.017	−3.441	−5.458	−2.235	−5.577	−3.307
StdDev_{Crisis}	17.861	8.476	10.287	5.655	12.438	9.367

Interpretation of the groups:

- Group 1: lowest expected return and highest level of risk
- Group 2: moderate expected return and low level of risk
- Group 3: moderate expected return and moderate level of risk
- Group 4: lowest level of risk
- Group 5: high expected return-high risk profile
- Group 6: highest expected return and moderate level of risk

expected return and the highest risk levels (group 1), and the portfolio with the highest expected return profile (group 6).

3.2 Analysis of the Dynamic Behavior

By considering the return series of the stocks classified into the six different groups, in the second step of the procedure we evaluate the dynamic behavior within each group using multivariate MS models specified in (2).[2]

In Tables 2, 3, and 4 we show the results for groups 5, 1, and 4, respectively, which are particularly interesting from a risk profile viewpoint.[3] Specifically, groups 1 and 5 are both characterized by the highest risk profiles but different expected return levels, whereas group 4 is characterized by the lowest level of risk. The results from Tables 2, 3, and 4, where regimes are ordered ascending according to the values of the switching intercepts, μ_{S_t}, show evidence of relevant differences across groups: specifically, (1) different number of regimes and switching and non-switching regressors, (2) diverse interpretation of the regime profiles, and (3) unsynchronized regime dynamics.

[2]Note that following a two-step approach implies adding misclassification error to the analysis. However, for the Gaussian mixture model with $M = 6$ components, the misclassification error at the first step is small, namely 0.002.

[3]The results for the other groups are not reported due to space constraints but are available from the authors upon request.

Table 2 Results for the MS model for group 5 (high return-high risk profile)
$r_t = \mu_{S_t} + \beta_1 pb_{t-1} + \beta_2 r_{t-1} + \varepsilon_t,\ \ \varepsilon_t \sim N(0, \Sigma_{S_t}),\ \ S_t = 1, \ldots, 5$
Stocks: *HP*: Hewlett-Packard; *JPM*: JP Morgan Chase & Co.; *INT*: Intel; *CIS*: Cisco Systems

		$S_t = 1$	$S_t = 2$	$S_t = 3$	$S_t = 4$	$S_t = 5$
HP	μ_{S_t}	-7.833^{***}	-4.317^{***}	0.637	7.848^{***}	6.656^{***}
	β_1	-1.233^{***}				
	β_2	-0.103^{**}				
JPM	μ_{S_t}	-7.239^{**}	-0.473	-0.755	7.673^{***}	8.386^{***}
	β_1	-3.646^{***}				
	β_2	-0.110^{***}				
INT	μ_{S_t}	-6.630^{**}	2.892^{*}	-1.003	-0.740	14.212^{***}
	β_1	-1.235^{***}				
	β_2	-0.024				
CIS	μ_{S_t}	-10.215^{***}	3.710^{**}	-0.225	4.981^{**}	12.588^{***}
	β_1	-0.367^{***}				
	β_2	0.020				
Σ_{S_t} †	σ^2_{HP,S_t}	175.00^{***}	58.73^{***}	41.69^{***}	95.91^{***}	160.96^{***}
	σ^2_{JPM,S_t}	258.38^{***}	8.73^{***}	44.55^{***}	71.41^{***}	29.33^{***}
	σ^2_{INT,S_t}	200.37^{***}	90.58^{***}	44.12^{***}	236.16^{***}	47.19^{***}
	σ^2_{CIS,S_t}	232.27^{***}	94.09^{***}	65.57^{***}	105.55^{***}	41.35^{***}
P	$S_{t-1} = 1$	0.612^{***}	0.078	0.002	0.002	0.307^{***}
	$S_{t-1} = 2$	0.080	0.508^{***}	0.002	0.246^{***}	0.164^{**}
	$S_{t-1} = 3$	0.044^{**}	0.000	0.948^{***}	0.001	0.007
	$S_{t-1} = 4$	0.002	0.521^{***}	0.002	0.474^{***}	0.002
	$S_{t-1} = 5$	0.178^{**}	0.004	0.186^{**}	0.201^{***}	0.431^{***}

*Significant at 10 %; **significant at 5 %; ***significant at 1 %;
†Covariances σ_{jj',S_t} not reported for space constraints

First, we consider the two groups of stocks characterized by high levels of risk, namely groups 1 and 5.

Table 2 reports the results for group 5 which includes $m_{y=5} = 4$ stocks: Hewlett–Packard (HP), JP Morgan Chase & Co. (JPM), Intel (INT), and Cisco Systems (CIS). For this group, we estimate a MS model with $K = 5$ latent states with $x_{1,t} = \varnothing$ (no switching regressors) and $x_{2,t} = \{pb_{t-1}, r_{t-1}\}$ (two non-switching regressors: the price/book value ratio at time $t-1$ and the first order autoregressive component).[4] The estimated values of μ_{S_t} and Σ_{S_t}, as well as the

[4]In the second step of the procedure, we select the best model according to different methods. In particular, as in the first step, we consider the AIC for determining the number of latent states K. Then, we also consider the significance levels of the regression coefficients to decide (1) whether a variable should be included or not in the MS model and (2) whether it should be included as switching or non-switching regressor. Furthermore, regressor selection and the overall fitting of the model is also evaluated using likelihood ratio tests [6] and Lagrange multiplier-type tests for omitted autocorrelation and omitted regressors [10].

Table 3 Results for the MS model for group 1 (lowest expected return and highest level of risk): $r_t = \mu_{S_t} + \beta_{1,(S_t)} pb_{t-1} + \beta_{2,(S_t)} r_{t-1} + \beta_{3,(S_t)} dp_{t-1} + \varepsilon_t$, $\varepsilon_t \sim N(0, \Sigma_{S_t})$, $S_t = 1, \ldots, 4$ Stocks: *ALC*: Alcoa; *BoA*: Bank of America

		$S_t = 1$	$S_t = 2$	$S_t = 3$	$S_t = 4$
ALC	μ_{S_t}	−11.55	−1.316	−0.475	2.139**
	β_1	−1.801**			
	β_2	−0.009			
	β_3	8.587			
BoA	μ_{S_t}	−21.69***	−4.623**	0.021	2.883***
	β_{1,S_t}	1.539	−15.88***	5.389***	−5.801***
	β_{2,S_t}	0.324***	−0.368***	−0.486***	0.135*
	β_{3,S_t}	−80.88***	61.42***	43.06***	−21.96**
Σ_{S_t}	σ^2_{ALC,S_t}	826.83**	175.63***	77.82***	48.03***
	σ^2_{BoA,S_t}	48.68**	67.61***	15.90***	47.85***
	σ_{ALC,BoA,S_t}	185.25**	−14.80	21.95***	14.75**
P	$S_{t-1} = 1$	0.388***	0.335**	0.064	0.213
	$S_{t-1} = 2$	0.081	0.740***	0.003	0.176**
	$S_{t-1} = 3$	0.024	0.026	0.695***	0.255***
	$S_{t-1} = 4$	0.030	0.003	0.272***	0.695***

*Significant at 10 %; **significant at 5 %; ***significant at 1 %

transition probabilities in matrix P in Table 2, allow us to interpret the features of the five latent states, i.e., the characteristics of the five regimes. In particular, $S_t = 1$ represents the *crisis* regime since is characterized by large negative intercepts, $\mu_{S_t=1}$, and the highest conditional variances, $\sigma^2_{S_t=1}$. On the other hand, state 5 is a *bull* regime. The transition probabilities in matrix P show that, conditionally at being in state 1 at time $t-1$, at time t the latent process may remain in the crisis regime with a probability of 0.612, or may switch directly to the bull regime with a probability of 0.307. The third latent state, $S_t = 3$, is characterized by small conditional means and relatively low variances. Its persistence probability, i.e., the probability of staying in the same regime from time $t-1$ to time t, is close to 0.95, which corresponds to an expected regime duration of approximately 19.4 months. Hence, this state can be interpreted as a (long) *lateral phase*. The states 2 and 4 are somehow connected since the latent Markov chain often visits these two states sequentially, see the transition probabilities in matrix P.

Table 3 shows the results for group 1, which includes $m_{y=1} = 2$ stocks, namely Alcoa (ALC) and Bank of America (BoA). For this group, the best MS model is characterized by $K = 4$ regimes and the presence of mixed switching and non-switching regressors, which are denoted by $\beta_{i,(S_t)}$ (see Table 3), where dp_{t-1} denotes the dividend-price ratio at time $t-1$. As can be noticed from the results reported in Table 3, in the first equation of the system (ALC) all the included regressors are non-switching and only the price/book value ratio regressor is found significant at a 5 % level. Conversely, the second equation in the system (BoA) requires all regressors to be switching. As for the interpretation, we observe that the

Table 4 Results for the MS model for group 4 (lowest risk profile)
$r_t = \mu_{S_t} + \beta_{1,(S_t)} pb_{t-1} + \beta_{2,(S_t)} r_{t-1} + \beta_{3,(S_t)} dp_{t-1} + \varepsilon_t,\ \varepsilon_t \sim N(0, \Sigma_{S_t}),\ S_t = 1, \ldots, 5$
Stocks: *3M*: 3M; *CHEV*: Chevron; *COC*: Coca Cola; *EXX*: Exxon Mobil; *J&J*: Johnson & Johnson; *P&G*: Procter & Gamble; *WMT*: Wal Mart Stores

		$S_t = 1$	$S_t = 2$	$S_t = 3$	$S_t = 4$	$S_t = 5$
3M	μ_{S_t}	−2.904***	−1.227	1.396***	2.516**	2.251**
	β_1	−0.156				
	β_2	−0.118**				
	β_3	14.91				
CHEV	μ_{S_t}	−0.549	−0.443	0.541	2.037***	2.518*
	β_1	−2.497***				
	β_{2,S_t}	−0.142	0.237***	−0.241***	−0.090	−0.316***
	β_3	11.87**				
COC	μ_{S_t}	−6.421***	−1.080	0.925**	4.830***	8.681***
	β_1	0.146***				
	β_2	−0.047				
	β_3	34.14***				
EXX	μ_{S_t}	0.019	−0.037	0.283	2.334***	4.806***
	β_{1,S_t}	−1.114	−0.510	−1.551***	−1.859**	−5.193***
	β_2	−0.060				
	β_3	−1.715				
J&J	μ_{S_t}	−3.525***	−2.832***	0.920**	5.070***	11.004***
	β_1	0.098				
	β_{2,S_t}	−0.073***	0.642***	−0.216***	−0.399***	−0.351***
	β_{3,S_t}	7.602	−35.24***	31.38***	57.90**	48.18***
P&G	μ_{S_t}	−5.804***	−1.276**	0.710*	5.622***	9.929***
	β_1	−0.002				
	β_2	−0.144***				
	β_3	19.83***				
WMT	μ_{S_t}	−2.452	−2.300***	1.104**	2.756***	6.908**
	β_1	0.315*				
	β_2	−0.117**				
	β_3	39.49**				
$\Sigma^{\dagger}_{S_t}$	σ^2_{3M,S_t}	30.08***	47.57***	17.59***	61.80***	12.46**
	σ^2_{CHEV,S_t}	53.66***	26.11***	27.94***	12.64***	25.77**
	σ^2_{COC,S_t}	56.74***	15.73***	12.27***	19.65***	39.90**
	σ^2_{EXX,S_t}	25.22***	43.38***	16.64***	14.85***	12.61**
	$\sigma^2_{J\&J,S_t}$	61.30***	2.99***	10.35***	6.69***	3.08**
	$\sigma^2_{P\&G,S_t}$	69.01***	10.55***	10.70***	6.74***	15.45**
	σ^2_{WMT,S_t}	87.90***	8.40***	23.78***	36.31***	88.80**
P	$S_{t-1} = 1$	0.528***	0.026	0.212***	0.056	0.178***
	$S_{t-1} = 2$	0.002	0.124*	0.792***	0.002	0.080*
	$S_{t-1} = 3$	0.061**	0.195***	0.540***	0.203***	0.001
	$S_{t-1} = 4$	0.206***	0.212***	0.299***	0.234***	0.049
	$S_{t-1} = 5$	0.170	0.003	0.069	0.754***	0.003

*Significant at 10 %; **significant at 5 %; ***significant at 1 %;
†Covariances σ_{jj',S_t} not reported for space constraints

coefficients β_{1,S_t} for states 2 and 4 are negative, as expected from a theoretical point of view: high (low) values of pb indicate a possible overvaluation (undervaluation) of the company which may lead to a decrease (increase) in expected returns. On the other hand, state 3 is characterized by $\beta_{1,S_t=3} > 0$, whereas $\beta_{1,S_t=1}$ is not significant. Therefore, during regime $S_t = 3$ that can be interpreted as a *lateral market phase*, the operators seem to expect that BoA stock price will continue to rise despite the overvaluation of the company. In other words, the stock is overvalued in operators' expectations and this mechanism may generate speculative bubbles. The coefficients β_{3,S_t} are expected to be positive since an increase (decrease) in the dividend yield (w.r.t. the stock price) usually leads to an increase (decrease) in the expected return. However, the results in Table 3 show negative values for states 1 and 4 which can be interpreted as the *crisis* and the *positive* regimes, respectively. Therefore, during the "most extreme" regimes, operators seem to expect less dividend yield at time t than the yield distributed at time $t-1$. This effect is understandable during *crisis* market phases but is quite surprising during *positive* phases. The coefficients of the autoregressive component, β_{2,S_t}, are negative for states 2 and 3 and positive for states 1 and 4. Thus, we find evidence of a positive autocorrelation during the *crisis* and *positive* market phases and the alternation of positive and negative returns during *lateral* phases represented by states 2 and 3. The transition probabilities in matrix P show a relative high level of regime-persistence: the probabilities of staying in the same regime from time $t-1$ to time t are close to 0.70 for all states, except for state 1, which has a probability of 0.388. Moreover, the significant off-diagonal values show that the latent Markov chain tends to switch to a neighboring state, except in the case of $S_{t-1} = 2$ where it may switch directly to $S_t = 4$ with probability of 0.176.

Finally, in Table 4, we analyze the results for the $m_{y=4} = 7$ stocks classified into group 4, which is characterized by the lowest level of risk. The best model is the MS model with $K = 5$ regimes which includes both switching and non-switching regressors (see Table 4). According to the estimates of μ_{S_t}, Σ_{S_t} and P, we can interpret the five different regimes. Specifically, we interpret state 1 as the *crisis/negative* regime, $S_t = 2$ as a *slightly negative* phase which precedes the *lateral* phase represented by state 3, while states 4 and 5 denote the *positive* phase and the *bull* regime, respectively. The $\beta_{i,(S_t)}$ coefficients can be easily interpreted as done above. The signs of the estimates of these coefficients are all consistent with the financial theory, except for β_1 in the COC equation, $\beta_{3,S_t=2}$ in the J&J equation and β_1 in the WMT equation, although the latter is significant only at a 10 % level.

The endogenous detection of financial market regimes and the estimation of regime-switching probabilities may substantially help monitoring the financial system and developing early warning predictive systems. As an example of the potential usefulness of this two-step approach in monitoring and disclosing financial crises, consider the estimated (smoothing) probabilities for the *crisis* regimes, i.e., the probability of being in the crisis state 1 at time t conditional on the observed return series, $\text{Prob}(S_t = 1|r_1, \ldots, r_T)$. In Fig. 1, we show these probabilities for the three MS models considered in Tables 2, 3, and 4 with respect to the period 2007–2012. As can be observed from Fig. 1, the group with the lowest risk profile (group

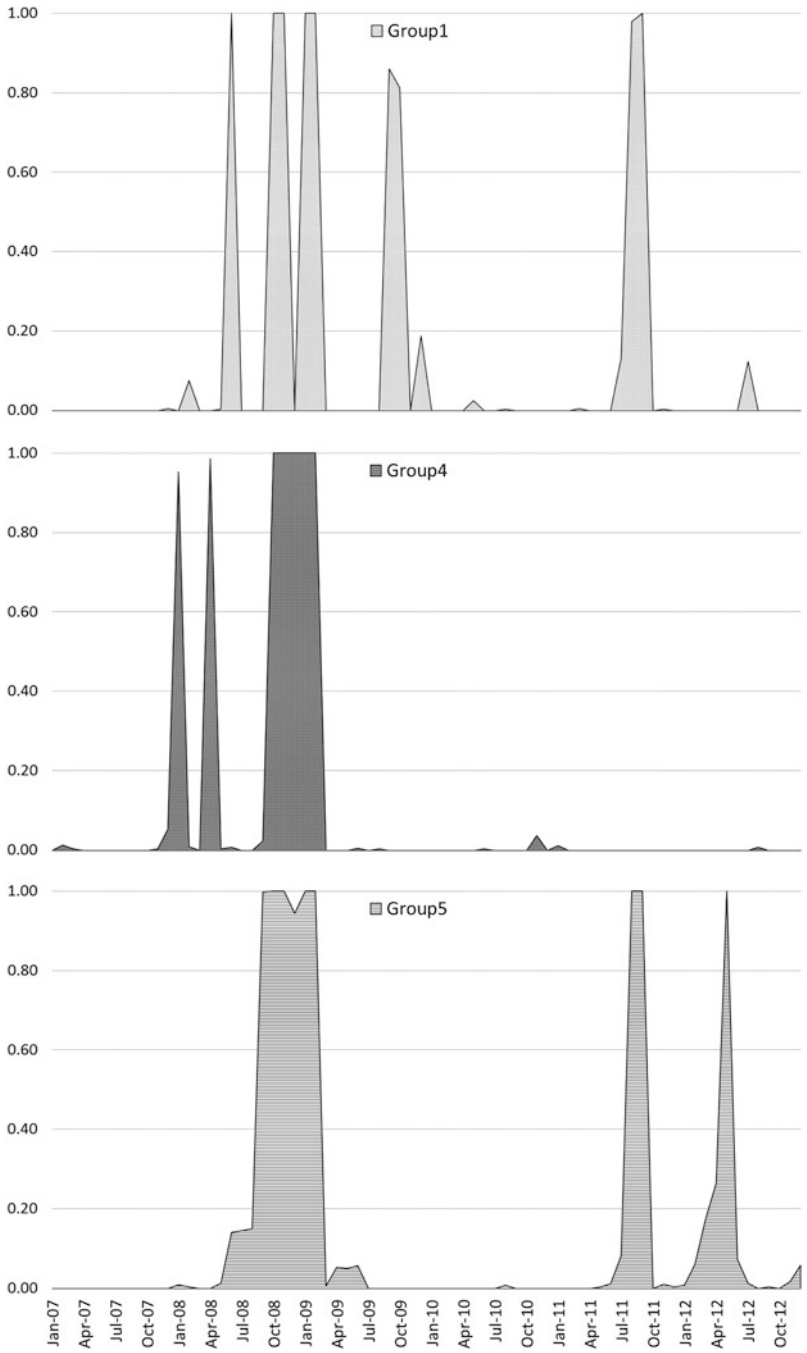

Fig. 1 Estimated (smoothing) probabilities for the *crisis* regime in the period 2007–2012

4) can be used to promptly detect the beginning of the so-called "sub-prime crisis" started at the end of 2007.[5]

Conclusions and Future Developments

In this paper we propose a two-step procedure involving latent variables for analyzing the distribution of stock returns. The aim of the first step is to obtain a time-invariant classification of the stocks which share a similar expected return-risk profile. In the second step we employ multivariate MS models to investigate the dynamic features of the different groups. By analyzing the monthly return distribution of the DJIA stocks, we find evidence of six groups characterized by different risk-return profiles. The analysis of the groups' dynamic behavior shows the presence of common dynamics within groups and contrasting behavior across groups, thus confirming the usefulness of the two-step procedure proposed in this paper. Indeed, the two-step approach provides much more flexibility than a single-step procedure based on a mixture of MS models, since allows the estimation of MS models which can be characterized by a different number of latent states (regimes). Therefore, the proposed two-step procedure enables the detection of the different dynamic features across groups which may not be detected using a single-step procedure.

Albeit very appealing for its potential, this approach presents some open issues which should be investigated in further research. Since the determination of the number of groups in the first step as well as the definition of the number of regimes in the second step of the procedure plays a crucial role in our analysis, in future work, we intend to consider and compare alternative indicators for the determination of M and K. Moreover, we will compare our proposal with alternative procedures, including a single-step procedure by estimating a mixture of MS models (see, e.g., [4]), thus checking how the two-step method is suitable and preferable. Furthermore, it can be of interest to consider alternative observed variables z_k in the mixture model estimated in the first step of the procedure. Finally, we also plan to consider other probability density functions for ε_t in the MS model specification, e.g., Student-t or GED distributions as discussed in Sect. 2.2.

Acknowledgements We thank Giampiero Gallo, Attilio Gardini and the three anonymous referees for their useful comments. Financial support from Italian PRIN 2010–2011 grant "Multivariate statistical models for risk assessment" is gratefully acknowledged.

[5]Note that the posterior probabilities for the crisis/negative regime across groups are not truly comparable but allows us to gather information on the development of the latest crisis, which can be extremely useful as early warning indicator.

References

1. Al-Anaswah, N., Wilfling, B.: Identification of speculative bubbles using state-space models with Markov-switching. J. Bank. Financ. **35**, 1073–1086 (2011)
2. De Angelis, L.: Latent class models for financial data analysis: some statistical developments. Stat. Methods Appl. **22**, 227–242 (2013)
3. De Angelis, L., Paas, L.J.: A dynamic analysis of stock markets using a hidden Markov model. J. Appl. Stat. **40**, 1682–1700 (2013)
4. Dias, J.G., Vermunt, J.K., Ramos, S.: Mixture hidden Markov models in finance research. In: Fink, A., Lausen, B., Seidel, W., Ultsch, A. (eds.) Advances in Data Analysis, Data Handling and Business Intelligence. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 451–459. Springer, Berlin (2010)
5. Gallo, G.M., Otranto, E.: Volatility spillovers, interdependence and comovements: a Markov switching approach. Comput. Stat. Data Anal. **52**, 3011–3026 (2008)
6. Garcia, R.: Asymptotic null distribution of the likelihood ratio test in Markov switching models. Int. Econ. Rev. **39**, 763–788 (1998)
7. Guidolin, M., Timmermann, A.: Asset allocation under multivariate regime switching. J. Econ. Dyn. Control **31**, 3503–3544 (2007)
8. Haas, M., Mittnik, S., Paolella, M.: A new approach to Markov-switching GARCH models. J. Financ. Econom. **2**, 27–62 (2004)
9. Hamilton, J.D.: A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica **57**, 357–384 (1989)
10. Hamilton, J.D.: Specification testing in Markov-switching time-series models. J. Econom. **70**, 127–157 (1996)
11. Krolzig, H.-M.: Markov Switching Vector Autoregressions: Modelling, Statistical Inference and Application to Business Cycle Analysis. Lecture Notes in Economics and Mathematical Systems, vol. 454. Springer, Berlin (1997)
12. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
13. Rydén, T., Terasvirta, T., Asbrink, S.: Stylized facts of daily return series and the hidden Markov model. J. Appl. Econom. **13**, 217–244 (1998)
14. Vermunt, J.K., Magidson, J.: LG-Syntax Users Guide: Manual for Latent GOLD 4.5 Syntax Module. Statistical Innovations, Belmont (2008)

A Multivariate Stochastic Volatility Model for Portfolio Risk Estimation

Andrea Pierini and Antonello Maruotti

Abstract

A Multivariate Latent Stochastic Volatility Factor Model is introduced for the estimation of volatility and optimal allocation of stocks portfolio in a Markowitz type portfolio. Returns on a set of 5 banks among the best capitalized banks' stocks traded on the Italian stock market (BIT) between 1 January 1986 and 31 August 2011 are modeled. Computational complexities arising in the estimation step are dealt by simulation-based methods, introducing a Griddy Gibbs sampler. The association structure among time-series is captured via a factor model, which reduces the computational burden required in the estimation step.

Keywords

Factor model • Griddy Gibbs • Markowitz portfolio • Stochastic volatility model

1 Introduction

Volatility modelling plays an important role in the analysis of financial time series. The persistence of volatility phenomenon is the most well-established effect exhibited by financial time series. Indeed, the variance of returns exhibits high serial

A. Pierini (✉)
Dipartimento di Economia, Universitá of Roma Tre, Via S. D'Amico 77, 00145 Roma, Italy
e-mail: andrea.pierini@uniroma3.it

A. Maruotti
S3RI and School of Mathematics, University of Southampton, Highifield Southampton SO17 1BJ, UK

Dipartimento di Scienze Politiche, Universitá di Roma Tre, Via Chiabrera 199, 00145 Roma, Italy
e-mail: a.maruotti@soton.ac.uk; antonello.maruotti@uniroma3.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_17, Published online: 12 November 2014

autocorrelation, which becomes evident by looking at periods of high volatility, with large changes in assets returns being followed by large ones as well, and at periods of low volatility in which small changes are followed by small ones. As this observation obviously is of great interest, capturing this effect could be challenging. This is the reason why stochastic volatility (SV) models have bee introduced and undergone a lot of research during the last two decades. Since the seminal papers by [14, 15], the univariate SV model has been widely used and several estimation methods have been introduced (see e.g. [3, 10, 13]). Nevertheless, as pointed out by e.g. [2], assets are linked together or influenced by common unobserved factors, which render the univariate approach too restrictive. It is then crucial to extend the univariate SV model to the multivariate case in order to capture the covariation effect. Several alternatives can be considered to describe the time evolution of the joint distribution of different assets (see e.g. [1, 7, 11]).

In this paper we aim at providing a multivariate SV model, based on a latent structure, in which covariation is accounted for via a factor model. Appropriately accounting for covariation is crucial in terms of portfolio diversification and asset allocation. Indeed, the ultimate goal of this paper is to provide indications on portfolio diversification with minimum risk, in a Markovitz framework.

Parameters estimation could be cumbersome and inference becomes therefore hard. This has lead to a substantial development of sampling-based methods in order to obtain parameter estimates, such as rejection sampling, Markov Chain Monte Carlo and Monte Carlo integration (see e.g. [3, 8, 12]). To reduce the computational burden often involved in sampling procedures, we adopt a nested Griddy Gibbs as sampler. In this way, we avoid the need of simulated density that mimic the shape of the preceding one and of guaranteeing the dominance (as imposed by [10]).

The proposed SV model is then applied to a subset of 5 banks among the best capitalized banks' stocks traded on the Italian stock market (BIT) between 1 January 1986 and 31 August 2011.

The paper is organized as follows. In Sect. 2 we introduce the model, while its computational aspects are discussed in Sect. 3. Section 4 briefly introduces the data, and the obtained results. Section "Conclusions" concludes.

2 Model Summary

In the following, we consider the SV parameterization introduced in [8, 9]. Let $\mathbf{r}_t = (r_t^{(1)}, \ldots, r_t^{(n)})$ an array of n stock returns, in which $r_t^{(i)}$ is the return for the i-th asset at time t and $\mathbf{a}_t = a(_{1t}, \ldots, a_{Kt})$ the vector of K common factors at time t, $t = 1, \ldots, T$. The SV parameterization we consider sets

$$\mathbf{r}_t = \boldsymbol{\mu}_t + \Theta \mathbf{a}_t + \mathbf{e}_t \tag{1}$$

$$\mu_t^{(i)} = \beta_0^{(i)} + \beta_1^{(i)} r_{t-1}^{(i)} + \ldots + \beta_p^{(i)} r_{t-p}^{(i)} \tag{2}$$

$$e_t^{(i)} = \sqrt{h_t^{(i)}}\epsilon_t^{(i)} \tag{3}$$

$$\log(h_t^{(i)}) = \alpha_0^{(i)} + \alpha_1^{(i)} \log(h_{t-1}^{(i)}) + v_t^{(i)}. \tag{4}$$

In this factor model, the matrix Θ is a constant $(n \times K)$ matrix of factor loading with $K < n$, $\boldsymbol{\mu_t} = (\mu_t^{(1)}, \ldots, \mu_t^{(n)})'$, the errors $\mathbf{e}_t \sim N(\mathbf{0}, \mathbf{E})$ are serially and mutually independent of all other error terms. The errors $\epsilon_t^{(i)}$ and $v_t^{(i)}$, $i = 1, \ldots, n$ are serially and mutually independent $N(0, 1)$ and $N(0, \sigma_v^{2,(i)})$, $E(\mathbf{a}_t) = \mathbf{0}$, $Cov(\mathbf{a}_t) = I$. We have also $|\alpha_1^{(i)}| < 1$, so that the factor log-volatility processes are stationary. Furthermore, $\beta^{(i)} = (\beta_1^{(i)}, \ldots, \beta_p^{(i)})$ are fixed regression parameters.

It follows from this model that the marginal distribution of the returns is multivariate Gaussian with mean $\boldsymbol{\mu_t}$ and covariance $\Theta \mathbf{H}_t \Theta' + \mathbf{E}$, where $\mathbf{H}_t = diag(h_t^{(1)}, \ldots, h_t^{(n)})$. Of course, it should also be noted that more complicated dynamics could be introduced in the latent SV process and for our purpose there is no need to estimate $\mathbf{a}_t$.

Jacquier et al. [9] proposed to use MCMC methods to estimate model parameters. However, the method has not been implemented for a multivariate financial portfolio. In the following we provide computational details to obtain parameter estimates.

3 Computational Details

Let us denote with $\omega^{(i)} = (\sigma_v^{2,(i)}, \alpha_0^{(i)}, \alpha_1^{(i)})'$ and $h^{(i)} = (h_1^{(i)}, \ldots, h_T^{(i)})$, $r^{(i)} = (r_1^{(i)}, \ldots, r_T^{(i)})$, $i = 1, \ldots, n$. The likelihood function can be written as:

$$f(r^{(i)} \mid \beta^{(i)}, \omega^{(i)}) = \int_{\mathbf{R}^T} f(r^{(i)} \mid X^{(i)}, \beta^{(i)}, h^{(i)}) f(h^{(i)} \mid \omega^{(i)}) dh^{(i)}. \tag{5}$$

To tackle this estimation problem we use the MCMC method as in [8].

Let's describe the steps of the algorithm that implements the MCMC method. *First step*: for each return i the optimal lag p of the Eq. (2) is selected, independently from the other returns, according to the AIC criterion, fixing $v_t^{(i)} = 0$.
Second step: for each return i the following likelihood function, related to the Eqs. (1)–(4) with $v_t^{(i)} = 0$, is maximized to find the MLE parameters as in [6]:

$$-L(\beta^{(i)}, \omega^{(i)}) = T^{-1} \sum_{t=1}^{n} a_t^{(i)} / \tilde{h}_t^{(i)} + \log(\tilde{h}_t^{(i)}) \tag{6}$$

where

$$\begin{aligned} a_t^{(i)} &= r_t^{(i)} - \mu_t^{(i)} \\ \tilde{h}_t^{(i)} &= \log(h_t^{(i)}) = \alpha_0^{(i)} + \alpha_1^{(i)} \tilde{h}_{t-1}^{(i)} \end{aligned} \tag{7}$$

These parameter estimations are chosen as the initial values for the conditional posterior distributions used by the MCMC iterations.
Third step: Steps 1,2 are repeated for each intrinsic values i corresponding to the return i, given by $P/E^{(i)} \times EPS^{(i)}$, where $P/E^{(i)}$ is the Price to Earnings ratio and $EPS^{(i)}$ is the Earning per share.

As the intrinsic value is considered to give indication on the return, the obtained parameter estimates are chosen as the prior distribution parameters used to define the conditional posterior distributions needed by the MCMC iterations.

Let's call these prior parameters as follow:
$\beta^{0,(i)} = (\beta_1^{0,(i)}, \ldots, \beta_p^{0,(i)})'$ and their variances $A^{0,(i)} = \text{diag}(\sigma^2_{\beta_1^{0,(i)}}, \ldots, \sigma^2_{\gamma^{0,(i)}})$ for the mean parameters, $\omega^{0,(i)} = (\sigma_v^{2,0,(i)}, \alpha_0^{0,(i)}, \alpha_1^{0,(i)})'$ and $C^{0,(i)} = \text{diag}(\sigma^2_{\alpha_0^{0,(i)}}, \sigma^2_{\alpha_1^{0,(i)}})$ for the volatility parameters.

Moreover the prior distributions are hypothesized multivariate normal for $\beta^{(i)} \sim N(\beta^{0,(i)}, A^{0,(i)})$ and $\alpha^{(i)} \sim N(\alpha^{0,(i)}, C^{0,(i)})$, inverted Chi squared for $\sigma_v^{2,(i)}$, that's to say $T\lambda/\sigma_v^{2,(i)} \sim \chi^2_T$, with λ a scale parameter.
Fourth step: We consider simulation-based methods. The MCMC Gibbs sampling estimation of the model (1)–(4), after combining the prior distributions with the likelihood using the Bayes' rule, consists in drawing random samples from the conditional posterior distributions

$$\begin{aligned} &f(\beta^{(i)}|r^{(i)}, x^{(i)}, h^{(i)}) \sim N(\beta^{*,(i)}, A^{*,(i)}) \\ &f(h_t^{(i)}|r^{(i)}, x^{(i)}, h^{(i)}, \beta^{(i)}, \omega^{(i)}) \\ &f(\alpha^{(i)}|h^{(i)}, \sigma_v^{2,(i)}) \sim N(\alpha^{*,(i)}, C^{*,(i)}) \\ &f(d/\sigma_v^{2,(i)}|h^{(i)}, \alpha^{(i)}) \sim \chi^2_{2T-1} \end{aligned} \tag{8}$$

in a sequence from initial value of the conditioning variables, with step by step substitutions of the new sampled values to the previous ones, until a number of iteration g is reached.

That's is to say, at the MCMC Gibbs iteration j with $j = 1, \ldots, g$:

1. we draw a random sample ${}_{[j]}\beta^{(i)}$ from: $f(\beta^{(i)}|r^{(i)}, x^{(i)}, {}_{[j-1]}\, h^{(i)})$
2. we draw a random sample ${}_{[j]}h_t^{(i)}$ from: $f(h_t^{(i)}|r^{(i)}, x^{(i)}, {}_{[j-1]}\, h_{-t}^{(i)}, {}_{[j]}\, \beta^{(i)}, {}_{[j-1]}\, \omega^{(i)})$
3. we draw a random sample ${}_{[j]}\alpha^{(i)}$ from: $f(\alpha^{(i)}|_{[j]}h^{(i)}, {}_{[j-1]}\, \sigma_v^{2,(i)})$
4. we draw a random sample ${}_{[j]}\sigma_v^{2,(i)}$ from: $f(d/\sigma_v^{2,(i)}|_{[j]}h^{(i)}, {}_{[j]}\, \alpha^{(i)})$

This completes a MCMC Gibbs iteration and current parameters values are

$$({}_{[j]}\beta^{(i)}, {}_{[j]}\, h^{(i)}_{1,\ldots,t}, {}_{[j]}\, \alpha^{(i)}, {}_{[j]}\, \sigma_v^{2,(i)})$$

In this way we obtain random samples $\{{}_{[j]}\beta^{(i)}\}_{j=g_0,\ldots,g}$, $\{{}_{[j]}h^{(i)}\}_{j=g_0,\ldots,g}$ $\{{}_{[j]}\alpha^{(i)}\}_{j=g_0,\ldots,g}$, $\{{}_{[j]}\sigma_v^{2,(i)}\}_{j=g_0,\ldots,g}$ that can be used to make inference.

Our estimations are the point estimation sample means of the previous random samples after eliminating the first $g_0 - 1$ values, that's to say

$$\begin{aligned}
\hat{\beta}^{(i)} &= \frac{1}{g-g_0} \sum_{j=g_0,\ldots,g\ [j]} \beta^{(i)} \\
\hat{h}_t^{(i)} &= \frac{1}{g-g_0} \sum_{j=g_0,\ldots,g\ [j]} h_t^{(i)} \\
\hat{\alpha}^{(i)} &= \frac{1}{g-g_0} \sum_{j=g_0,\ldots,g\ [j]} \alpha^{(i)} \\
\hat{\sigma}_v^{2,(i)} &= \frac{1}{g-g_0} \sum_{j=g_0,\ldots,g\ [j]} \sigma_v^{2,(i)}
\end{aligned}$$

The value of g_0 is chosen so that the estimation $(\hat{\beta}^{(i)}, \hat{h}^{(i)}, \hat{\alpha}^{(i)}, \hat{\sigma}_v^{2,(i)})$ of the parameters $(\beta^{(i)}, h^{(i)}, \alpha^{(i)}, \sigma_v^{2,(i)})$ is stable in the sense that after g_0 the means obtained by adding one by one the successive random sample of the Gibbs are almost identical.

In the Eq. (8) the Bayes' rule gives $A^{*,(i)} = \left(\sum_{t=1}^{T} x_{0,t}^{(i)} x_{0,t}^{(i)\prime} + (A^{0,(i)})^{-1}\right)^{-1}$,

$x_{0,t}^{(i)} = x_t^{(i)} / \sqrt{h_t^{(i)}}$

$\beta^{*,(i)} = A^{*,(i)} \left(\sum_{t=1}^{T} x_{0,t}^{(i)} r_{0,t}^{(i)} + (A^{0,(i)})^{-1} \beta^{0,(i)}\right)$, $r_{0,t}^{(i)} = r_t^{(i)} / \sqrt{h_t^{(i)}}$,

$C^{*,(i)} = \left(\sum_{t=2}^{T} y_t^{(i)} y_t^{(i)\prime} / \sigma_v^{2,(i)} + (C^{0,(i)})^{-1}\right)^{-1}$, $y_t^{(i)} = (1, ln(h_t^{(i)}))'$

$\alpha^{*,(i)} = C^{*,(i)} \left(\sum_{t=2}^{T} y_t^{(i)} ln(h_t^{(i)}) / \sigma_v^{2,(i)} + (C^{0,(i)})^{-1} \alpha^{0,(i)}\right)$

$d = T\lambda + \sum_{t=2}^{T} v_t^{2,(i)}$, $r_t^{(i)}$ is the compound return, $x_{j,t}^{(i)}$ is its lagged value.

The posterior distribution $f(h_t^{(i)} | r^{(i)}, x^{(i)}, h^{(i)}, \beta^{(i)}, \omega^{(i)})$ is a non standard one even if its density is known up to a normalizing constant [8].

Therefore a nested Gibbs sampler of type Griddy is implemented in the following way:

1. a grid of values for $h_t^{(i)}$ is selected, say, $h_{t,1}^{(i)} \leq h_{t,2}^{(i)} \leq \ldots \leq h_{t,m}^{(i)}$; its posterior distribution is evaluated on this values to obtain $w_s = f(h_{t,s}^{(i)} | r^{(i)}, x^{(i)}, h_{-t}^{(i)}, \beta^{(i)}, \omega^{(i)})$, $s = 1, \ldots, m$.
2. The w_s are used to obtain an approximation of the inverse cumulative distribution function of $f(h_{t,s}^{(i)} | r^{(i)}, x^{(i)}, h_{-t}^{(i)}, \beta^{(i)}, \omega^{(i)})$.
3. A uniform random variable between 0 and 1 is drawn and transformed via the preceding step 2 to obtain a random drawn for $h_t^{(i)}$.

Fifth step: the estimation $(\hat{v}_{i,i}(T+1|\Omega_T))_{i\in\{1,\ldots,n\}}$ of the volatility matrix $(v_{i,i}(T+1|\Omega_T))_{i\in\{1,\ldots,n\}}$, that's to say $v_{i,i}(T+1) = h_{T+1}^{(i)}$,will be obtained in the following way, at the iteration j of the Gibbs sampler, $j = g_0, \ldots, g$:

1. we draw a random sample $v_{T+1}^{(i)}$ from $N(0, {}_{[j]}\sigma_v^{2,(i)})$ and the Eq. (4) with ${}_{[j]}\beta^{(i)}$ is used to compute ${}_{[j]}h_{T+1}^{(i)}$;

2. we draw a random sample $\epsilon_{T+1}^{(i)}$ from $N(0,1)$ to obtain $e_{T+1}^{(i)} = \sqrt{{}_{[j]}h_{T+1}^{(i)}}\epsilon_{T+1}^{(i)}$ and the Eq. (2) with ${}_{[j]}\alpha^{(i)}$ is used to compute ${}_{[j]}r_{T+1}^{(i)}$;
   In this way we obtain a random sample $\{{}_{[j]}h_{T+1}^{(i)}\}_{j=g_0,\ldots,g}$ and a random sample $\{{}_{[j]}r_{T+1}^{(i)}\}_{j=g_0,\ldots,g}$ that can be used to make inference.
   Our estimation is the point forecast of the previous two random samples, that's to say $\hat{h}_{T+1}^{(i)} = \frac{1}{g-g_0}\sum_{j=g_0,\ldots,g}{}_{[j]}h_{T+1}^{(i)}$ and $\hat{r}_{T+1}^{(i)} = \frac{1}{g-g_0}\sum_{j=g_0,\ldots,g}{}_{[j]}r_{T+1}^{(i)}$.

Sixth step: To estimate the off-diagonal elements of V, we consider the multivariate Latent Factor model (see [4]).
Seventh step: The Markowitz problem can be foretasted at time $T+1$ using the preceding estimation of $v_{i,j}$ and $r^{(i)}$ that we called $\hat{v}_{i,j}(T+1)$ and $\hat{r}_{T+1}^{(i)}$, by solving through a quadratic programming method, the following:

$$min_{\gamma\in\mathbf{R}}\{\gamma'\hat{V}(T+1)\gamma \ : \gamma'\cdot\underline{1} = 1, \gamma'\cdot\hat{r}_{T+1} = R_p, \gamma \geq 0\} \tag{9}$$

where $\gamma = (\gamma_1,\ldots,\gamma_n)$ and $R_p \in [min_{\{i:i=1,\ldots,n\}}\hat{r}_{T+1}^{(i)}, max_{\{i:i=1,\ldots,n\}}\hat{r}_{T+1}^{(i)}]$.

4 Data and Results

The model is applied to a subset of the entire stocks' universe among the series of data regarding the best capitalized 5 banks' stocks traded on the Italian stock market (BIT) between 1 January 1986 and 31 August 2011. Data are shown in Fig. 1.

This figure shows the histograms and the fitted normal density. Even if the empirical distributions of the returns are symmetry and uni-modal, the fitted normal curves are not enough similar to the empirical counterparts in the tails. Tests of Jarque–Bera for normality reject the null hypothesis H_0 of normality at 95 %.

Moreover tests Ljung–Box for the squared of residuals $e_t^{(i)}$ of Eqs. (1)–(2), reject the null hypothesis H_0 of no ARCH effects at 95 %. Therefore it is necessary to include the Eqs. (3)–(4), that's to say the SV model part, in order to obtain an unconditional distribution of the residuals $e_t^{(i)}$ that has heavier tails, an excess of kurtosis with respect to a gaussian distribution. This is in agreement with the financial data in hand.
With respect to the model (1)–(2) setting $v_t^{(i)} \equiv 0$,where $r_t^{(i)}$ is the intrinsic value, the explanatory variables $x_{j,t}^{(i)}$ are the lagged values $r_{t-j}^{(i)}$ and the exogenous variable $z_t^{(i)}$ is the market index, we give the optimal lags in Table 1, which minimize the AIC.

The posterior distribution parameters, calculated as means of the last $g-g_0 = 100$ iterations of sampling from the MCMC posterior conditional distribution are as follows (Tables 2 and 3):

From Table 2 it can be seen that $\hat{\beta}_0^{(i)} \simeq 0$ so there is no constant term in the mean Eqs. (1)–(2). As $\hat{\beta}_k^{(i)} \neq 0, k > 0$ it can be said that return is dependent on its past

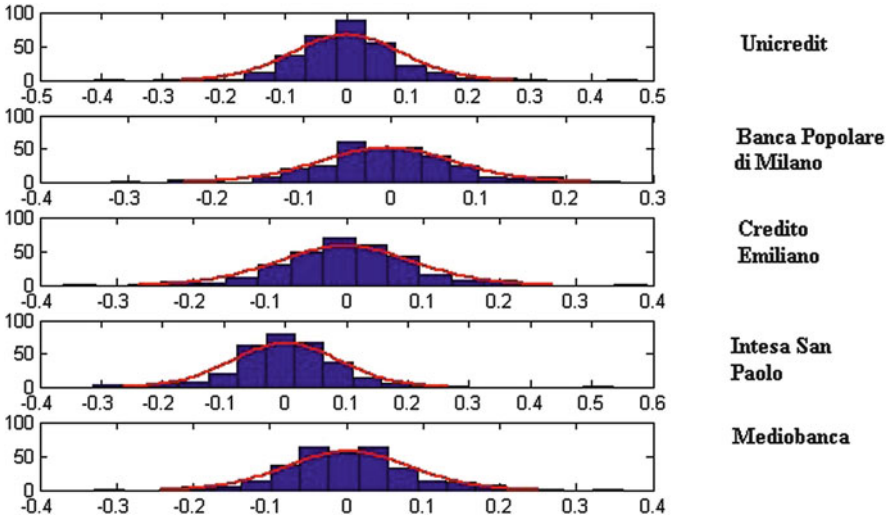

Fig. 1 Data description

Table 1 Optimal AIC lags

	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Opt. lag
Unicredit	−722.66	−722.05	−720.16	−719.8	−718.43	1
BPM	758.76	758.63	760.19	761.65	760.82	2
Credito Emiliano	530.94	531.82	533.74	535.72	536.67	1
Intesa	219.81	175.05	167.33	161.86	102.65	5
Mediobanca	606.37	607.66	599.84	588.03	580.52	5

Table 2 $\hat{\beta}^{*,(i)} = (\hat{\beta}_0^{(i)}, \ldots, \hat{\beta}_p^{(i)})'$ posterior distribution parameters (stock i)

i	$\hat{\beta}_0^{(i)}$	$\hat{\beta}_1^{(i)}$	$\hat{\beta}_2^{(i)}$	$\hat{\beta}_3^{(i)}$	$\hat{\beta}_5^{(i)}$	$\hat{\beta}_5^{(i)}$
1	−0.0012	0.0322	0	0	0	0
2	−0.0039	−0.2394	−0.1512	0	0	0
3	−0.0018	−0.1528	0	0	0	0
4	−0.0026	−0.6606	−0.0108	−0.0783	0.1415	0.3238
5	0.0016	−0.0050	−0.0874	−0.1363	−0.1751	−0.2782

values. Moreover, from Table 3 it can be seen that the volatility is dependent from its past as each $\hat{\alpha}_1^{(i)} \neq 0$ in the Eq. (4). Lastly $\hat{\sigma}_v^{2,(i)} > 0$ is not negligible so the SV model, which introduces $v_t^{(i)}$, is capable to improve a pure ARCH model for the volatility.

The estimated variance-covariance (*risk*) matrix is given in Table 4.

As expected, it can be seen in Table 4 that $0.47 < \hat{v}_{i,j} < 0.70, i \neq j$, so there are positive correlations among the series. Indeed the series belong to the same sector.

Table 3 $\hat{\omega}^{(i)} = (\hat{\sigma}_v^{2,(i)}, \hat{\alpha}_0^{(i)}, \hat{\alpha}_1^{(i)})'$ posterior distribution parameters (stock i)

i	$\hat{\sigma}_v^{2,(i)}$	$\hat{\alpha}_0^{(i)}$	$\hat{\alpha}_1^{(i)}$
1	2.3661	−0.00023698	0.98773
2	3.1517	−0.0063111	0.9798
3	2.2379	−0.0028054	0.99625
4	1.6529	−0.044743	0.86632
5	2.1015	−0.077245	0.7201

Table 4 Covariance risk matrix estimation

$$\begin{pmatrix} \hat{v}_{1,1} & \cdots & \hat{v}_{1,n} \\ \vdots & \ddots & \vdots \\ \hat{v}_{n,1} & \cdots & \hat{v}_{n,n} \end{pmatrix}_{T+1|\Omega_T} =$$

2.3211	0.6257	0.59352	0.69577	0.6879
0.6257	1.6836	0.47016	0.55623	0.54937
0.59352	0.47016	1.7268	0.56166	0.60608
0.69577	0.55623	0.56166	1.6892	0.6187
0.6879	0.54937	0.60608	0.6187	1.6936

Table 5 Latent factor loadings estimates

Series	$\hat{\Theta}_1$	$\hat{\Theta}_2$
Unicredit	0.8176	0.33716
BPM	0.65679	0.26308
Credito Emiliano	0.37386	0.85376
Intesa	0.70746	0.34807
Mediobanca	0.66952	0.41672

It can be seen that the variances dominate the covariances which are all positive. Thus, it is possible to find portfolios that have lower risk than either single asset. Among those ones we choose the minimum risk portfolio.

In order to provide further insights on the association structure we may look at the estimated latent factors (Table 5). The first latent factor can be associated with Unicredit, BPM and Intesa San Paolo, whilst the second one is mainly related to Credito Emiliano. This can be seen in Fig. 2, where we call Component i the $\hat{\Theta}_i, i = 1, 2$ and the numbers are the stocks in the same order of Table 5.

The possible double clustering suggested by the Fig. 2 could drive another portfolio diversification in order to take into consideration the different dependency (loading) each stock's return has of the common latent factors.

Of course, as a central issue, we have to solve the quadratic programming of Markowitz in order to optimize our portfolio. We propose the following optimal (with minimum risk) fractions to invest in each stock as the numerical solution of the optimization problem given in the Eq. (9) (Table 6).

By depicting the optimal fractions in Fig. 3, it can be seen that a good diversification among the stocks is obtained.

The optimal fractions $\hat{\gamma}_{(i)}$ give a risk of 0.90724 and a monthly portfolio return of 0.035194. Different portfolio choice are possible that can get greater return or

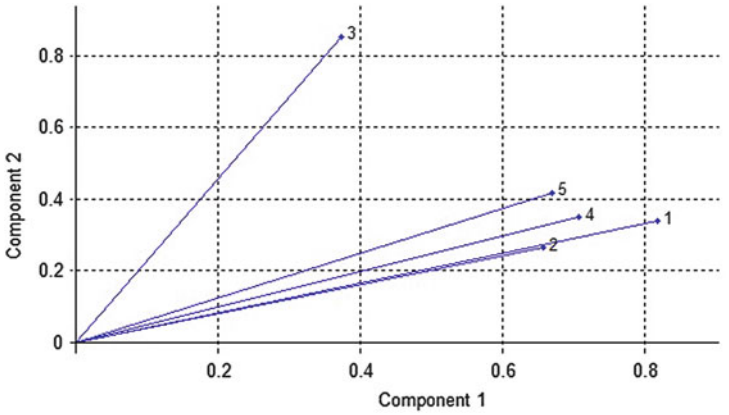

Fig. 2 Latent factor loadings

Table 6 $\hat{\gamma}_{(i)}$ optimal (with minimum risk) fractions

Series	Unicredit	BPM	Credito Emiliano	Intesa	Mediobanca
$\hat{\gamma}_{(i)}$	0.10474	0.2493	0.23337	0.21036	0.20222

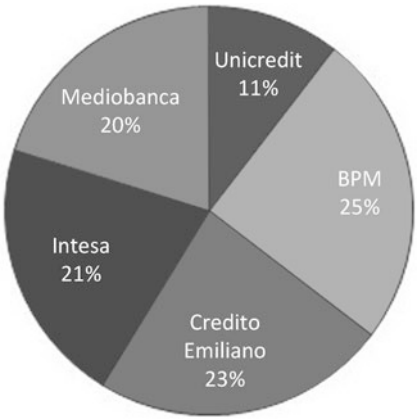

Fig. 3 Optimal fractions

better latent shocks warranty at the expense of a greater risk. Thus in a risk averse view, an investor should choose the minimum risk portfolio suggested.

Conclusions

We have conducted an empirical investigation of stochastic volatility of major Italian banks, by introducing a computational feasible algorithm based on simulation techniques. The proposed estimation methodology is easily implementable and this is an important step forward on multivariate volatility estimation, since the likelihood function of stochastic volatility models is not easily calculable. The procedure proposed in this work attempts to combine the simplicity of the factor model with the sophistication of stochastic volatility procedures. Open problems remain, primarily in the modelling of multivariate heavy-tailed or skewed error distributions, as well as the computational burden required in the estimation step in the modelling of high dimensional data. In time, further significant developments can be achieved by introducing a time-varying latent structure such as parsimonious hidden Markov models, which are able to reduce the curse of dimensionality of the considered problem and account for well-know stylized facts arising in the stock returns modelling.

References

1. Asai, M., McAleer, M., Yu, J.: Multivariate stochastic volatility: a review. Econ. Rev. **25**, 145–175 (2006)
2. Aydemir, A.B.: Volatility Modelling in Finance. In: Knight J., Satchell, S. (eds.) Forcasting Volatility in Financial Markets, pp. 1–46, Butterworth-Heinemann, Oxford (1998)
3. Broto, C., Ruiz, E.: Estimation methods for stochastic volatility models: a survey. J. Econ. Surv. **18**, 613–649 (2004)
4. Chib, S., Omori, Y., Asai, M.: Multivariate stochastic volatility. In: Andersen, T.G. et al. (eds.) Handbook of Financial Time Series, pp. 365–400, Springer, New York (2009)
5. Connor, G.: The three types of factor models. Financ. Analysts J. **51**, 42–46 (1995)
6. Francq, C., Zakoian, J.M.: GARCH models: structure, statistical inference and financial applications, Wiley and Sons (2010)
7. Harvey, A.C., Ruiz, E., Shephard, N.: Multivariate stochastic variance models. Rev. Econ. Stud. **61**, 247–64 (1994)
8. Jacquier, E., Polson, N., Rossi, P.: Bayesian analysis of stochastic volatility models. J. Bus. Econ. Stat. **12**, 371–417 (1994)
9. Jacquier, E., Polson, N.G., Rossi, P.E.: Stochastic volatility: univariate and multivariate extensions. CIRANO Working paper 99s-26, Montreal (1999)
10. Langrock, R., MacDonald, I., Zucchini, W.: Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models. J. Empirical Finance **19**, 147–161 (2011)
11. Pitt, M.K., Shephard, N.: Time varying covariances: a factor stochastic volatility approach. In: Bernardo J.M., Berger J.O., Dawid A.P., Smith A.F.M. (eds.) Bayesian Statistics, vol. 6, pp. 547–570. Oxford University Press, Oxford (1999)
12. Shephard, N., Pitt, M.K.: Likelihood analysis of non-gaussian measurment time series. Biometrika **84**, 653–667 (1997)
13. Shephard, N.: Stochastic Volatility: Selected Readings. Oxford University Press, Oxford (2004)
14. Taylor, S.J.: Financial returns modelled by the product of two stochastic processes—a study of the daily sugar prices 1961–1975. In: Anderson, O.D. (ed.) Time Series Analysis: Theory and Practice, vol. 1. North-Holland, Amsterdam (1982)
15. Taylor, S.J.: Modeling Financial Time Series. Wiley, Chichester (1986)

A Thick Modeling Approach to Multivariate Volatility Prediction

Alessandra Amendola and Giuseppe Storti

Abstract

This paper proposes a modified approach to the combination of forecasts from multivariate volatility models where the combination is performed over a restricted subset including only the best performing models. Such a subset is identified over a rolling window by means of the Model Confidence Set (MCS) approach. The analysis is performed using different combination schemes, both linear and non linear, and considering different loss functions for the evaluation of the forecasting performance. An application to a vast dimensional portfolio of 50 NYSE stocks shows that (a) in non-extreme volatility periods the use of forecast combinations allows to improve over the predictive accuracy of the single candidate models (b) performing the combination over the subset of most accurate models does not significantly reduce the accuracy of the combined predictor.

Keywords

Forecast combination • Multivariate volatility • Thick modeling • Weights estimation

1 Introduction

The econometric and statistical literature has been recently paying much attention to the analysis of model uncertainty in multivariate volatility models. So far, most of the research efforts in this field have been dedicated to the evaluation of the

A. Amendola (✉) • G. Storti
Department of Economics and Statistics (DiSES) & Statlab, University of Salerno,
Via Giovanni Paolo II, 132, 84084, Fisciano (SA), Italy
e-mail: alamendola@unisa.it; storti@unisa.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_18, Published online: 21 November 2014

forecasting performance focusing on problems such as the choice of the loss function and of the volatility proxy used for the evaluation of forecasts. In particular, Patton and Sheppard [14] and Laurent et. al [13] have analyzed the effects that the quality of the proxy can have on the ranking of forecasting models implied by a given loss function. Both these papers, however, do not investigate the possibility of combining volatility forecasts from different models as a way for improving the forecast accuracy.

In this paper we focus on the application of forecast combination techniques as a tool for dealing with model uncertainty in multivariate volatility prediction for vast dimensional portfolios. Under this respect it is important to remark that the dimension is a critical variable. The risk of model misspecification is indeed particularly sizeable in large dimensional problems. In this setting, it is well known that the need for reducing the number of parameters usually requires the formulation of highly restrictive assumptions on the volatility dynamics that, in most cases, are applied without any prior testing (see e.g. Pesaran et al. [16]).

Despite the undoubted relevance of this issue, the statistical problems related to the application of forecast combination techniques in multivariate volatility prediction have been largely left unexplored by the statistical and econometric literature. As a consequence of this, in most applications the standard approach is still to select the "best" specification from those that are available according to some sensible criterion. Granger and Jeon [10] refer to this strategy as a *thin* modelling approach, opposed to a *thick* modelling strategy that does not require the identification of a single best performing model but combines forecasts from different alternative specifications. Adopting a thin modelling strategy can lead to an information loss due to the fact that, discarding all the suboptimal models, we also neglect the additional information set on which they depend. On the other side, a problem with the thick approach is that, if the number of competing candidate predictors is large, combining forecasts requires the estimation of a large number of parameters. This cannot be done efficiently unless a large number of past forecasts is available. Also, combining forecasts generated from a possibly large set of different models requires taking some additional choices related to the combination scheme, which is the blending function through which individual forecasts are combined, and to the procedure used for the estimation of combination weights. Finally, in the specific case of volatility, as it will be later discussed in more detail, a further source of uncertainty is related to the choice of the proxy used for approximating the latent volatility.

In addition to the problems we have just discussed, there is also a concern about robustness since the presence of some outlying predictors, e.g. very bad performers, could introduce some bias in the estimation of the combination weights assigned to each model. In order to deal with this issue a possible solution could be to consider, for forecast combination, only a subset of models that are performing significantly

better than the others according to some appropriately chosen criterion. Under this respect, De Pooter et al. [5] suggest to select these models by the Model Confidence Set (MCS) approach [11].

The combination of volatility forecasts, in an univariate setting, has already been investigated by Amendola and Storti [1] who have proposed a GMM procedure for the estimation of the combination weights. This procedure has been later generalized to a multivariate setting [2]. More recently, the same authors have proposed and empirically compared some alternative combination strategies for multivariate volatility forecasts by considering the same dataset including 50 NYSE stocks which has been used in this paper [3]. Aim of this work is to investigate the profitability, for the prediction of vast dimensional conditional variance matrices, of forecast combination schemes in which the combination involves only a restricted set of best performing models rather than all the potential candidate models. In particular, we extend to the multivariate volatility case, the approach that has been proposed by De Pooter et al. [5]. This choice is expected to give the most relevant advantages in applications where a large set of candidate models is available. More precisely, we expect it to reduce the computational time required as well as to increase the efficiency in the estimation of the combination parameters.

The paper is structured as follows. Section 2 illustrates the combination functions used in the paper while the estimation of combination weights is discussed in Sect. 3. Section 4 presents the results of an empirical application to stock market data and concludes.

2 Reference Model and Combination Functions

2.1 The Data Generating Process

The data generating process (DGP) is assumed to be given by

$$r_t = S_t z_t \qquad \text{t=1,...,T} \tag{1}$$

where $z_t \overset{iid}{\sim} (0, I_k)$, S_t is any $(k \times k)$ positive definite (p.d.) matrix such that $S_t S_t^{'} = \tilde{H}_t = Var(r_t | I^{t-1})$, $\tilde{H}_t = C(H_{1,t}, \ldots, H_{n,t}; w)$, $H_{j,t}$ is a symmetric p.d. $(k \times k)$ matrix and I^t denotes the information available at time t. In practice the $H_{j,t}$ are forecasts of the conditional covariance matrix of r_t generated by the j-th candidate model. The function $C(.)$ is an appropriately chosen *combination function* and w is a vector of combination parameters. The weights assigned to each candidate model depend on the values of the elements of w but do not necessarily coincide with them. The DGP in (1) is very general and nests a wide range of multivariate volatility models as special cases.

2.2 The Linear Combination Function

Among all the possible choices of $C(.)$, the most common is the *linear* combination function

$$\tilde{H}_t = w_1 H_{1,t} + \ldots + w_n H_{n,t} \qquad w_j \geq 0 \tag{2}$$

where w is the vector of combination weights. A potential drawback of this specification is related to the fact that we need to assume that the weights w_j are non-negative, in order to guarantee that the estimated combined predictor $\tilde{H}_t$ is a well defined positive definite conditional variance matrix. However, such an assumption can be highly restrictive. It automatically implies that the combined volatility forecast of a single asset i is directly proportional to the volatilities predicted by each of the candidate models. The same obviously holds for conditional covariances. Also, in this paper we do not impose any constraint on the value of the sum of the weights w_j. In the early time series literature on forecasts combination it was customary to impose the convexity constraint $\sum_{j=1}^{n} w_j = 1$ which implies that $\tilde{H}_t$ is defined as a convex linear combination of candidate forecasts. However it can be easily shown that ignoring this constraint allows to correct for the presence of bias in the candidate predictors.

It is worth noting that Eq. (2) could be further generalized by including a $(k \times k)$ matrix of intercepts on the right hand side. A parsimonious way of doing this would be to use matrices of rank one

$$\tilde{H}_t = \underline{a}\underline{a}' + w_1 H_{1,t} + \ldots + w_n H_{n,t} \qquad w_j \geq 0 \tag{3}$$

where $\underline{a}$ is a k-dimensional vector of parameters. This solution has two important advantages. First, it keeps the number of parameters reasonable and linear in n. Second, $\tilde{H}_t$ is guaranteed to be positive definite if at least one of the $H_{j,t}$ is positive definite. A more parsimonious solution is to adopt a targeting strategy in which the intercept matrix $(\underline{a}\underline{a}')$ in (3) is replaced by the matrix

$$\bar{H} = \frac{1}{T} \sum_{t=1}^{T} \left(\Sigma_t - \sum_{j=1}^{n} w_j H_{j,t} \right) \tag{4}$$

where Σ_t is an appropriately chosen volatility proxy, such as a realized covariance matrix, and the weights w_j are constrained to satisfy the condition $\min(\text{eig}(\bar{H})) \geq 0$. This solution would not require the estimation of any additional parameters other than the model weights.

2.3 The Square-Root Combination Function

Alternatively, in order to get rid of the positivity constraint on the w_j, the linear combination function can be replaced by an alternative scheme known as the *square root* combination function. This is based on a linear combination of Cholesky factorizations of the candidate forecasts of the conditional covariance matrix. Differently from the previous linear function, the square root combination (for S_t) is not directly performed on the $H_{j,t}$ but on the $S_{j,t}$

$$S_t = w_1 S_{1,t} + \ldots + w_n S_{n,t} \tag{5}$$

with $\tilde{H}_t = S_t S_t'$ and $H_{j,t} = S_{j,t} S_{j,t}'$. As in the linear case, in low dimensional problems equation (5) could be further generalized by adding a matrix of intercepts. This could be introduced in a similar fashion as already discussed for the linear combination functions. In particular the intercept matrix could be specified as a rank one matrix of parameters or alternatively a two-stage targeting procedure could be used. In the latter case the matrix of intercepts could be expressed as

$$\bar{S} = \frac{1}{T} \sum_{t=1}^{T} \left(L_t - \sum_{j=1}^{n} w_j S_{j,t} \right) \tag{6}$$

where the matrix L_t is obtained from the Cholesky decomposition of Σ_t. It is important to remark that, differently from what observed for the linear case, we don't need to impose any constraints on the estimated w_j.

3 Estimation of the Combination Parameters

The combination functions depend on the vector of unknown parameters given by $w = (w_1, \ldots, w_n)'$. An obvious approach to the estimation of these parameters is based on a direct minimization of an appropriately defined loss function with respect to the unknown combination parameters. For this purpose a wide range of different loss functions could be used but, in any case, a proxy of the latent volatility is required. A measure of forecasts accuracy is then based on the comparison of a proxy of the unobserved conditional covariance matrix, such as the Realized Covariance estimator, with the combined predictor $\tilde{H}_t$.

In general, lack of accuracy of the chosen volatility proxy could result into a biased measure of predictive accuracy and, henceforth, into a biased model ranking. However, Laurent et al. [13], generalizing a similar result obtained by Patton [15] for

univariate volatility prediction, have identified a class of *robust* loss function where robustness should be evaluated in terms of invariance of the models ranking with respect to inaccuracies in the volatility estimator used for the evaluation of forecast performance.

More precisely, the estimates of the combination weights w_j are obtained by solving the following optimization problem:

$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{t=1}^{T} L(\Sigma_t; w),$$

where $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_k)'$, $L(\cdot)$ is an appropriately chosen loss function and $\{\Sigma_t\}_{t=1}^{T}$ is a symmetric positive semi-definite realized covariance matrix of dimension n recorded at time t

$$\Sigma_t = \sum_{i=1}^{n_\Delta} r_t^{(i)} (r_t^{(i)})'$$

where n_Δ is the number of equally spaced intervals of length Δ within the trading day and $r_t^{(i)}$ denotes the log-return over the i-th interval.

For the evaluation of multivariate volatility forecasts we focus on the Euclidean loss function

$$LE = vech(\Sigma_t - H_t)' vech(\Sigma_t - H_t)$$

A discussion on the properties of different loss functions in predicting multivariate volatility can be found in the paper by Laurent et al. [13] who show that this function produces a ranking of the competing forecasts that is robust to the choice of the volatility proxy.

4 Empirical Results

In this section we present the results of an application to a portfolio of 50 NYSE stocks whose symbols have been reported in Table 1.

The dataset we use is composed of price quotations observed every minute, from 9.30 a.m. to 4.00 p.m., from October 1, 1997 to July 18, 2008.[1] The raw-returns are aggregated over 5 min intervals in order to obtain a time series of daily realized covariance matrices. Furthermore, we use the available data to compute a time series of daily open-to-close returns over the period of interest. Our choice of using open-to-close returns follows the approach of Andersen et al. [4] who argue that the overnight return can be interpreted as a deterministically occurring jump. Hence the

[1] The data are available online at www.tickdata.com.

Table 1 Symbols identifying the 50 stocks NYSE included in the analyzed portfolio

ABT	AZO	CAH	CAG	F
AFL	AVY	CTL	COST	GCI
APD	BHI	CTAS	DOV	GPS
AA	BAC	C	DOW	GE
ALL	BAX	CLX	DTE	GIS
AXP	BDX	CMS	EMN	GPC
AIG	BBBY	KO	EIX	HNZ
ADI	BMY	CL	ETR	HPQ
APOL	CPB	CMA	FDO	HD
T	COF	CSC	FISV	HON

open-to-close return can be considered as the daily return adjusted for the overnight jump.[2]

Our aim is (a) comparing the performances of different models in generating one-step-ahead predictions of the conditional covariance matrix of daily returns (b) evaluating the ability of forecast combinations to improve the performance of the single candidate models (c) assessing the profitability of thick modelling schemes based on the combination of forecasts generated from the subset of the best performing models identified using the MCS approach.

The candidate models that have been considered for forecast evaluation can be classified into two groups. The first group includes MGARCH models that do not exploit intra-daily information and are fitted to time series of daily returns: the Dynamic Conditional Correlation (DCC) model [6], the BEKK model [8], the standard RiskMetrics (RM) and a Moving Covariance (MC) predictor. Differently, the second group includes models directly fitted to time series of realized covariance matrices: the Conditionally Autoregressive Wishart (CAW) model [9] and a realized version of the RM (RRM) and MC estimators (RMC). The DCC and BEKK models of order (1,1), following Engle et al. [7], have been estimated by Gaussian composite quasi maximum likelihood while the CAW model has been estimated by maximizing a Wishart quasi likelihood function. The value of the smoothing parameter of the RM and RRM estimators has been set equal to $\lambda = 0.94$, to meet the indications of the RiskMetrics technical document [12]. The length of the moving window for the calculation of the MC and RMC estimator has been set equal to $m = 100$ which is a recurrent value among practitioners. The model parameters are re-estimated approximately every month (22 days) over a rolling window of length equal to 500 days.

The accuracy in predicting the conditional covariance matrix is assessed using the Euclidean loss function and taking the 5-minute realized covariance matrix as a proxy of the latent volatility. The unconditional volatility level within the

[2]The choice of a discretization interval equal to 5 min is a standard practice in the literature since it usually guarantees a reasonable compromise between bias and variability of the realized covariance estimator.

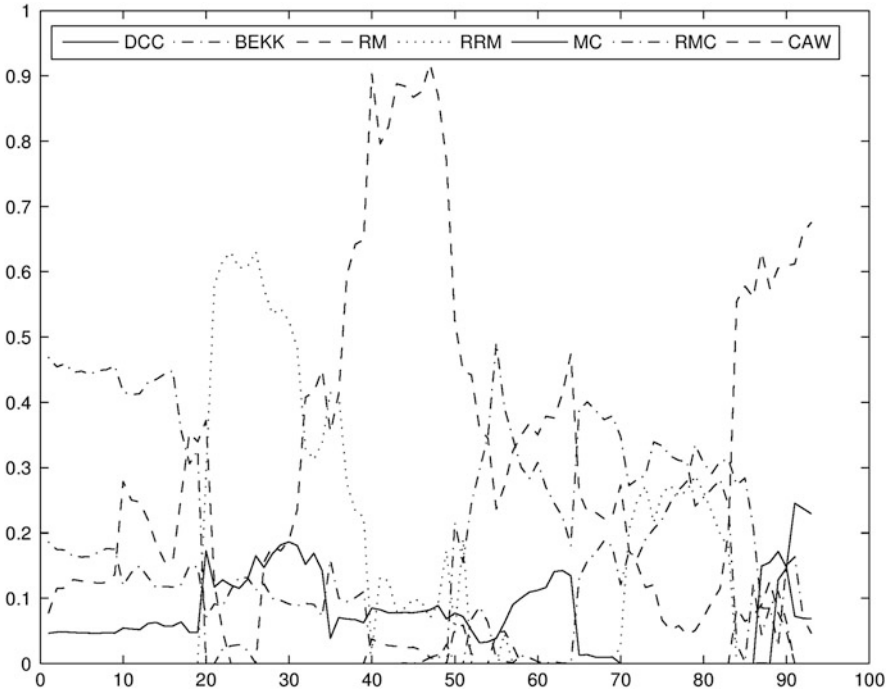

Fig. 1 Estimated weights of the linear combined predictor including all models (over 93 re-estimations)

out-of-sample forecasting period is subject to different level shifts. In particular we identify a high volatility period, associated to the dot-com bubble, from April 13, 2000 to the end of March 2003, for a total of 737 observations, a long low volatility period ending in July 2007, for a total of 1065 observations and, finally, a short high-volatility period at the end of the sample period, loosely related to the sub-prime crisis, from August 2007 to July 18, 2008, for a total of 244 observations. Since it is well known that the dynamic properties of volatilities and correlations depend on the unconditional level of volatility itself, a given model could be differently able in predicting the conditional covariance matrix in calm and storm periods. For this reason we have separately analyzed the forecast accuracy of our candidate models and combined predictors in the three above described periods. For the sake of brevity, we have not reported all the estimated combination weights but only those related to the forecast combination considering all models and not including an intercept. The estimated weights have been reported in Fig. 1, for the linear combination scheme, and Fig. 2, for the Cholesky combination. In both cases our analysis reveals that the weights assigned to the candidate models involved in the combination are remarkably varying over time adapting to the sharp changes in the volatility and correlation patterns that characterize the data of interest. This finding confirms our intuition that the structure of the volatility process and, hence, of the

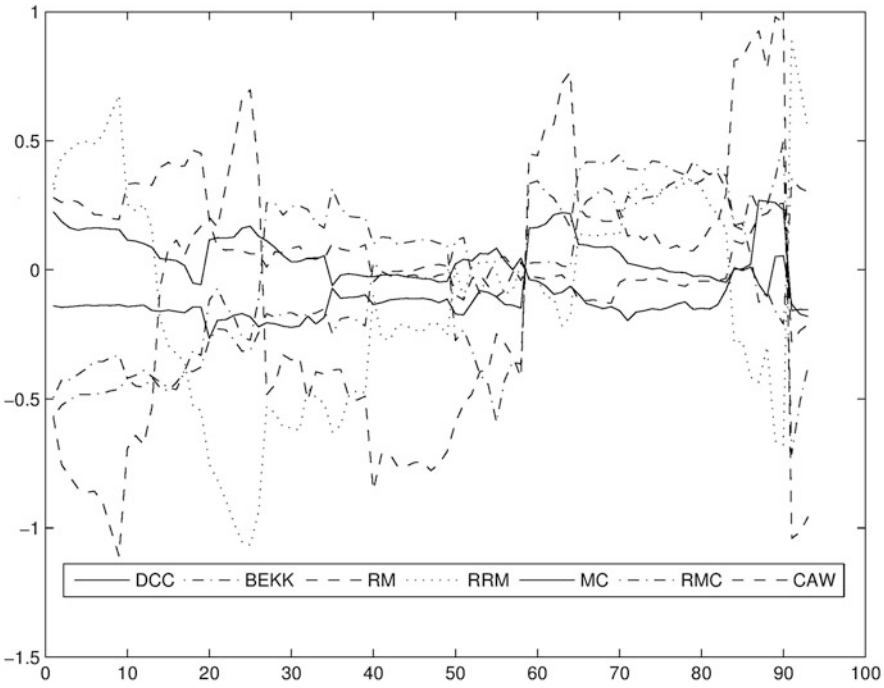

Fig. 2 Estimated weights of the Cholesky combined predictor including all models (over 93 re-estimations)

optimal combined predictor has been changing over time. For the linear combination scheme, the results in Fig. 1 show that, on average, the CAW is the most influential model in the central and final part of the forecasting period while, during the first part of the period, its role is taken by the RMC, first, and the RRM predictor, later. The RM and MC predictors are virtually excluded from the optimal predictor while the BEKK and DCC based on daily the returns are regularly present in the optimal combination even if their weights are in most cases lower than those assigned to the predictors based on realized covariances. The weights from the Cholesky combination scheme (Fig. 2) also show that the structure of the optimal predictor is far from being stable over time. However, their interpretation is not as immediate as that of the linear combination weights.

The results of the Model Confidence Set (MCS) based on the LE loss function have been summarized in Table 2. Their analysis reveals evidence that, over period 2, the combined predictors are slightly improving over the candidate models. The situation is reversed in the high volatility periods 1 and 3. Furthermore, it is evident that the thick predictors based only on the models included in the MCS are characterized by performances very similar to those of their more complex counterparts based on the whole set of candidate models.

Table 2 Average LE loss function values ($\times 10^4$) over period 1 (high volatility, dot com), 2 (low volatility), 3 (high volatility, financial crisis) and whole forecasting period (all). (^{t}): combination of the best performing models included in the MCS for that period. (^{a}): combination over all the candidate models. ($_i$): denotes combined predictors including a matrix of intercepts. ($_c$): denotes linear combination of Cholesky decompositions of candidate predictors (see Eq. (2)). #(MCS): number of models included in the MCS. (*): predictor included in the MCS at level $\alpha = 0.25$

Predictor	Period 1	Period 2	Period 3	All
DCC	4.1022	0.4073	1.6786*	1.8899
BEKK	4.0704	0.3937	1.5653	1.8578
RM	4.2977	0.3904	1.5378	1.9347
RRM	4.2924	0.3831	1.5712	1.9330
MC	3.8501*	0.3661*	1.4711*	1.7529
RMC	3.9997	0.3663*	1.5700	1.8187
CAW	3.8294*	0.3668	1.4218*	1.7399*
LE^t	3.8637	0.3661	1.4402*	1.7541
LE^t_i	4.0356	0.3953	1.4243*	1.8293
LE^t_c	3.8599	0.3656*	1.4392*	1.7523
$LE^t_{c,i}$	4.2740	0.4052	1.5046*	1.9299
LE^a	3.8633	0.3654*	1.4346*	1.7529
LE^a_i	4.1761	0.4414	1.4403*	1.9058
LE^a_c	3.8888	0.3653*	1.4234*	1.7607
$LE^a_{c,i}$	4.7827	0.9341	1.6600	2.4070
#(MCS)	2	5	10	1

Acknowledgements The authors gratefully acknowledge financial support from MIUR within the PRIN project 2010–2011 (prot. 2010J3LZEN): *Forecasting economic and financial time series: understanding the complexity and modelling structural change*.

References

1. Amendola, A., Storti, G.: A GMM procedure for combining volatility forecasts. Comput. Stat. Data Anal. **52**(6), 3047–3060 (2008)
2. Amendola, A., Storti, G.: Combination of multivariate volatility forecasts. SFB 649 Discussion Papers, DP2009-007. Humboldt University, Berlin, Germany (2009)
3. Amendola, A., Storti, G.: Model uncertainty and forecast combination in high dimensional multivariate volatility prediction. In: Proceedings of COMPSTAT 2012, ISI/IASC, 27–38 (2012)
4. Andersen, T.G., Bollerslev, T., Frederiksen, P., Nielsen, O.: Continuous-time models, realized volatilities, and testable distributional implications for daily stock returns. J. Appl. Econometrics **25**(2), 233–261 (2010)
5. De Pooter, M., Ravazzolo, F., van Dijk, D.: Term structure forecasting using macro factors and forecast combination. Board of Governors of the Federal Reserve System, Discussion Paper 993 (2010)
6. Engle, R.F.: Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. J. Bus. Econ. Stat. **20**(3), 339–350 (2002)
7. Engle, R.F., Shephard, N., Sheppard, K.: Fitting vast dimensional time-varying covariance models. Economics Series Working Papers 403. University of Oxford, Oxford (2008)
8. Engle, R.F., Kroner, K.F.: Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. Econ. Theor. **11**(1), 122–150 (1995)
9. Golosnoy, V., Gribisch, B., Liesenfeld, R.: The conditional autoregressive Wishart model for multivariate stock market volatility. J. Econometrics **167**, 211–223 (2011)
10. Granger, C.W.J., Jeon, Y.: Thick modeling. Econ. Model. **21**, 323–343 (2004)

11. Hansen, P.R., Lunde, A., Nason, J.M.: The model confidence set. Econometrica **79**, 453–497 (2011)
12. J.P. Morgan Guaranty Trust Company: RiskMetrics Technical Document, 4 edn. (1996)
13. Laurent, S., Rombouts, J.V.K., Violante, F.: On loss functions and ranking forecasting performances of multivariate volatility models. J. Econometrics **173**(1), 1–10 (2013)
14. Patton, A.J., Sheppard, K.: Evaluating volatility and correlation forecasts. In: Andersen, T.G., Davis, R.A., Kreiss, J.P., Mikosch, T. (eds.) Handbook of Financial Time Series. Springer, Berlin (2009)
15. Patton, A.J.: Volatility forecast comparison using imperfect volatility proxies. J. Econometrics **160**(1), 246–256 (2011)
16. Pesaran, M.H., Schleicher, C., Zaffaroni, P.: Model averaging in risk management with an application to futures markets. J. Empir. Finance **16**(2), 280–305 (2009)

Exploring Compositional Data with the Robust Compositional Biplot

Karel Hron and Peter Filzmoser

Abstract

Loadings and scores of principal component analysis are popularly displayed together in a planar graph, called biplot, with an intuitive interpretation. In case of compositional data, multivariate observations that carry only relative information (represented usually in proportions or percentages), principal component analysis cannot be used for the original compositions. They first need to be transformed using the centered logratio (clr) transformation. If outlying observations occur in compositional data, even the clr (compositional) biplot can lead to useless conclusions. A robust alternative biplot can be computed by using the isometric logratio (ilr) transformation, and by robustly estimating location and covariance. The robust compositional biplot has a big potential in many applications, like in geology, analytical chemistry, or social sciences.

Keywords

Compositional data • Compositional biplot • Principal component analysis • Robust statistics

K. Hron (✉)
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, Olomouc, Czech Republic

Department of Geoinformatics, Faculty of Science, Palacký University, 17. listopadu 50, Olomouc, Czech Republic
e-mail: hronk@seznam.cz

P. Filzmoser
Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstrasse 8-10, Vienna, Austria
e-mail: P.Filzmoser@tuwien.ac.at

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_1, Published online: 28 October 2014

1 Compositional Data and Their Geometry

Compositional data frequently occur in many applied fields. They are characterized by the fact that not the absolute reported information of the variables is relevant, but their ratios contain the important information [15]. When analyzing a chemical composition of a rock, not the absolute values of the mass of the compounds (which depend on the size of the sample), but ratios provide a relevant picture of the multivariate data structure. Compositional data (or compositions for short) are popularly represented by proportions or percentages, i.e. as data with a constant sum constraint. Any reasonable analysis of compositions should follow properties like scale invariance (the information in a composition does not depend on the particular units in which the composition is expressed) and subcompositional coherence (information conveyed by a full composition should not be in contradiction with that coming from a sub-composition), see e.g. [5] for details. Since the specific properties of compositions naturally induce their own geometry (the Aitchison geometry [7]) on the simplex, i.e. the sample space of compositions, the main effort is devoted to express the compositions in orthonormal coordinates, where the usual Euclidean rules already hold [6], and accommodate the standard statistical methods for their analysis. Namely, standard statistical methods completely ignore the above requirements since they rely on the usual Euclidean geometry in the real space [4]. Because all the relevant information in compositional data is contained in ratios between the parts, it is natural that zero compositional parts are not allowed for the analysis. According to the origin of zero values, either as a result of an imprecise measurement of a trace element in the composition (i.e. rounding zeros) or the result of structural processes (structural zeros), special care has to be taken prior to a further processing of the observations [3, 14].

The Aitchison geometry forms a Euclidean vector space of dimension $D-1$, that makes it possible to express the compositions in coordinates with respect to an orthonormal basis on the simplex. The corresponding mapping $h : \mathcal{S}^D \to \mathbf{R}^{D-1}$, that results in the real vector $h(\mathbf{x}) = \mathbf{z} = (z_1, \ldots, z_{D-1})'$, moves the Aitchison geometry to the standard Euclidean geometry in the real space isometrically. For this reason the mapping h is usually referred to as the isometric logratio (ilr) transformation [7]. Among infinitely many possibilities how to form the orthonormal basis on the simplex and construct the orthonormal coordinates, one popular choice results in the $(D-1)$-dimensional real vector $\mathbf{z} = (z_1, \ldots, z_{D-1})'$,

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \; i = 1, \ldots, D-1. \tag{1}$$

Note that different ilr transformations are orthogonal rotations of each other [7].

For most statistical methods, an interpretation of the compositional data analysis in orthonormal coordinates is fully satisfactory. An exception is the biplot of principal component analysis which is related to the centered logratio (clr) transformation [1], resulting for a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ in a real vector

$$\mathbf{y} = (y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)'.$$

Elements of $\mathbf{y}$ represent coefficients with respect to a generating system of compositions, i.e. the covariance matrix of a random composition $\mathbf{y}$ is positive semidefinite. Consequently, the clr transformed data are not appropriate for a robust statistical analysis, because the popular robust estimators can cope just with regular observations. Fortunately, there exists a linear relation between the clr coefficients and orthonormal coordinates [like those from Eq. (1)], $\mathbf{y} = \mathbf{V}\mathbf{z}$. The columns of the $D \times (D-1)$-matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ are formed by the clr transformation of the orthonormal basis vectors, resulting in coordinates $\mathbf{z}$, concretely

$$\mathbf{v}_{D-i} = \sqrt{\frac{i}{i+1}} \left(0, \ldots, 0, 1, -\frac{1}{i}, \ldots, -\frac{1}{i} \right)', \; i = 1, \ldots, D-1.$$

The above properties of the matrix $\mathbf{V}$ imply isometry of the clr transformation.

Although measures of location and variability of a random composition can even be expressed directly on the simplex, it is usually preferred to capture location and variability of compositions directly in coordinates using the expectation $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. If a sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is given for the coordinate $\mathbf{z}$, one usually arrives at the arithmetic mean $\bar{\mathbf{z}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i$ and the sample covariance matrix $\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})'$.

However, also in the case of compositional data, outlying observations may completely destroy results of a statistical analysis, compared to those obtained from the homogeneous majority of observations in a data set. In addition, the specific geometry of compositional data induces a different view of outliers compared to the usual case. For example, now obviously an observation with high absolute values on the compounds (parts) must not necessarily be an outlier, if the corresponding ratios between its parts follow the dominant data behavior. For this reason, not only the classical statistical methods, but even the robust ones cannot be applied directly to raw compositional data. Particularly, this would lead to a mismatch of regular and outlying observations.

Because of different representations of compositions in coordinates, the affine equivariance of the corresponding robust estimators is crucial. The location estimator $\mathbf{t}$ and the covariance estimator $\mathbf{C}$ are called affine equivariant, if for a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of n observations in $\mathbf{R}^{D-1}$, any nonsingular $(D-1) \times (D-1)$ matrix $\mathbf{A}$ and for any vector $\mathbf{b} \in \mathbf{R}^{D-1}$ the conditions

$$\mathbf{t}(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{b}) = \mathbf{A}\mathbf{t}(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{b},$$
$$\mathbf{C}(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{b}) = \mathbf{A}\mathbf{C}(\mathbf{x}_1, \ldots, \mathbf{x}_n)\mathbf{A}'$$

are fulfilled.

In the following the Minimum Covariance Determinant (MCD) estimator [13] is used, which is advantageous in particular for computational reasons [16]. The MCD estimator shares the property of affine equivariance for both the resulting location and covariance estimator. Conceptually, it looks for a subset h out of n observations with the smallest determinant of their sample covariance matrix. A robust estimator of location is the arithmetic mean of these observations, and a robust estimator of covariance is the sample covariance matrix of the h observations, multiplied by a factor for consistency at normal distribution. The subset size h can vary between half the sample size and n. It will determine the robustness of the estimates and also their efficiency.

2 Principal Component Analysis and the Compositional Biplot

Principal component analysis (PCA) cannot directly be used for raw compositional data; in addition, the proper estimation of location ($\mathbf{t}$) and covariance ($\mathbf{C}$) plays an important role. Let $\mathbf{C} = \mathbf{G}_\mathbf{z}\mathbf{L}\mathbf{G}_\mathbf{z}'$ be a spectral decomposition of the estimated covariance matrix $\mathbf{C}$, with the diagonal matrix $\mathbf{L}$ of eigenvalues and the matrix $\mathbf{G}_\mathbf{z}$ of eigenvectors of $\mathbf{C}$. Then PCA results in a linear transformation

$$\mathbf{z}_i^* = \mathbf{G}_\mathbf{z}'(\mathbf{z}_i - \mathbf{t}), \tag{2}$$

of the coordinates into new variables (principal components) such that the first principal component has the largest possible variance (accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. Although both scores $\mathbf{z}_i^*$ and loadings (columns of the matrix $\mathbf{G}_\mathbf{z}$) of the principal components could also be interpreted directly in orthonormal coordinates, it is rather common to transform the loadings back to the clr space, $\mathbf{G}_\mathbf{y} = \mathbf{V}\mathbf{G}_\mathbf{z}$, where the matrix $\mathbf{V}$ comes from the linear relationship between ilr and clr transformations (see Sect. 1) and the affine equivariance property of the MCD estimator is utilized [9]. The scores in the clr space are identical to the scores of the ilr space, except that the additional last column of the clr score matrix has entries of zero. Finally, the transformed loadings and scores are used to construct the biplot of clr transformed compositional data [2, also referred to as "compositional biplot"]. Although the purpose of the compositional biplot is the same as for the standard covariance biplot [11], i.e. to provide a planar graph that represents a rank-two approximation of both the observations (PCA scores, plotted as points) and variables (loadings, rays) of multivariate data, its interpretation is different: The main interest

is in the links (distances between vertices of the rays); concretely, for the rays i and j $(i, j = 1, \ldots, D)$ the link approximates the log-ratio variance $\mathrm{var}(\ln \frac{x_i}{x_j})$, forming the variation matrix [1]

$$\mathbf{T} = \left(\mathrm{var}(\ln \frac{x_i}{x_j}) \right)_{i,j=1}^{D}.$$

This matrix can be estimated in a classical or in a robust way, by using the relation

$$\mathbf{T} = \mathbf{J}\mathrm{diag}(\mathbf{VCV}') + \mathrm{diag}(\mathbf{VCV}')\mathbf{J} - 2\mathbf{VCV}',$$

where $\mathbf{J}$ denotes a $D \times D$ matrix of ones and $\mathbf{C}$ stands once again for the estimated covariance matrix in orthonormal coordinates (preferably the robust one using the MCD estimator).

Hence, from the form of the log-ratio variances we can see that when the vertices coincide, or nearly so, then the ratio between x_i and x_j is constant, or nearly so. Consequently, this characteristic replaces the thinking in terms of correlation coefficients between two coordinates (standard variables). In addition, directions of the rays in the biplot signalize where observations with dominance of the corresponding compositional part are located.

3 Example

The theoretical results are applied to a real-world compositional data set, describing relative contributions of six main leasure time activities (sport, reading, TV, PC, music, outdoor activities) of 30 younger school age children (10–11 years old); the original data (expressed in minutes) are quoted in [17]. The corresponding robust compositional biplot is displayed in Fig. 1. From the data structure, some outlying observations are clearly detectable (like 13, 16, 24, 25); nevertheless, they do not influence the resulting loadings and scores of robust principal components in the clr space. From the links between the vertices we can see that TV watching and working with PC are quite strongly related, a similar conclusion can be observed also from the link between sport and the other outdoor activities. This is quite a logical output as both couples of leasure time activities are of similar nature and one can thus expect a kind of stable proportional distribution between them. On the other hand, music obviously represents quite an exceptional leasure time activity, not related to the others.

For the sake of comparison, we form also the corresponding classical compositional biplot by replacing the robust estimates of location and covariance in orthonormal coordinates by the arithmetic mean and the sample covariance matrix, see Fig. 2. The first interesting feature to be observed from Fig. 2 is that the first two classical principal components explain just 57.17 % of the total variability of the compositional data set, compared to 75.66 % for the robust case. Only the

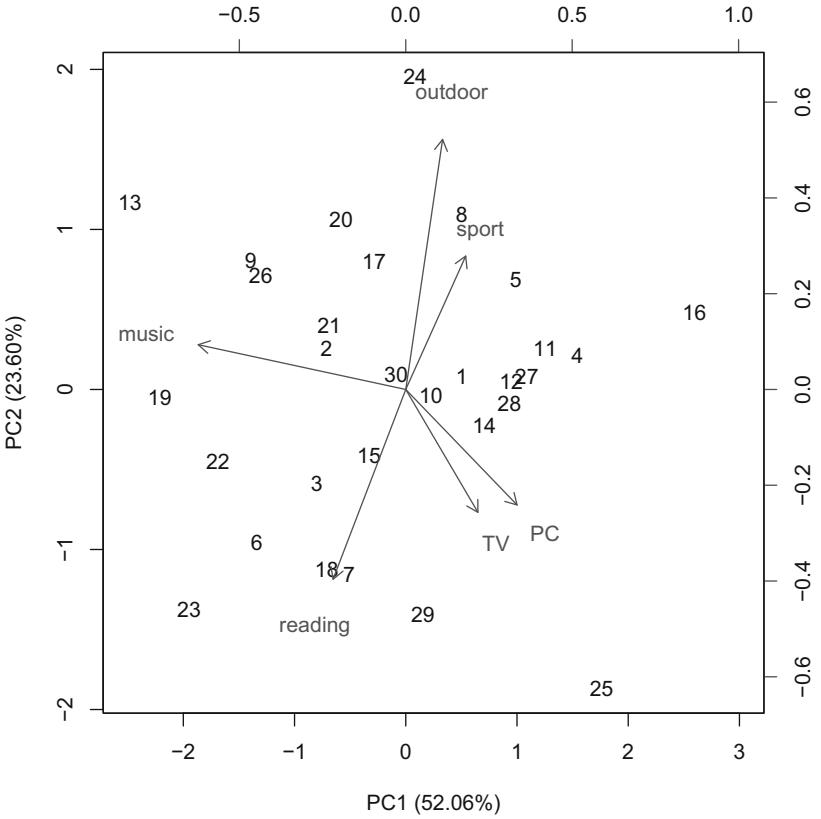

Fig. 1 Robust compositional biplot of leasure time activities of younger school age children

two compositional parts *music*, *reading* are well reflected by this biplot, but the information of the other parts is poorly represented. This is due to outliers which have attracted the first two principal component directions. Also the positions of outlier observations changed comparing to the robust compositional biplot. While the role of 13, 16, 25 remained unchanged, the observation 24 moved into the main data cloud. On the other hand, new "outliers" arise (like 1, 7, 9, 17), nevertheless, their position is rather driven by the true outliers, detectable using the robust version of the biplot. Thus, this biplot does not provide information of the multivariate data structure of the data majority, but is heavily influenced by some outliers. An interpretation would thus also be misleading, and it is even counter-intuitive.

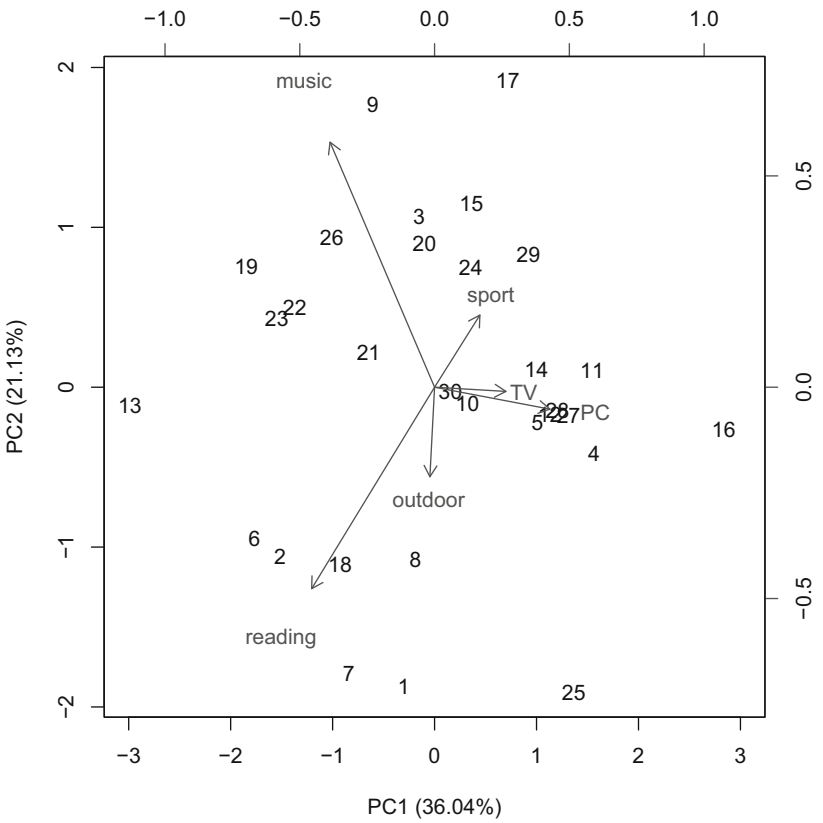

Fig. 2 Classical compositional biplot of leasure time activities of younger school age children

Conclusions

The robust compositional biplot allows to display the inherent multivariate structure of compositional data in form of a planar graph, even in presence of outliers. The robust compositional biplot can also be used for the interpretation of outlying observations [10]. Although the clr transformation of compositions is still preferable for this purpose, as an alternative also orthonormal coordinates may be used to construct a compositional biplot (even with the standard interpretation according to [11]), if some prior information the studied problem is available. For example, one possibility is to construct the ilr coordinates with interpretation in sense of balances between groups of compositional parts, see [8] for details. The compositional biplot (and preferably its robust version) can be applied to real-world problems from various applied fields, and the obtained results usually nicely follow previous expert knowledge [12].

Acknowledgements The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness—European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic) and the grant IGA PrF 2014 028 Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman and Hall, London (1986)
2. Aitchison, J., Greenacre, M.: Biplots of compositional data. Appl. Stat. **51**, 375–392 (2002)
3. Aitchison, J., Kay, J.W.: Possible solutions of some essential zero problems in compositional data analysis. In: Proceedings of the CoDaWork 2003 Conference, University of Girona, Girona (2003)
4. Eaton, M.: Multivariate Statistics: A Vector Space Approach. Wiley, New York (1983)
5. Egozcue, J.J.: Reply to "On the Harker Variation Diagrams;..." by J.A. Cortés. Math. Geosci. **41**, 829–834 (2009)
6. Egozcue, J.J., Pawlowsky-Glahn, V.: Simplicial geometry for compositional data. In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds.) Compositional Data in the Geosciences: From Theory to Practice, pp. 145–160. Geological Society, London (2006)
7. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric logratio transformations for compositional data analysis. Math. Geol. **35**, 279–300 (2003)
8. Egozcue, J.J., Pawlowsky-Glahn, V.: Groups of parts and their balances in compositional data analysis. Math. Geol. **37**, 795–828 (2005)
9. Filzmoser, P., Hron, K., Reimann, C.: Principal component analysis for compositional data with outliers. Environmetrics **20**, 621–635 (2009)
10. Filzmoser, P., Hron, K., Reimann, C.: Interpretation of multivariate outliers for compositional data. Comput. Geosci. **39**, 77–85 (2012)
11. Gabriel, K.R.: The biplot graphic display of matrices with application to principal component analysis. Biometrika **58**, 453–467 (1971)
12. Hron, K., Jelínková, M., Filzmoser, P., Kreuziger, R., Bednář, P., Barták, P.: Statistical analysis of wines using a robust compositional biplot. Talanta **90**, 46–50 (2012)
13. Maronna, R., Martin, R.D., Yohai, V.J.: Robust Statistics: Theory and Methods. Wiley, New York (2006)
14. Martín-Fernández, J.A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J.: Model-based replacement of rounded zeros in compositional data: classical and robust approaches. Comput. Stat. Data Anal. **56**, 2688–2704 (2012)
15. Pawlowsky-Glahn, V., Buccianti, A.: Compositional Data Analysis: Theory and Applications. Wiley, Chichester (2011)
16. Rousseeuw, P., Van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics **41**, 212–223 (1999)
17. Těthalová, L.: Statistical analysis of compositional data using the CoDaPack software (in Czech). Bachelor thesis, Palacký University (2013)

Sparse Orthogonal Factor Analysis

Kohei Adachi and Nickolay T. Trendafilov

Abstract

We propose a sparse orthogonal factor analysis (SOFA) procedure in which the optimal loadings and unique variances are estimated subject to additional constraint which directly requires some factor loadings to be exact zeros. More precisely, the constraint specifies the required number of zero factor loadings without any restriction on their locations. Such loadings are called sparse which gives the name of the method. SOFA solutions are obtained by minimizing a FA loss function under the sparseness constraint making use of an alternate least squares algorithm. We further present a sparseness selection procedure in which SOFA is performed repeatedly by setting the sparseness at each of a set of feasible integers. Then, the SOFA solution with the optimal sparseness can be chosen using an index for model selection. This procedure allows us to find the *optimal* orthogonal confirmatory factor analysis model among all possible models. SOFA and the sparseness selection procedure are assessed by simulation and illustrated with well known data sets.

Keywords

Confirmatory factor analysis • Factor analysis • Optimal sparseness selection • Sparse loadings • Sparse principal component analysis

K. Adachi (✉)
Graduate School of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita, Osaka 565-0871, Japan
e-mail: adachi@hus.osaka-u.ac.jp

N.T. Trendafilov
Department of Mathematics and Statistics, The Open University, Buckinghamshire, UK

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_2, Published online: 28 October 2014

1 Introduction

Factor analysis (FA) is a model that aims to explain the interrelationships among observed variables by a small number of latent variables called common factors. The relationships of the factors to observed variables are described by a factor loading matrix. FA is classified as exploratory (EFA) or confirmatory (CFA). In EFA, the loading matrix is unconstrained and has rotational freedom which is exploited to rotate the matrix so that some of its elements approximate zero. In CFA, some loadings are constrained to be zero and the loading matrix has no rotational freedom [9].

One refers to a loading matrix with a number of exactly zero elements as being *sparse*, which is an indispensable property for loadings to be interpretable. In EFA, a loading matrix is rotated toward a sparse matrix, but the literal sparseness is not attained, since rotated loadings cannot exactly be equal to zero. Thus, the user must decide which of them can be viewed as approximately zeros. On the other hand, some loadings are fixed exactly to zero in CFA. However, the problem in CFA is that the number of zero loadings and their locations must be chosen by users. That is, the users' subjective decision is needed in both EFA and CFA.

In order to overcome these difficulties, we propose a new FA procedure, which is neither EFA nor CFA. The optimal orthogonal factor solution is estimated such that it has a sparse loading matrix with a suitable number of zero elements. Note that, their locations are also estimated in an optimal way. The procedure to be proposed consists of the following two stages:

(a) The optimal solution is obtained for a specified number of zero loadings.
(b) The optimal number of zero loadings is selected among possible numbers.

Stages (a) and (b) would be described in Sects. 2–4, respectively.

In the area of principal component analysis (PCA), many procedures, called sparse PCA, have been proposed in the last decade (e.g. [8, 13, 16]). As in our FA procedure, they obtain sparse loadings. However, besides the difference between PCA and FA, our approach does not rely on penalty functions, which is the standard way to induce sparseness in the existing sparse PCA.

2 Sparse Factor Problem

The main goal of FA is to estimate the p-variables $\times m$-factors matrix $\mathbf{\Lambda}$ containing loadings and the $p \times p$ diagonal matrix $\mathbf{\Psi}^2$ including unique variances from the n-observation $\times p$-variables ($n > p$) column-centered data matrix $\mathbf{X}$. For this goal, FA can be formulated by a number of different objective functions, among which we choose the least squares function

$$f = ||\mathbf{X} - \mathbf{F}\mathbf{\Lambda}' - \mathbf{U}\mathbf{\Psi}||^2 \tag{1}$$

recently utilized in several works [1,4,14,15]. Here, $\mathbf{F}$ is the $n \times m$ matrix containing common factor scores and $\mathbf{U}$ is the $n \times p$ matrix of unique factor scores. The factor score matrices are constrained to satisfy

$$n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, n^{-1}\mathbf{U}'\mathbf{U} = \mathbf{I}_p, \text{ and } \mathbf{F}'\mathbf{U} = {}_m\mathbf{O}_p \tag{2}$$

with $\mathbf{I}_m$ the $m \times m$ identity matrix and ${}_m\mathbf{O}_p$ the $m \times p$ matrix of zeros.

We propose to minimize (1) over $\mathbf{F}$, $\mathbf{U}$, $\mathbf{\Lambda}$, and Ψ subject to (2) and

$$SP(\mathbf{\Lambda}) = q, \tag{3}$$

where $SP(\mathbf{\Lambda})$ expresses the sparseness of $\mathbf{\Lambda}$, i.e., the number of its elements being zero, and q is a specified integer.

The reason for our choosing loss function (1) is that we can define

$$\mathbf{A} = n^{-1}\mathbf{X}'\mathbf{F} \tag{4}$$

to decompose (1) as

$$f = ||\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\mathbf{\Psi} - (\mathbf{F}\mathbf{\Lambda}' - \mathbf{F}\mathbf{A}')||^2 = ||\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\Psi||^2 + n||\mathbf{\Lambda} - \mathbf{A}||^2. \tag{1'}$$

This equality is derived from the fact that $(\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\Psi)'(\mathbf{F}\mathbf{\Lambda}' - \mathbf{F}\mathbf{A}') = n\mathbf{A}\mathbf{\Lambda}' - n\mathbf{A}\mathbf{A}' - n\mathbf{A}\mathbf{\Lambda}' + n\mathbf{A}\mathbf{A}' = {}_p\mathbf{O}_p$ is given using (2) and (4). In (1′) only a simple function $||\mathbf{\Lambda} - \mathbf{A}||^2$ is relevant to $\mathbf{\Lambda}$ and thus can be easily minimized over $\mathbf{\Lambda}$ subject to (3) as seen in the next section. It is difficult for other objective functions of FA to be rewritten into simple forms as (1′). For example, the likelihood function for FA includes the determinant of a function of $\mathbf{\Lambda}$ which is difficult to handle.

3 Algorithm

For minimizing (1) subject to (2) and (3), we consider alternately iterating the update of each parameter matrix.

First, let us consider updating $\mathbf{\Lambda}$ so that (1) or (1′) is minimized subject to (3) while $\mathbf{F}$, $\mathbf{U}$, and Ψ are kept fixed. This amounts to minimizing $g(\mathbf{\Lambda}) = ||\mathbf{\Lambda} - \mathbf{A}||^2$ under (3), since of (1′). Using $\mathbf{\Lambda} = (\lambda_{ij})$ and $\mathbf{A} = (a_{ij})$, we can rewrite $g(\mathbf{\Lambda})$ as

$$g(\mathbf{\Lambda}) = \sum_{(i,j)\in\mathbf{N}} a_{ij}^2 + \sum_{(i,j)\in\mathbf{N}^\perp} (\lambda_{ij} - a_{ij})^2 \geq \sum_{(i,j)\in\mathbf{N}} a_{ij}^2, \tag{5}$$

where $\mathbf{N}$ denotes the set of the q pairs of (i, j) for the loadings λ_{ij} to be zero and $\mathbf{N}^\perp$ is the complement to $\mathbf{N}$. The inequality in (5) shows that $g(\mathbf{\Lambda})$ attains its lower limit $\Sigma_{(i,j)\in N} a_{ij}^2$ when the loading λ_{ij} with $(i, j) \in \mathbf{N}^\perp$ is set equal to a_{ij}. Further,

the limit $\Sigma_{(i,j)\in N} a_{ij}^2$ is minimal when $\mathbf{N}$ contains the (i, j) for the q smallest a_{ij}^2 among all squared elements of $\mathbf{A}$. The optimal $\mathbf{\Lambda} = (\lambda_{ij})$ is thus given by

$$\lambda_{ij} = \begin{cases} 0 \ \text{ iff} \, a_{ij}^2 \leq a_{[q]}^2 \\ a_{ij} \ \ \text{otherwise} \end{cases} \tag{6}$$

with $a_{[q]}^2$ the q-th smallest value among all a_{ij}^2.

Next, let us consider the update of the diagonal matrix Ψ. Substituting (2) in (1) simplifies the objective function to

$$f = n\text{tr}\mathbf{S} + n\text{tr}\mathbf{\Lambda}\mathbf{\Lambda}' + n\text{tr}\mathbf{\Psi}^2 - 2n\text{tr}\mathbf{X}'\mathbf{F}\mathbf{\Lambda}' - 2\text{tr}\mathbf{X}'\mathbf{U}\mathbf{\Psi} \tag{1''}$$

with $\mathbf{S} = n^{-1/2}\mathbf{X}'\mathbf{X}$ the sample covariance matrix. Since (1″) can further be rewritten as $||n^{1/2}\mathbf{\Psi} - n^{-1/2}\text{diag}(\mathbf{X}'\mathbf{U})||^2 + c$ with c a constant irrelevant to $\mathbf{\Psi}$, the minimizer is found to be given by

$$\mathbf{\Psi} = \text{diag}(n^{-1}\mathbf{X}'\mathbf{U}), \tag{7}$$

when $\mathbf{F}, \mathbf{U}$, and $\mathbf{\Lambda}$ are considered fixed.

Finally, let us consider minimizing (1) over $n \times (m + p)$ block matrix $[\mathbf{F}, \mathbf{U}]$ subject to (2) with $\mathbf{\Psi}$ and $\mathbf{\Lambda}$ kept fixed. We note that (1″) is rewritten as $f = c^* - 2n\text{tr}\mathbf{B}'\mathbf{X}'[\mathbf{F}, \mathbf{U}]$ with $\mathbf{B} = [\mathbf{\Lambda}, \mathbf{U}]$ an $p \times (m + p)$ matrix and c^* a constant irrelevant to $[\mathbf{F}, \mathbf{U}]$. As proved in Appendix 1, f is minimized for

$$n^{-1}\mathbf{X}'[\mathbf{F}, \mathbf{U}] = \mathbf{B}'^{+}\mathbf{Q}\mathbf{\Delta}\mathbf{Q}', \tag{8}$$

where $\mathbf{B}'^{+}$ is the Moore-Penrose inverse of $\mathbf{B}'$ and $\mathbf{Q}\mathbf{\Delta}\mathbf{Q}'$ is obtained through the eigenvalue decomposition (EVD) of $\mathbf{B}'\mathbf{S}\mathbf{B}$:

$$\mathbf{B}'\mathbf{S}\mathbf{B} = \mathbf{Q}\mathbf{\Delta}^2\mathbf{Q}', \tag{9}$$

with $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_p$ and Δ^2 the positive definite diagonal matrix. Rewriting (8) as $[n^{-1}\mathbf{X}'\mathbf{F},\ n^{-1}\mathbf{X}'\mathbf{U}] = \mathbf{B}'^{+}\mathbf{Q}\mathbf{\Delta}\,Q'$ and comparing it with (4) and (7), one finds:

$$\mathbf{A} = \mathbf{B}'^{+}\mathbf{Q}\mathbf{\Delta}\mathbf{Q}'\mathbf{H}_m \tag{4'}$$

$$\mathbf{\Psi} = \text{diag}(\mathbf{B}'^{+}\mathbf{Q}\mathbf{\Delta}\mathbf{Q}'\mathbf{H}^p) \tag{7'}$$

where $\mathbf{H}_m = [\mathbf{I}_m,\ {}_m\mathbf{O}_p]'$ and $\mathbf{H}^p = [{}_p\mathbf{O}_m,\ \mathbf{I}_p]'$.

The above equations show that $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ can be updated if only the sample covariance matrix $\mathbf{S}(= n^{-1}\mathbf{X}'\mathbf{X})$ is available. In other words, the updating of $[\mathbf{F}, \mathbf{U}]$ can be avoided when the original data matrix $\mathbf{X}$ is not given, That is, the decomposition (9) gives the matrices $\mathbf{Q}$ and $\mathbf{\Delta}$ needed in (4′) and (7′), with (4′) being used for (6). Further, the resulting loss function value can be computed without the

use of $\mathbf{X}$: (6) implies $\mathbf{\Lambda}'\mathbf{A} = \mathbf{\Lambda}'\mathbf{\Lambda}$, and substituting this, (4), and $\mathbf{B} = [\mathbf{\Lambda}, \mathbf{U}]$ into (1″) leads to $f = n\mathrm{tr}\mathbf{S} + n\mathrm{tr}\mathbf{\Lambda}\mathbf{\Lambda}' - 2n\mathrm{tr}\mathbf{\Lambda}'\mathbf{A} - n\mathrm{tr}\mathbf{\Psi}^2 = n\{\mathrm{tr}\mathbf{S} - \mathrm{tr}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2)\} = n(\mathrm{tr}\mathbf{S} - \mathrm{tr}\mathbf{B}\mathbf{B}')$. Then, the standardized loss function value

$$f_S(\mathbf{B}) = 1 - \mathrm{tr}\mathbf{B}\mathbf{B}'/\mathrm{tr}\mathbf{S}, \tag{10}$$

which takes a value within [0,1], can be used for convenience instead of f.

The optimal $\mathbf{B} = [\mathbf{\Lambda}, \mathbf{\Psi}]$ is thus given by the following algorithm:

Step 1. Initialize $\mathbf{\Lambda}$ and $\mathbf{\Psi}$.
Step 2. Set $\mathbf{B} = [\mathbf{\Lambda}, \mathbf{\Psi}]$ to perform EVD (9).
Step 3. Obtain $\mathbf{A}$ by (4′) to update $\mathbf{\Lambda}$ with (6).
Step 4. Update $\mathbf{\Psi}$ with (7′).
Step 5. Finish if convergence is reached; otherwise, go back to Step 2.

The convergence of the updated parameters in Step 5 is defined as the decrease of (10) being less than 0.1^7. To avoid missing the global minimum, we run the algorithm multiple times with random start in Step 1. The procedure for selection of the optimal solution is described in Appendix 2. We denote the resulting solution of $\mathbf{B}$ as $\hat{\mathbf{B}}_q = [\hat{\mathbf{\Lambda}}_q, \hat{\mathbf{\Psi}}_q]$, where the subscript q indicates the particular number of zeros used in (3).

4 Sparseness Selection

Sparseness can be restated as parsimony: the greater $SP(\mathbf{\Lambda})$ implies that fewer parameters are to be estimated and the resulting loss function value is greater. Thus, the sparseness selection means to choose a FA model with the optimal combination of the attained loss function value and parsimony. For such model selection, we can use the information criteria [10] which are defined using maximum likelihood (ML) estimates. Although ML method is not used in our algorithm, we assume that $\hat{\mathbf{B}}_q = [\hat{\mathbf{\Lambda}}_q, \hat{\mathbf{\Psi}}_q]$ is equivalent to the ML-CFA solution which maximizes log likelihood $L(\mathbf{\Lambda}, \mathbf{\Psi}) = -0.5n\{\log|\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2| + \mathrm{tr}\mathbf{S}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2)^{-1}\}$ with the locations of the zero loadings constrained to be those of $\hat{\mathbf{\Lambda}}_q$. This assumption would be validated empirically in the next section. Under this assumption, we propose to use an information criterion BIC [10] for choosing the optimal q. BIC can be expressed as

$$BIC(q) = -2L(\hat{\mathbf{\Lambda}}_q, \hat{\mathbf{\Psi}}_q) - q\log n + c^{\#} \tag{11}$$

for $\hat{\mathbf{B}}_q$ with $c^{\#}$ a constant irrelevant to q. The optimal sparseness is thus defined as

$$\hat{q} = \underset{q_{\min} \le q \le q_{\max}}{\arg\min} \; BIC(q) \tag{12}$$

and $\hat{\mathbf{B}}_{\hat{q}}$ is chosen as the final solution $\hat{\mathbf{B}}$. Here, we set $q_{\min} = m(m-1)/2$, since it prevents $\mathbf{\Lambda}$ from rotational indeterminacy if q goes below it. On the other hand, we set $q_{\max} = p(m-1)$, since it prevents $\mathbf{\Lambda}$ from having an empty column if q were greater than the limit.

5 Simulation Study

We performed a simulation study to assess the proposed procedure with respect to (1) identifying the true sparseness and the locations of zero loadings; (2) goodness of the recovery of parameter values; (3) sensitivity to local minima; (4) whether SOFA solutions are equivalent to the solutions of the ML-CFA procedure with the locations of the zero elements in $\mathbf{\Lambda}$ set to those obtained by SOFA.

We used the five types of the true $\mathbf{\Lambda}$ shown in Table 1, which are desired to be possessed by FA solutions. The first three types have simple structure, while the remaining two have bi-factor simple structure as defined by [6]. For each type, we generated 40 sets of $\{\mathbf{\Lambda}, \mathbf{\Psi}, \mathbf{S}\}$ by the following steps: (1) each diagonal element of $\mathbf{\Psi}$ was set to $u(0.1^{1/2}, 0.7^{1/2})$, where $u(\alpha, \beta)$ denotes a value drawn from the uniform distribution of the range $[\alpha, \beta]$. (2) A nonzero value in $\mathbf{\Lambda}$ was set to $u(0.4, 1)$, while an element denoted by "r" in Table 1 was randomly set to zero or $u(0.4, 1)$. (3) $\mathbf{\Lambda}$ was normalized so as to satisfy $\mathrm{diag}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2) = \mathbf{I}_p$. (4) Setting $n = 200p$, we sampled each row of $\mathbf{X}$ from the centered p-variate normal distribution with its covariance matrix $\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2$. (5) Inter-variable correlation matrix $\mathbf{S}$ was obtained from $\mathbf{X}$. For the resulting data sets, we carried out SOFA: its algorithm was run multiple times for each of $q = q_{\min}, \ldots, q_{\max}$ until the two equivalently optimal

Table 1 Three loading matrices of simple structure (left) and two ones of bi-factor structure (right), with nonzero and zero elements denoted by # and blank cells, respectively

#	r		#	r			#		r	r		#	#			#	#		
#	r		#				#		r	r		#	#			#	#		
#			#				#		r	r		#	#			#	#		
#			#				#		r	r		#	r			#	r		
r	#		r	#			r	#		r		#	r			#	r	r	
r	#		r	#			r	#		r		#	r			#		#	
	#		r	#			r	#		r		#		#		#		#	
	#		r	#			r	#		r		#		#		#		#	
			r	r	#		r	r	#			#		#		#		r	
				r	#		r	r	#			#		r		#		r	r
				r	#		r	r	#			#		r		#			#
				r	#		r	r	#			#		r		#			#
								r	r	#						#			#
								r	r	#						#			r
								r	r	#						#	r		r
								r	r	#									

Table 2 Percentiles of index values for assessing the SOFA solutions

		(B) Rate		(C) Difference to the true value		(D) Difference to ML-CFA	
Percentile	(A) BES	R_{00}	$R_{\#\#}$	$\mathbf{\Lambda}$	$\mathbf{\Psi}^2$	$\mathbf{\Lambda}$	$\mathbf{\Psi}^2$
5	−0.133	0.843	0.972	0.013	0.023	0.002	0.004
25	−0.031	0.968	1.000	0.017	0.032	0.003	0.005
50	0.000	1.000	1.000	0.021	0.038	0.004	0.006
75	0.000	1.000	1.000	0.026	0.046	0.006	0.008
95	0.000	1.000	1.000	0.040	0.056	0.009	0.011

solutions are found by the procedure in Appendix 2. As done there, we use L_q for the number of runs necessitated.

To asses the sensitivity of SOFA to local minima, we counted L_q and averaged it over q for each data set. The sensitivity is indicated by L_q as described in Appendix 2. The quartiles of the averaged L_q values over 200 data sets were 89, 120, and 155: the second quartile 120 implies that the $120 - 2 = 118$ solutions (except two equivalently optimal solutions) are local minimizers among 120 solutions for a half of data sets. This demonstrates high sensitivity to local minima. Nevertheless, good performances of the proposed procedure are shown next.

For each of 200 data sets, we obtained some index values to assess the correctness of the $\hat{q}$ selected by BIC and the corresponding optimal solution $\hat{\mathbf{B}}_{\hat{q}} = [\hat{\mathbf{\Lambda}}_{\hat{q}}, \hat{\mathbf{\Psi}}_{\hat{q}}]$. The percentiles of the index values over the 200 cases are shown in Panels (A), (B), and (C) of Table 2. The first index is BES $= (\hat{q} - q)/q$ which assesses the relative bias of the estimated sparseness from the true q. The percentiles in Panel (A) show that sparseness was satisfactorily estimated, though it tended to be underestimated. The indices R_{00} and $R_{\#\#}$in Panel (B) are the rates of the zero and non-zero elements in the true $\mathbf{\Lambda}$ correctly identified by $\hat{\mathbf{\Lambda}}$. Panel (B) shows that non-zero elements have been exactly identified. The indices in Panel (C) are mean absolute differences $||\hat{\mathbf{\Lambda}}_{\hat{q}} - \mathbf{\Lambda}||_1/(pm)$ and $||\hat{\mathbf{\Psi}}^2_{\hat{q}} - \mathbf{\Psi}^2||_1/p$, where $||\cdot||_1$ denotes the sum of the absolute values of the elements of the argument. The percentiles of the differences show that the parameter values were recovered very well.

For each data set, ML-CFA was also performed with the locations of the zero loadings fixed at those in $\hat{\Lambda}_{\hat{q}}$. For ML-CFA, we used the EM algorithm with [2] formulas. Let $\mathbf{\Lambda}_{\mathrm{ML}}$ and $\mathbf{\Psi}_{\mathrm{ML}}$ denote the resulting $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. Panel (D) in Table 2 shows the percentiles of $||\hat{\mathbf{\Lambda}}_{\hat{q}} - \mathbf{\Lambda}_{\mathrm{ML}}||_1/(pm)$ and $||\hat{\mathbf{\Psi}}^2_{\hat{q}} - \mathbf{\Psi}^2_{\mathrm{ML}}||_1/p$. There, we find that the differences were small enough to be ignored, which validate the use of ML-based BIC in SOFA.

6 Examples

We illustrate SOFA with two famous examples. The first one is a real data set known as [6] twenty four psychological test data, which contain the scores of $n = 145$ students for $p = 24$ problems. The correlation matrix is available in [5], p. 124.

Table 3 Solution for 24 psychological test data with empty cells standing for zero

Abilities	Variables (problems)	$\mathbf{\Lambda}$				ψ_i^2
		1	2	3	4	
Spatial perception	Visual perception	0.67				0.52
	Cubes	0.43				0.79
	Paper form board	0.52		−0.19		0.66
	Flags	0.54				0.68
Verbal processing	General information	0.56	0.59			0.31
	Paragraph comprehension	0.58	0.58			0.31
	Sentence completion	0.55	0.64			0.26
	Word classification	0.62	0.35			0.47
	Word meaning	0.59	0.60			0.26
Speed of performances	Addition	0.26	0.16	0.80		0.25
	Code	0.42		0.47	0.26	0.50
	Counting dots	0.37		0.62		0.45
	Straight-curved capitals	0.56		0.38		0.51
Memory	Word recognition	0.36			0.46	0.64
	Number recognition	0.34			0.45	0.67
	Figure recognition	0.54	−0.15		0.35	0.55
	Object-number	0.36		0.20	0.52	0.54
	Number-figure	0.45		0.27	0.33	0.59
	Figure-word	0.43			0.22	0.74
Mathematics	Deduction	0.66				0.54
	Numerical puzzles	0.58		0.30		0.55
	Problem reasoning	0.65				0.56
	Series completion	0.74				0.43
	Arithmetic problems	0.54	0.21	0.40		0.49

From the EFA solution for the matrix, [7] found bi-factor structure using their proposed bi-factor rotation with $m = 4$. We analyzed the correlation matrix by SOFA with the same number of factors. The optimal $SP(\mathbf{\Lambda}) = 48$ was found by BIC. The solution is shown in Table 3. Its first column shows the abilities made up by [5], p. 125, which are considered necessary for solving the corresponding groups of problems. This grouping can be used to give clear interpretation of $\hat{\mathbf{\Lambda}}$: the first, second, third, and fourth factors stand in turn for the general ability related to all problems, the skill of verbal processing, the speed of performances, and the accuracy of memory, respectively. It matches the bi-factor structure found by [7]. However, our result allows us to interpret the factors simply by observing the nonzero loadings, while [7] obtain reasonable interpretation only after considering the loadings greater than or equal to 0.3 in magnitude. This choice is subjective and likely to lead to suboptimal and misleading solutions.

Table 4 Solution for the box problem with empty cells standing for zero

	$\mathbf{\Lambda}$			
Variables	x	y	z	ψ_i^2
x^2	0.95			0.08
y^2		0.96		0.08
z^2			0.94	0.09
xy	0.67	0.61		0.17
xz	0.64		0.64	0.17
yz		0.66	0.63	0.15
$(x^2 + y^2)^{1/2}$	0.69	0.64		0.10
$(x^2 + z^2)^{1/2}$	0.68		0.64	0.12
$(y^2 + z^2)^{1/2}$		0.66	0.67	0.11
$2x + 2y$	0.68	0.67		0.08
$2x + 2z$	0.67		0.68	0.08
$2y + 2z$		0.66	0.68	0.09
$\log x$	0.89			0.19
$\log y$		0.87		0.23
$\log z$			0.88	0.21
xyz	0.47	0.49	0.54	0.22
$(x^2 + y^2 + z^2)^{1/2}$	0.57	0.52	0.54	0.10
e^x	0.71			0.48
e^y		0.68		0.52
e^z			0.71	0.49

The second example considers [12] box problem which gives simulated data traditionally used as a benchmark for testing FA procedures. As described in Appendix 3, we followed [12] to generate 20 variables using functions of 3×1 common factor vector $[x, y, z]'$, with the functions defined as in the first column of Table 4. Those procedures gave the correlation matrix (Table 5) to be analyzed. The ideal solution for this problem is the one such that variables load the factor(s) used for defining the variables: for example, the fourth variable should ideally load x and y. The SOFA solution with $SP(\mathbf{\Lambda}) = 27$ selected by BIC is shown in Table 4, where we find that the ideal loadings were obtained.

Table 5 Correlation coefficients multiplied by 100 for the box problem

Var.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	100																			
2	−5	100																		
3	−2	0	100																	
4	62	57	−2	100																
5	59	−3	60	42	100															
6	−4	64	59	40	41	100														
7	63	61	−2	82	42	40	100													
8	65	1	61	47	81	40	46	100												
9	−3	62	63	40	42	83	41	44	100											
10	62	63	2	87	44	45	90	48	45	100										
11	63	−2	63	44	86	42	45	89	45	46	100									
12	−4	62	63	41	40	86	40	44	90	45	45	100								
13	82	−6	1	57	60	−3	59	62	−2	59	64	−4	100							
14	−5	80	2	57	0	58	57	2	60	59	0	63	−5	100						
15	−2	1	79	1	58	60	−2	55	62	2	61	64	1	2	100					
16	41	47	50	66	72	73	61	64	65	65	68	67	43	44	50	100				
17	52	50	51	68	67	64	74	76	72	75	74	71	48	46	48	77	100			
18	74	1	−6	48	41	−5	53	47	−4	50	42	−1	52	3	−7	28	43	100		
19	−7	72	6	35	−2	51	43	0	50	41	−3	45	−7	49	5	34	37	−5	100	
20	1	4	76	2	41	48	1	48	50	4	46	50	−1	7	51	40	43	−3	8	100

7 Discussions

In this paper, we proposed a new FA procedure named SOFA (sparse orthogonal factor analysis), which is neither EFA nor CFA. In SOFA, FA loss function (1) is minimized over loadings and unique variances subject to the direct sparseness constraint on loadings. This minimization algorithm alternately estimates the locations of the zero loadings and the values of the nonzero ones. Further, the best sparseness is selected using BIC. The simulation study demonstrated that the true sparseness and parameter values are recovered well by SOFA, and the examples illustrated that SOFA produces reasonable sparse solutions.

As stated already, a weakness of the rotation methods in EFA is that the user must decide which rotated loadings can be viewed as potential zeros. Another weakness of the rotation methods is that they do not involve the original data, because the rotation criteria are functions of the loading matrix only [3]. Thus, the rotated loadings may possess structure which is not relevant to the true loadings of the underlying data. In contrast, SOFA minimizes (1) so that the FA model is optimally fitted to the data set under the sparseness constraint, and thus can find the loadings underlying the data set with a suitable sparseness.

Our proposed procedure of SOFA with the sparseness selection by BIC allows us to find an optimal orthogonal solution with the best sparseness. If one tries to find that optimal solution by CFA without any prior knowledge about the solution, CFA must be performed over all possible models, i.e., over all possible locations of q zero loadings with changing q from $q_{\min}$ to $q_{\max}$. That is, the number of the models to be tested is so enormous that it is unfeasible. An optimal model can, however, be found by our procedure. It is thus regarded as an automatic finder of an optimal orthogonal CFA model.

A drawback of SOFA is that its solutions are restricted to the orthogonal ones without inter-factor correlations. It thus remains for future studies to develop a sparse oblique FA procedure with the correlations included in parameters.

Acknowledgements The works were partially supported by Grant #4387 by The Great Britain Sasakawa Foundation.

Appendix 1: Update of $n^{-1}X'$ [F,U]

We prove that $c^* - \mathrm{tr}\mathbf{B}'\mathbf{X}'[\mathbf{F},\mathbf{U}]$ is minimized, or equivalently, $\mathrm{tr}\mathbf{B}'\mathbf{X}'[\mathbf{F},\mathbf{U}]$ is maximized, for (8) subject to (2), supposing that the rank of $\mathbf{XB}$ is p. First, let us consider maximizing $\mathrm{tr}\mathbf{B}'\mathbf{X}'[\mathbf{F},\mathbf{U}]$ under the constrains in (2) summarized in $n^{-1}[\mathbf{F},\mathbf{U}]'[\mathbf{F},\mathbf{U}] = \mathbf{I}_{m+p}$. The maximizer is given by

$$[\mathbf{F},\mathbf{U}] = n^{1/2}\mathbf{P}\mathbf{Q}' + n^{1/2}\mathbf{P}_\perp\mathbf{Q}'_\perp \tag{13}$$

through the singular value decomposition of $n \times (m + p)$ matrix $n^{-1/2}\mathbf{XB}$;

$$n^{-1/2}\mathbf{XB} = [\mathbf{P}, \mathbf{P}_\perp] \begin{bmatrix} \mathbf{\Delta} & \\ & {}_m\mathbf{O}_m \end{bmatrix} \begin{bmatrix} \mathbf{Q}' \\ \mathbf{Q}'_\perp \end{bmatrix} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}'. \tag{14}$$

Here, $[\mathbf{P}, \mathbf{P}_\perp]$ and $[\mathbf{Q}, \mathbf{Q}_\perp]$ are $n \times (p + m)$ and $p \times (p + m)$ orthonormal matrices, respectively, whose blocks $\mathbf{P}$ and $\mathbf{Q}$ consist of p columns, and $\mathbf{\Delta}$ is a $p \times p$ diagonal matrix [11]. Next, let us note that the rank of $\mathbf{XB}$ being p implies $\mathbf{B}$ being of full-row rank, which leads to $\mathbf{B}\mathbf{B}^+ = \mathbf{I}_p$. Using this fact in (14) we have $n^{-1}\mathbf{X} = n^{-1/2}\mathbf{P}\mathbf{\Delta}\mathbf{Q}'\mathbf{B}^+$, which is transposed and post-multiplied by (13) to give (8), since of $\mathbf{P}'\mathbf{P}_\perp = {}_p\mathbf{O}_{p-m}$. Further, (8) is obtained with (9) followed from (14).

Appendix 2: Multiple Runs Procedure

The initial $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ in the SOFA algorithm (Sect. 3) are chosen randomly. Each diagonal element of $\mathbf{\Psi}$ is initialized at $u(0.1^{1/2}, 0.7^{1/2})$ with $u(\alpha, \beta)$ a value drawn from the uniform distribution of the range $[\alpha, \beta]$. Each loading of $\mathbf{\Lambda} = (\lambda_{ij})$ is set to $u(0.3, 1)$, and the value $\lambda^2_{[q]}$ is obtained that is the q-th smallest among all λ^2_{ij}, which is followed by transforming the loadings with $\lambda^2_{ij} \leq \lambda^2_{[q]}$ into zeros. Further, the initial $\mathbf{\Lambda}$ is normalized so as to satisfy $\mathrm{diag}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}^2) = \mathbf{I}_p$.

Let $\mathbf{B}_{ql} = [\mathbf{\Lambda}_{ql}, \mathbf{\Psi}_{ql}]$ denote the solution of $\mathbf{B}$ resulting from the l-th run of the SOFA algorithm for $SP(\mathbf{\Lambda})$ set at a specified q, with $l = 1, \ldots, L_q$. We regard $\mathbf{B}_{ql*} = [\mathbf{\Lambda}_{ql*}, \mathbf{\Psi}_{ql*}]$ with $l^* = \arg\min_{1 \leq l \leq L_q} f_S(\mathbf{B}_{ql})$ as the optimal solution $\hat{\mathbf{B}}_q$, and define $\mathbf{B}_{ql}$ being a local minimizer as $\Delta(\mathbf{B}_{ql}, \mathbf{B}_{ql*}) = 0.5(||\mathbf{\Lambda}_{ql} - \mathbf{\Lambda}_{ql*}||_1/mp + ||\mathbf{\Psi}_{ql} - \mathbf{\Psi}_{ql*}||_1/p) > 0.1^3$, with $||\cdot||_1$ denoting the sum of the absolute values of the elements of the argument. Here, the suitable L_q (number of runs) is unknown beforehand. We thus employ a strategy in which L_q is initialized at an integer and increased until $\{\mathbf{B}_{ql};\ l = 1, \ldots, L_q\}$ include the two equivalently optimal solutions $\mathbf{B}_{ql*}$ and $\mathbf{B}_{ql^\#}$ satisfying $\Delta(\mathbf{B}_{ql*}, \mathbf{B}_{ql^\#}) \leq 0.1^3$ and $l^* = \mathrm{argmin}_{1 \leq l \leq L} f(\mathbf{\Theta}_l)$ with $l^\# \neq l^*$. This procedure is formally stated as follows:

1. Set $L_q = 50$ and obtain $l^* = \arg\min_{1 \leq l \leq L_q} f_S(\mathbf{B}_{ql})$
2. Go to 6, if $\Delta(\mathbf{B}_{ql*}, \mathbf{B}_{ql^\#}) \leq 0.1^3$ is satisfied for $l^* \neq l^\#$; otherwise, go to 3.
3. Set $L_q := L_q + 1$, and let $\mathbf{B}_{ql^\#}$ be the output from another run.
4. Exchange $\mathbf{B}_{ql*}$ for $\mathbf{B}_{ql^\#}$ if $f_S(\mathbf{B}_{ql^\#}) < f_S(\mathbf{B}_{ql*})$.
5. Go to 6 if $\Delta(\mathbf{B}_{ql*}, \mathbf{B}_{ql^\#}) \leq 0.1^3$ or $L_q = 200$; otherwise, go back to 3.
6. Finish with $\hat{\mathbf{B}}_q$ set at $\mathbf{B}_{ql*}$.

In this procedure, except $\mathbf{B}_{ql*}$ and $\mathbf{B}_{ql^\#}$, the rest $L_q - 2$ solutions are local minimizers, thus the L_q value clearly indicates the sensitivity to local minima.

Appendix 3: Box Problem Data

In the box problem, the 3×1 common factor score vector $\mathbf{f} = [x,\ y,\ z]'$ is supposed to yield 20×1 observation vector $\mathbf{x}$ with $\mathbf{x} = \boldsymbol{\phi}(x,\ y,\ z) + \boldsymbol{\Psi}\mathbf{u}$, where $\boldsymbol{\phi}(x, y, z)$ is the vector function with its 20 elements defined as in the first column of Table 4. The original [12] box data matrix is 20×20, whose rows are 20 realizations of $\mathbf{x}' = \boldsymbol{\phi}'(x, y, z)$ without unique factor $\boldsymbol{\Psi}\mathbf{u}$. Here, x, y, z was set to the lengths, widths, and heights of boxes, from which the name of the problem originates. However, the 20×20 data matrix does not suit the cases of $n > p$ considered in this paper. We thus simulated the 400×20 $\mathbf{X}$ based on $\mathbf{x} = \boldsymbol{\phi}(x, y, z) + \boldsymbol{\Psi}\mathbf{u}$ with the following steps: First, we set x, y, and z at $u(1, 10)$ to have $400 \times 20\boldsymbol{\Phi}$ whose rows are the realizations of $\boldsymbol{\phi}'(x, y, z)$. Second, we sampled each element of $\mathbf{u}$ from the standard normal distribution to have 400×20 $\mathbf{U}$ with its rows $\mathbf{u}'$ and set the diagonal elements of $\boldsymbol{\Psi}$ to $0.1^{1/2}$. Third, we standardized the columns of $\boldsymbol{\Phi}$ so that their average and variance were zero and one, and had $\mathbf{X} = \boldsymbol{\Phi} + \mathbf{U}\boldsymbol{\Psi}$ whose inter-column correlations are shown in Table 5.

Bibliography

1. Adachi, K.: Some contributions to data-fitting factor analysis with empirical comparisons to covariance-fitting factor analysis. J. Jpn. Soc. Comput. Stat. **25**, 25–38 (2012)
2. Adachi, K.: Factor analysis with EM algorithm never gives improper solutions when sample covariance and initial parameter matrices are proper. Psychometrika **78**, 380–394 (2013)
3. Browne, M.W.: An overview of analytic rotation in exploratory factor analysis. Multivariate Behav. Res. **36**, 111–150 (2001)
4. de Leeuw, J.: Least squares optimal scaling of partially observed linear systems. In: van Montfort, K., Oud, J., Satorra, A. (eds.) Recent Developments of Structural Equation Models: Theory and Applications, pp. 121–134. Kluwer Academic, Dordrecht (2004)
5. Harman, H.H.: Modern Factor Analysis, 3rd edn. University of Chicago Press, Chicago (1976)
6. Holzinger, K.J., Swineford, F.: A study in factor analysis: the stability of a bi-factor solution. University of Chicago: Supplementary Educational Monographs No. 48 (1939)
7. Jennrich, R.I., Bentler, P.M.: Exploratory bi-factor analysis. Psychometrika **76**, 537–549 (2011)
8. Jolliffe, I.T., Trendafilov, N.T., Uddin, M.: A modified principal component technique based on the LASSO. J. Comput. Graphical Stat. **12**, 531–547 (2003)
9. Mulaik, S.A.: Foundations of Factor Analysis, 2nd edn. CRC, Boca Raton (2010)
10. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**, 461–464 (1978)
11. ten Berge, J.M.F.: Least Square Optimization in Multivariate Analysis. DSWO Press, Leiden (1993)
12. Thurstone, L.L.: Multiple Factor Analysis. University of Chicago Press, Chicago (1947)
13. Trendafilov, N.T.: From simple structure to sparse components: a review. Comput. Stat. **29**, 431–454 (2014)
14. Trendafilov, N.T., Unkel, S.: Exploratory factor analysis of data matrices with more variables than observations. J. Comput. Graphical Stat. **20**, 874–891 (2011)
15. Unkel, S., Trendafilov, N.T.: Simultaneous parameter estimation in exploratory factor analysis: an expository review. Int. Stat. Rev. **78**, 363–382 (2010)
16. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. J. Comput. Graphical Stat. **15**, 265–286 (2006)

Adjustment to the Aggregate Association Index to Minimise the Impact of Large Samples

Eric J. Beh, Salman A. Cheema, Duy Tran, and Irene L. Hudson

Abstract

The past few decades have seen a great deal of attention given to the development of techniques to analysis the association between aggregated categorical data. One of the most recent additions to this analysis has been the development of the aggregate association index (AAI). One feature of the AAI is that its magnitude is affected by the sample size; as the sample size increases so too does the AAI, even when the marginal proportions remain unchanged. In this article, we propose adjustments to the AAI to overcome the effect of increasing sample size. The Adjusted AAI is shown to be more stable than the original AAI in response to any increase in the sample size. Fisher's criminal twin data (Fisher, J. R. Stat. Assoc. Ser. A **98**, 39–82, 1935) is used to demonstrate the adjustments.

Keywords

Aggregate association index • 2×2 tables • Pearson's chi-squared statistics

1 Introduction

Discussions concerning the utility of aggregate data in 2×2 contingency tables have a long history in the literature dating back to the social sciences [15]. It became a topic of significant statistical discussion when [8], p. 48, discussed aggregate data for a 2×2 table by saying "let us blot out the contents of the table, leaving only the marginal frequencies". In his discussion Fisher concluded that marginal data is "ancillary information" in terms of estimating cell frequencies. Yates [19],

E.J. Beh • S.A. Cheema (✉) • D. Tran • I.L. Hudson
School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW, Australia
e-mail: eric.beh@newcastle.edu.au; salman.cheema@uon.edu.au; duytungtran@yahoo.com; irenelena.hudson@gmail.com

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_24, Published online: 12 November 2014

p. 447 agreed with Fisher to some extent with the exception of extreme marginal distributions. Plackett [16] and Berkson [4] determined the efficacy of marginal data by disagreeing with Fisher.

More recently, the statistical, and allied disciplines, have seen an explosion of new topics and techniques for the analysis of aggregate data all of which lies within the framework of ecological inference. Fréchet [9] provided the upper and lower bounds of cell values of contingency tables using marginal data. Recently, Dobra and Fienberg [5] extended those bounds for higher dimensional contingency tables. Goodman [11] is considered as the first serious attempt to model aggregate data. Kousser [14] and Freedman et al. [10] presented some adaptations of Goodman's approach. King's parametric and non-parametric models are considered as the breakthrough in modelling aggregate data [13]. More recently [17] homogeneous modelling approach is also important. However, all of the ecological inference techniques require assumptions about the individual level data that cannot be rigorously tested. Hudson et al. [12] demonstrated the effectiveness of a variety of ecological inference strategies by considering early New Zealand gendered election data and concluded that "these assumptions are either unrealistic or untestable". Wakefield [18], among many, is also another note able contributor in the area of ecological inference.

Recently, Beh [1, 2], proposed the aggregated association index (AAI), which, quantifies the strength of the association between the categorical variables when only the marginal information, or aggregate data, is available. Rather than estimating the cells (or some function of them) of multiple 2×2 contingency tables, the purpose of the AAI is to quantify the likelihood that a statistically significant association exists between the two dichotomous variables. Unlike the various ecological inference techniques that are available, the AAI may be applied to the analysis of a single 2×2 table.

One unfortunate characteristic of the AAI, whose basis is Pearson's chi-squared is that it increases as the sample size increases, given that Pearson's chi-squared statistic is susceptible to changes in the sample size of the contingency table. Therefore, the true nature of the association between the variables can be masked by the magnitude of the sample size. Recently, Beh et al. [3] presented two adjustments to reduce the effect of increased sample size on the magnitude of the AAI. This paper further explores and establishes a new adjustment to the AAI which is shown to be more efficient at reducing the impact of the sample size on the index, and may be considered when the relative marginal frequencies remain constant. A simple empirical study of the AAI and its adjustments will be provided using [8] twin criminal data which motivated his seminal discussion of the analysis of aggregate data.

2 The Aggregate Association Index

2.1 Notation

Suppose, n_o, is the original sample size of a 2×2 contingency table where n_{ij} denotes its (ij)th cell frequency. Therefore, let $p_{ij} = n_{ij}/n_o$ be the proportion of

Table 1 A general 2×2 contingency table

	Column 1	Column 2	Total
Row 1	n_{11}	n_{12}	$n_{1.}$
Row 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n_o

classifications made into this cell for $i = 1, 2$ and, $j = 1, 2$. The ith row jth column marginal frequencies are denoted by $n_{i.} = \sum_{j=1}^{2} n_{ij}$ and $n_{.j} = \sum_{i=1}^{2} n_{ij}$ respectively so that $\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} = \sum_{i=1}^{2} n_{i.} = \sum_{j=1}^{2} n_{.j} = n_o$. Thus, let $p_{i.} = n_{i.}/n_o$ and $p_{.j} = n_{.j}/n_o$ be the ith row marginal and jth column marginal proportions respectively. Let, also, $e_{ij} = n_{i.}n_{.j}/n_o$, represents the expected cell frequency for (ij)th cell, when there is no relationship between two variables. Table 1 gives the general form of a 2×2 contingency table.

When the cell frequencies of Table 1 are unknown, hence only the aggregate data is available, [6] considered the upper and lower bounds of n_{11}, as,

$$A_1 = \max(0, n_{.1} - n_{2.}) \leq n_{11} \leq \min(n_{.1}, n_{1.}) = B_1.$$

Rather than considering, n_{11}, much of the attention given to the ecological inference techniques focuses to date on the conditional proportion $P_1 = n_{11}/n_{1.}$, which is the conditional probability of the classification of an individual into "Column 1" given that it has been classified in "Row1". Therefore the bounds of P_1 are

$$L_1 = \max\left(0, \frac{p_{.1} - p_{2.}}{p_{1.}}\right) \leq P_1 \leq \min\left(\frac{p_{.1}}{p_{1.}}, 1\right) = U_1. \tag{1}$$

Beh [2] showed that when only marginal information is available, and a test of association is made at the a level of significance, the bounds of P_1 can be narrowed to

$$L_\alpha(n_o) = \max\left(0, p_{.1} - p_{2.}\sqrt{\frac{\chi^2_\alpha}{n_o}\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right)}\right)$$

$$< P_1 < \min\left(1, p_{.1} + p_{2.}\sqrt{\frac{\chi^2_\alpha}{n_o}\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right)}\right) = U_\alpha(n_o), \tag{2}$$

where χ^2_α is the $1 - \alpha$ percentile of the chi-squared distribution with 1 degree of freedom. Since we consider the case where each of the cell frequencies of Table 1 is unknown, the proportion of interest, P_1, is therefore also unknown. Despite this, [1, 2] demonstrated that Pearson's chi-squared statistic can be expressed as a quadratic

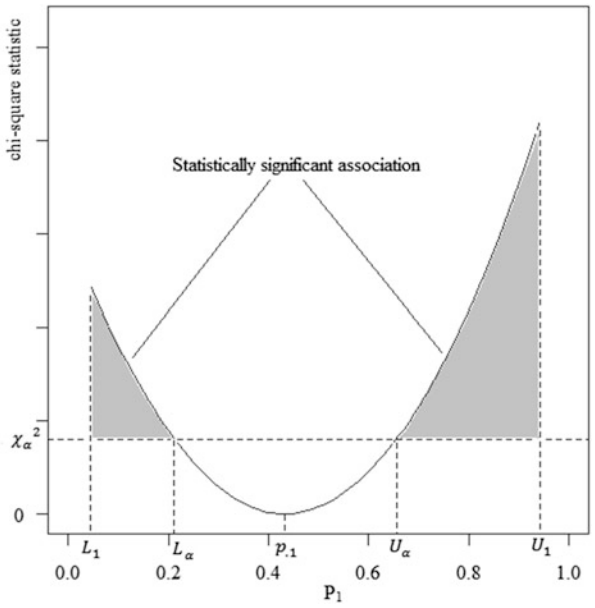

Fig. 1 A graphical display of the AAI

function of this proportion such that,

$$X^2\left(P_1|p_{1.}, p_{.1}\right) = n_o\left(\frac{P_1 - p_{.1}}{p_{2.}}\right)^2\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right) \quad . \tag{3}$$

2.2 The Index

Figure 1 provides a graphical representation of the quadratic relationship between Pearson's chi-squared statistic, (3), and the bounds of (1) and (2). Note that U_α and L_α in Fig. 1 refer to the extremes of (2). The null hypothesis of independence between the dichotomous variables is rejected when the observed Pearson chi-squared value (at some value of P_1) exceeds the critical value of χ^2_α. Therefore, the region under the curve, defined by (3), but lying above the critical line of χ^2_α, indicates where a statistically significant association exists between the two variables, considering the marginal proportion only. The relative size of this region, when compared with the total area under the curve, is quantified by

$$A_\alpha = 100\left[1 - \frac{\chi^2_\alpha\left\{\left(L_\alpha\left(n_o\right) - L_1\right) + \left(U_1 - U_\alpha\left(n_o\right)\right)\right\}}{k n_o\left(\left(U_1 - p_{.1}\right)^3 - \left(L_1 - p_{.1}\right)^3\right)}\right.$$

$$-\frac{\left\{\left(U_\alpha\left(n_o\right)-p_{.1}\right)^3-\left(L_\alpha\left(n_o\right)-p_{.1}\right)^3\right\}}{\left(\left(U_1-p_{.1}\right)^3-\left(L_1-p_{.1}\right)^3\right)}\Bigg], \tag{4}$$

where $k=\frac{1}{3p_{.2}^2}\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right)$ and $0 \leq A_\alpha < 100$; see [2]. Equation (4) is referred to as the aggregate association index, or more simply the AAI. It quantifies, for a given α, how likely a particular set of fixed marginal frequencies will enable the user to conclude that there exists a statistically significant association between the variables. If $A_\alpha \approx 100$ then, given only the aggregate data, it is highly likely that a significant association exists. However, if $A_\alpha \approx 0$ it is highly unlikely that such an association exists.

Equation (3) shows that the magnitude of Pearson's chi-squared statistic is highly dependent on the sample size, n_o. For example, if the original sample size of Table 1 is increased by a factor of $C > 1$ so that $n = Cn_o$, then Pearson's statistic increases by a factor of C. This has been long understood and prompted Pearson to consider his phi-squared statistic. Everitt [7], p. 56, and many others, also discussed this feature of the statistic. The impact of the sample size on AAI can be observed by Eq. (4). As the sample size increases, $U_\alpha\left(n_o\right)$ and $L_\alpha\left(n_o\right)$ approaches to $p_{.1}$. Therefore, AAI approaches 100 as the sample size increases. We now propose a simple strategy to ensure that the AAI is less affected by any increase in sample size, when the marginal proportions are constant.

3 Adjusted Aggregate Association Index

The AAI defined by (4), can be expressed alternatively as, follows, see [3]

$$\begin{aligned} A_\alpha = 100\Bigg[1-f\left(n_o\right)\left(\frac{U_1-L_1}{U_\alpha\left(n_o\right)-L_\alpha\left(n_o\right)}\right) \\ \times\left\{\frac{\chi_\alpha^2\left\{\left(L_\alpha\left(n_o\right)-L_1\right)+\left(U_1-U_\alpha\left(n_o\right)\right)\right\}}{kn_o\left(\left(U_1-p_{.1}\right)^3-\left(L_1-p_{.1}\right)^3\right)}\right. \\ \left.+\frac{\left\{\left(U_\alpha\left(n_o\right)-p_{.1}\right)^3-\left(L_\alpha\left(n_o\right)-p_{.1}\right)^3\right\}}{\left(\left(U_1-p_{.1}\right)^3-\left(L_1-p_{.1}\right)^3\right)}\right\}\Bigg], \end{aligned} \tag{5}$$

where

$$f\left(n_o\right)=\frac{U_\alpha\left(n_o\right)-L_\alpha\left(n_o\right)}{U_1-L_1} \quad . \tag{6}$$

Suppose the level of significance, α, at which a test of independence is made remains fixed, as does the relative marginal proportions for the row and column categories.

Multiplying the sample size by $C > 1$ does not change the relative marginal information, although it does impact on the sample size and on the row and column totals of the contingency table. Increasing the original sample size of Table 1, n_o, by multiplying it by $C > 1$ will result in a new sample size $n = Cn_o$ and an increased Pearson's chi-squared statistic. Thus, given a sample size that is a multiple of C, we have,

$$A_\alpha(C) = 100\left[1 - f(Cn_o)\left(\frac{U_1 - L_1}{U_\alpha(Cn_o) - L_\alpha(Cn_o)}\right)\right.$$
$$\times\left\{\frac{\chi^2_\alpha\{(L_\alpha(Cn_o) - L_1) + (U_1 - U_\alpha(Cn_o))\}}{kCn_o\left((U_1 - p_{.1})^3 - (L_1 - p_{.1})^3\right)}\right.$$
$$\left.\left.+\frac{\left\{(U_\alpha(Cn_o) - p_{.1})^3 - (L_\alpha(Cn_o) - p_{.1})^3\right\}}{\left((U_1 - p_{.1})^3 - (L_1 - p_{.1})^3\right)}\right\}\right] \tag{7}$$

and,

$$f(Cn_o) = \frac{U_\alpha(Cn_o) - L_\alpha(Cn_o)}{U_1 - L_1}. \tag{8}$$

As the sample size, n_o, increases by a factor of $C > 1$, this narrows the interval (2) and therefore decreases the magnitude of $f(Cn_o)$, hence AAI increases, even though the relative marginal information remains unchanged. Specifically, as $C \to \infty$, $f(Cn_o) \to 0^+$, and $A_\alpha(C) \to 100$. Similarly, as $C \to 0^+$, $f(Cn_o) \to 1^+$, and $A_\alpha(C) \to 0$. Therefore, to help minimise the impact increasing sample size has on AAI, different specifications of $f(n_o)$, subject to $0 \le f(n_o) \le 1$, may be considered as an alternative to (6). As a result, this adjusts AAI according to the choice of $f(Cn_o)$ and leads to our adjusted AAI

$$A^{/}_\alpha = 100\left[1 - f^{/}(n_o)\left(\frac{U_1 - L_1}{U_\alpha(n) - L_\alpha(n)}\right)\right.$$
$$\times\left\{\frac{\chi^2_\alpha\{(L_\alpha(n) - L_1) + (U_1 - U_\alpha(n))\}}{kn\left((U_1 - p_{.1})^3 - (L_1 - p_{.1})^3\right)}\right.$$
$$\left.\left.+\frac{\left\{(U_\alpha(n) - p_{.1})^3 - (L_\alpha(n) - p_{.1})^3\right\}}{\left((U_1 - p_{.1})^3 - (L_1 - p_{.1})^3\right)}\right\}\right], \tag{9}$$

where $f^{/}(n_o)$ is the adjustment of (6) and may be subjectively, or objectively, determined so that $0 \le f^{/}(n_o) \le 1$.

One possible adjustment to consider is a subjective choice of $f^{/}(n_o)$ that remains constant for all C. A conservative value, and one that [3] used, is $f^{/}(n_o) = 0.5$.

Another adjustment relies on a rearrangement of Eq. (2): as, $U_\alpha(n_o) - L_\alpha(n_o) = 2p_{2.}\sqrt{\frac{\chi^2_\alpha}{n_o}\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right)}$. Therefore, a second candidate for adjustment to (6) is to consider

$$f^{/}(n_o) = \frac{2p_{2.}}{U_1 - L_1}\sqrt{\frac{\chi^2_\alpha}{n_o}\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right)} \quad .$$

In this case, the adjusted AAI, (7), is

$$A^{/}_\alpha = 100\left[1 - \sqrt{\frac{n}{n_o}}\left\{\frac{\chi^2_\alpha\{(L_\alpha(n) - L_1) + (U_1 - U_\alpha(n))\}}{kn\left((U_1 - p_{.1})^3 - (L_1 - p_{.1})^3\right)}\right.\right.$$
$$\left.\left. + \frac{\left\{(U_\alpha(n) - p_{.1})^3 - (L_\alpha(n) - p_{.1})^3\right\}}{\left((U_1 - p_{.1})^3 - (L_1 - p_{.1})^3\right)}\right\}\right]. \tag{10}$$

Given the adjustment in (10), the relationship between the original AAI, (4), and its adjusted AAI , (10), may be expressed as,

$$A^{/}_\alpha = A_\alpha\sqrt{\frac{n}{n_o}} - 100\left(\sqrt{\frac{n}{n_o}} - 1\right).$$

Thus, if the original sample size is increased by a factor of C, where $C > 1$, so that $n = Cn_o$, then

$$A^{/}_\alpha = A_\alpha\sqrt{C} - 100\left(\sqrt{C} - 1\right) \quad . \tag{11}$$

Hence, $A^{/}_\alpha < A_\alpha$ for any reasonably large C.

The following adjustment, now incorporates the rate of change of (3), the transformation of Pearson's chi-square statistic, with respect to the sample size. The rate of change of (3) can be quantified as,

$$\frac{d}{dn_o}X^2(P_1|p_{1.}, p_{.1}) = \left(\frac{P_1 - p_{.1}}{p_{2.}}\right)^2\left(\frac{p_{1.}p_{2.}}{p_{.1}p_{.2}}\right).$$

In order to take account of the impact of the sample size, we consider the ratio of the average rate of change with respect to original and increased sample size. Let us

denote this ratio by

$$D = \frac{(U_\alpha(n_o) - p_{.1})^2 + (L_\alpha(n_o) - p_{.1})^2}{(U_\alpha(n) - p_{.1})^2 + (L_\alpha(n) - p_{.1})^2}, \tag{12}$$

where the numerator of (12) is the average rate of change of (3), with respect to the original sample size, over the corresponding interval (2) of P_1. Similarly, the denominator considers this average rate of change given the increased sample size.

By simplifying Eq. (12), we see that it is simply the ratio of increased sample size, n, to the original sample size, n_o, where, $D = C = n/n_o$. Thus, a generalised version of adjustment (11), is

$$A_\alpha^{//} = A_\alpha\sqrt{C} - 100\left(\sqrt{C} - 1\right) - \sqrt{D}, \tag{13}$$

where $D = 0$ if $n_o = n$ and $D = C$ if $n = Cn_o$ when $C > 1$. It is important to note that this adjustment acts as a penalty on the transformation of Pearson's chi-squared statistic to reduce the impact of increased sample size on the original AAI, when the marginal proportions are unchanged.

When there is no increase in the sample size from its original size, n, this adjustment ensures that the adjusted AAI is equivalent to the original AAI, A_α It also preserves the association structure by ensuring that any increase in sample size does not increase the Pearson chi-squared statistic or the AAI. Mathematically, this characteristic is established by evaluating the effect of increased sample size, as function of $\sqrt{C}$, on, A_α, $A_\alpha^{/}$ and $A_\alpha^{//}$. By considering the rate of change, with respect to $\sqrt{C}$, of the AAI, A_α, given by Eq. (5), and the adjusted AAI's of $A_\alpha^{/}$ and $A_\alpha^{//}$, given by Eqs. (11) and (13) respectively, then

$$\frac{\partial A_\alpha}{\partial\sqrt{C}} > \frac{\partial A_\alpha^{/}}{\partial\sqrt{C}} > \frac{\partial A_\alpha^{//}}{\partial\sqrt{C}}.$$

As such any increase in the AAI due to an increase original sample size, $n = Cn_o$, is reduced for $A_\alpha^{//}$ when compared with the original AAI and Eq. (11)

We may also note that $\frac{\partial A_\alpha^{/}}{\partial\sqrt{C}} = A_\alpha - 100 \leq 0$ since $0 \leq A_\alpha < 100$. Similarly, $\frac{\partial A_\alpha^{//}}{\partial\sqrt{C}} = A_\alpha - 100 - 1 < \frac{\partial A_\alpha^{/}}{\partial\sqrt{C}} < \frac{\partial A_\alpha}{\partial\sqrt{C}}$. We shall now empirically demonstrate this feature.

4 Empirical Study

Consider Table 2 as was originally studied by [8]. It cross-classifies 30 criminal twins according to whether they are a monozygotic twin or a dizygotic twin, and also informs on whether their same sex twin has been convicted of a criminal offence.

Table 2 Fisher [8] criminal twin data

	Convicted	Not convicted	Total
Monozygotic	10	3	13
Dizygotic	2	15	17
Total	12	18	30

Pearson's chi-squared statistic for Table 1 is 13.032, with a p-value of 0.0003, which shows that there is a statistically significant association between the two dichotomous variables. For this data $P_1 = 10/13 = .7692$, hence 77 % of those monozygotic criminal twins in the sample have a same sex sibling who has also been convicted of a crime.

Beh [2] analysed Fisher's twin data and established that, if only the aggregate data were known, and testing the association at the 5 % level of significance, the AAI of Eq. (4) is 61.83. Therefore, based only on analysis of the aggregate data of Table 2, there is a 61.83 % chance of a statistically significant association between the variables. This index can also be viewed using the adjusted AAI of Eq. (10)—or, equivalently (11) since $C = 1$ (see [3]).

Suppose we now consider the case where larger samples than $n_o = 30$ were selected, but where the marginal proportions of Table 2 remain the same. Figure 2 graphically shows the impact of the AAI as increases for C ranging from 1 to 20, equivalent to a sample size ranging from 30 to 600. When, $n = 30$, $A_{.05} = A^{/}_{.05} = 61.83$, and since $C = 1$, then $D = 0$, so that $A^{//}_{.05} = 61.83$ This shows that the adjusted versions of the AAI are identical to the original AAIAs the sample size increases, such as for $n = 600$, the magnitude of the original AAI increases to $A_{.05} = 97.49$ indicating that it is now extremely likely that an association exists between the variables of Table 2 (given only the aggregate data).

Our aim is therefore to stabilise AAI as the sample size increases. This will allow us to obtain a clearer indication of the true nature of the association by reducing the impact of the magnitude of n, and can be achieved by the two proposed adjusted AAI's. As the sample size increases, these adjusted versions of A_α do increase, but more slowly than original A_α. For example, at $n = 600\,(C = 20)$, the first adjustment given by Eq. (11), yields $A^{/}_{.05} = 88.77$ Note that, this adjusted AAI can be obtained from the original AAI by considering Eq. (11):

$$A^{/}_{.05} = 97.49\sqrt{20} - 100\left(\sqrt{20} - 1\right) = 88.77$$

The newly proposed adjustment, $A^{//}_{.05}$ leads to 84.29 extent of association given only the marginal data. This can directly be calculated via Eq. (13), which gives,

$$A^{//}_{.05} = 97.49\sqrt{20} - 100\left(\sqrt{20} - 1\right) - \sqrt{20} = 84.29$$

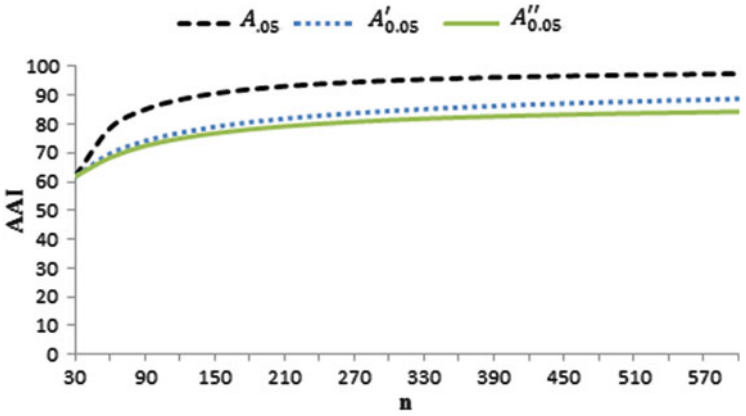

Fig. 2 Comparison of, $A^{/}_{.05}$, using the first adjustment (*blue line*), and the second adjustment, $A^{//}_{.05}$, (*green line*), with the original AAI, $A_{.05}$ (*dashed line*) as n increases

Figure 2 shows that the rate of change of both, $A^{/}_{.05}$ and $A^{//}_{.05}$is more stable, and less than, the original AAI as C increases from 1 to 20.

5 Discussion

In this article we have presented two adjustments of the original AAI of a 2×2 contingency table that help to stabilise the association index for increases in the sample size. We have demonstrated this empirically using [8] classic twin example and shown that both adjustments do not inflate the magnitude of the index as severely as the original index given increased sample size, One may view these adjustments as simple ad hoc strategies for minimising the impact of sample size when assessing the statistical significance of the association between two dichotomous variables. However, this study provides only an introduction into how adjustments to the AAI can be made. More comprehensive research still needs to be undertaken to reveal the features of these, and other, adjustments. For example, one area of that requires further investigation is establishing an $f^{/}(n_o)$ that formally minimises the rate of change of $A^{/}_{\alpha}$, and hence provides a more stable index, as the sample size increases.

Bibliography

1. Beh, E.J.: Correspondence analysis of aggregate data: the 2×2 table. J. Stat. Plan. Inference **138**, 2941–2952 (2008)
2. Beh, E.J.: The aggregate association index. Comput. Stat. Data Anal. **54**, 1570–1580 (2010)
3. Beh, E.J., Cheema, S.A., Tran, D., Hudson, I.L.: Adjusting the aggregate association index for large samples. In: Proceedings of advances on latent variables/methods models and applications, Brescia, Italy (2013)

4. Berkson, J.: In dispraise of the exact test: do the marginal totals of the 2 × 2 table contain relevant information respecting the table proportion. J. Stat. Plan. Inference **2**, 27–42 (1978)
5. Dobra, A., Fienberg, S.E.: Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. Proc. Natl. Acad. Sci. **97**, 11885–11892 (2002)
6. Duncan, O.D., Davis, B.: An alternative to ecological correlation. Am. Soc. Rev. **18**, 665–666 (1953)
7. Everitt, B.S.: The Analysis of Contingency Tables. Wiley, New York (1977)
8. Fisher, R.A.: The logic of inductive inference (with discussion). J. R. Stat. Assoc. Ser. A **98**, 39–82 (1935)
9. Fréchet, M.: Les probabilitisé, Associées a un Systéme d'Événments Compatibles et Dé pendants, Premiere Partie. Hermann and Cie, Paris (1940)
10. Freedman, D.A., Klein, S.P., Sacks, J, Smyth, C.A., Everett, C.G.: Ecological regression and voting rights. Eval. Rev. **15**, 673–711 (1991)
11. Goodman, L.: Ecological regressions and the behavior of individuals. Am. Soc. Rev. **18**, 663–666 (1953)
12. Hudson, I.L., Moore, L., Beh, E.J., Steel, D.G.: Ecological inference techniques: an empirical evaluation using data describing gender and voter turnout at New Zealand elections 1893–1919. J. R. Stat. Soc. **173**, 185–213 (2010)
13. King, G.: A Solution to Ecological Inference Problem. Princeton University Press, Princeton (1997)
14. Kousser, M. J.: Ecological regression and analysis of past politics. J. Interdiscip. Hist. **4**, 237–262 (1973)
15. Ogburn, W.F., Goltra, I.: How women vote: a study of an election in Portland, Oregon. Political Sci. Q. **34**, 413–433 (1919)
16. Plackett, R.L.: The marginal totals of a 2 × 2 table. Biometrika **64**, 37–42 (1977)
17. Steel, D.G., Beh, E.J., Chambers, R.L.: The information in aggregate data. In: King, G., Rosen, O., Tanner, M. (eds.) Ecological Inference: New Methodological Strategies, pp. 51–68. Cambridge University Press, Cambridge (2004)
18. Wakefield, J.: Ecological inference for 2 × 2 tables. J. R. Stat. Soc. Ser. A **167**, 385–445 (2004)
19. Yates, F.: Tests of significance for 2 × 2 contingency tables (with discussion). J. R. Stat. Soc. Ser. A **147**, 426–463 (1984)

Graphical Latent Structure Testing

Robin J. Evans

Abstract

Many models with latent structure are just semi-algebraic sets, and have recently begun to be studied from this perspective; this has shed much light on the dimension, identifiability, and asymptotic statistical properties of these models. Though most of the attention has been on equality constraints, some progress has also been made on evaluating inequalities which might be used to test such models. However, the mathematical complexity of these approaches seems to have led to a gap between our theoretical understanding and the manner in which these models are applied in practice. In this paper we make a plea for some focus on finding simpler (in particular more graphical) and more computationally feasible ways to express such constraints, even at the cost of a loss of statistical power. Recent advances for directed acyclic graph models with latent variables and phylogenetic models are given as illustrations.

Keywords

Graphical models • Inequalities • Latent variables • Phylogenetic trees

1 Introduction

Models based on unobserved or latent variables are widely used in psychology, epidemiology, genetics, economics and other disciplines. In cases where the latent variables themselves are of direct interest, it is generally necessary to fit such models by explicitly including the latent structure and using, for example, the EM-algorithm or an MCMC method with imputation. Problems with such methods are well documented, and include non-identifiability of parameters in the latent structure,

R.J. Evans (✉)
Department of Statistics, University of Oxford, Oxford, UK
e-mail: robin.evans@stats.ox.ac.uk

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_10, Published online: 28 October 2014

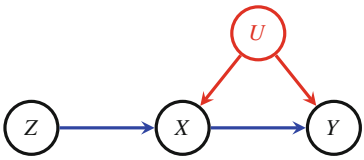

Fig. 1 The instrumental variables (IV) model; U is unobserved

multimodal likelihoods, and non-standard asymptotic properties of seemingly sensible estimators [see, e.g. 10, 12].

However, there are many examples in which the latent structure acts only as a nuisance, inhibiting standard inferential methods through confounding or selection bias. In such cases we advocate a constraint-based approach, meaning that testing and inference should focus on the implications of the model on the observable margin of the data. Even in cases in which it is necessary to model latent variables explicitly, constraint-based tests provide a useful method for model checking. This paper considers the observable implications of directed acyclic graph models with various kinds of latent variables, and discusses some of the recent advances in our understanding of these models.

As an example which is increasingly widely used in epidemiology and genetics, consider the discrete instrumental variables (IV) model, pictured in Fig. 1 [8]. Given random variables X, Y, Z, U under some joint distribution, the graph encodes the assumption that Z and U are marginally independent, and that Y is independent of Z conditional on X and U (for more details on directed graph models, see Sect. 2). Typically, Z is randomized (or assumed to be randomized), and U represents all possible sources of confounding between X and Y, whether understood or not. Interest usually lies in the strength of the effect of X upon Y, which cannot be estimated directly because of the confounding.

It usually makes little sense to try to model U explicitly, since we have no sense of what state-space might be suitable; indeed when the observed variables are binary, even if we assume the same is true of U, then the full model is unidentifiable. Instead we can ask whether there are any constraints over the observable margin (Z, X, Y) which might allow us to test the validity of our modelling assumptions. In fact it is well known that, in the case of discrete Z, X, Y, and making no assumption about the state-space of U, the observed probability distribution obeys the inequality constraints

$$\max_x \sum_y \max_z P(X = x, Y = y \mid Z = z) \leq 1. \tag{1}$$

This is the *instrumental inequality*, first derived by Pearl [17], and shown to be complete in the binary case by Bonet [5]; it thus provides the only test of the binary IV model, without making further assumptions.

In practice, however, this simple test is not widely applied in the applied literature [13]. This paper seeks to present the instrumental inequality as a special case of the much more general phenomenon of testable constraints which arise from latent structure, and which do not involve explicit modelling of latent variables. We believe

that more attention should be placed on using such observable constraints to validate models, especially before attempting to fit or interpret latent variables. Similar sentiments about the advantages of avoiding explicit modelling have been expressed by, for example, Allman and Rhodes [1] and Silva and Ghahramani [20]. In order to facilitate this approach it is essential that methods for finding constraints are easy to understand and computationally feasible; in particular, we advocate graphical methods for finding constraints.

To illustrate these ideas we present two examples of model classes in which there has been much recent progress towards deriving constraints graphically: marginalized directed acyclic graphs (with no assumption made about the latent variables), and phylogenetic tree models. The remainder of the paper is organized as follows: Sect. 2 looks at constraints on margins of directed acyclic graphs, and Sect. 3 at phylogenetic trees; Sect. 4 considers some other examples, and Sect. 5 contains a discussion.

2 Directed Acyclic Graphs with Latent Variables

A *directed acyclic graph* $\mathcal{G}$ is a pair $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a collection of *vertices*, and $\mathcal{E}$ is a collection of ordered pairs of distinct vertices, or *edges*. If $(X, Y) \in \mathcal{E}$ we write $X \to Y$, and say that X is a *parent* of Y. The set of parents of Y is denoted $\mathrm{pa}_{\mathcal{G}}(Y)$. A *path* is a sequence of adjacent edges in a graph, without repetition of vertices; for example, the graph in Fig. 1 contains the path $\pi_1 : Z \to X \leftarrow U \to Y$. A path is *directed* from X to Y if all the arrows point away from X and towards Y. A directed graph is *acyclic* if there is no directed path from a vertex to any of its parents.

We can associate each vertex X with a random variable under some multivariate distribution P; let P admit a density f. For convenience, we abuse notation and use X to refer to both the vertex and random variable. Similarly for sets of vertices such as $\boldsymbol{A}$ or $\mathrm{pa}_{\mathcal{G}}(X)$, we will use the same notation for the collection of random variables associated with those vertices. The *factorization criterion* for DAGs says that P is in the model corresponding to the DAG $\mathcal{G}$ if the joint density factorizes into univariate conditional distributions as $\prod_{X \in \mathcal{V}} f(X \mid \mathrm{pa}_{\mathcal{G}}(X))$.

For a path, π, internal vertices on π with two adjacent arrowheads are called *colliders* (on π); other internal vertices are *non-colliders*. On the path π_1 defined above, X is a collider, and U a non-collider. A path from V to W is *blocked* by a set of vertices $\boldsymbol{C}$ if either:

(i) there is a non-collider on π in $\boldsymbol{C}$; or
(ii) there is a collider on π which is neither in $\boldsymbol{C}$, nor is there any directed path from the collider to $\boldsymbol{C}$.

Definition 1 For disjoint sets of vertices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, we say that $\boldsymbol{A}$ and $\boldsymbol{B}$ are *d-separated* by $\boldsymbol{C}$, if every path from any vertex in $\boldsymbol{A}$ to any vertex in $\boldsymbol{B}$ is blocked by $\boldsymbol{C}$. A probability distribution P obeys the *global Markov property* for a DAG $\mathcal{G}$ if whenever $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{C}$ in $\mathcal{G}$, then $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C}\ [P]$.

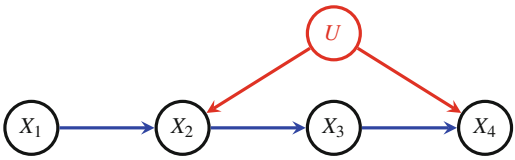

Fig. 2 A graph with a nested constraint on the observed distribution; U is unobserved

It is well known that (when a joint density exists) d-separation is equivalent to the factorization criterion [23]. In particular, all constraints implied by a DAG model on fully observed random variables may be interpreted as conditional independences.

2.1 Introducing Latent Vertices

If some of the variables in a DAG are unobserved, we may be interested in the implications of the underlying graph for the observable margin. Let $\mathscr{U} \subset \mathscr{V}$ denote the set of latent or unobservable vertices; the observable margin is then the distribution over $\mathscr{V} \setminus \mathscr{U}$. In this section we will make no assumption about the state-space of the latent variables. Some conditional independences will still be observable: specifically, we can determine whether or not $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C}$ if all the variables in $\boldsymbol{A} \cup \boldsymbol{B} \cup \boldsymbol{C}$ are observed. In the graph in Fig. 2, X_1 and X_3 are d-separated by X_2, and since all these variables are observed, the independence $X_1 \perp\!\!\!\perp X_3 \mid X_2$ holds in the observed marginal distribution.

The observed vertices of a graph may be partitioned into *districts*; X and Y lie in the same district if there is a path between X and Y on which no two adjacent vertices are both observable. The graph in Fig. 2 has three districts, $\{X_2, X_4\}$, $\{X_1\}$ and $\{X_3\}$. Latent variables and their incident edges will be drawn in red (see Fig. 1); districts are then joined by red paths. Due to the arbitrary state-space of the latent variables, without loss of generality we may consider only graphs in which none of the latent variables have any parents.

2.2 Nested Constraints

Constraints other than conditional independences also arise, specifically *nested constraints* and inequalities on the observed distribution. Nested constraints can be understood graphically by iteratively removing vertices with no children (marginalizing) and splitting graphs into districts and their parents [see 19, 21]. Both these operations lead to corresponding distributions which are identifiable from the observed distribution, and therefore any independences which hold in the graphs so obtained may be tested with data.

Here we give an example derivation of a nested constraint: full details are found in Shpitser et al. [19]. Consider the district $\{X_2, X_4\}$ from the graph in Fig. 2, together with its parents $\{X_1, X_3\}$; edges not within the district, nor directed from a parent to the district are removed. In this case the edge $X_2 \to X_3$ is dropped,

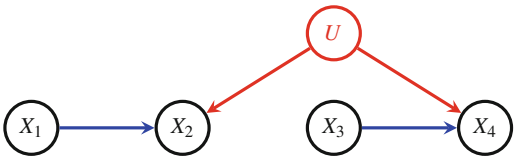

Fig. 3 The graph from Fig. 2 after isolating the district $\{X_2, X_4\}$

leaving the graph in Fig. 3. This operation corresponds to dividing the joint density by $f(x_1) \cdot f(x_3 \mid x_2)$, to yield

$$f^*(x_2, x_4 \mid x_1, x_3) \equiv \frac{f(x_1, x_2, x_3, x_4)}{f(x_1) \cdot f(x_3 \mid x_2)}, \tag{2}$$

which is a new conditional probability density on $X_2, X_4 \mid X_1, X_3$; let the associated distribution be P^* (marginal distributions for X_1 and X_3 may be chosen arbitrarily to form a full joint distribution).

In the graph in Fig. 3, X_1 is d-separated from X_4 by X_3, so it must hold that $X_4 \perp\!\!\!\perp X_1 \mid X_3$ under P^* (indeed, this can be seen directly using the factorization criterion). All quantities on the right hand-side of (2) can be estimated from data, so one can test the constraint using, for example, a likelihood ratio test between the model where the constraint is enforced and the model where it is not.

These nested constraints strictly generalize conditional independence, and we conjecture that the constraints enumerated by the Markov properties of Shpitser et al. [19] are complete, in the sense that there are no further algebraic constraints on the joint distribution induced by these latent variable models.

2.3 Separation Criterion for Inequalities

Marginalized DAG models also induce inequality constraints. In general and at present, deriving such bounds exhaustively is difficult and prohibitively computationally expensive other than for very small graphs; see, for example, Bonet [5]. Since finding these bounds is an example of the NP-complete problem of determining membership of projections of a convex polytope, there is reason to believe that fast methods may not be obtainable in general [22].

Kang and Tian [15] give an algorithm for obtaining inequalities on causal effects which can be used to derive bounds on the observed marginal distribution. Evans [11] gives a graphical method based on a separation criterion (presented below), analogous to d-separation. Neither of these methods will find *all* inequalities, but both represent a step towards reducing the difficulties in finding such constraints.

Definition 2 Let $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} \subseteq \mathscr{V}$ be disjoint sets of observed vertices in a DAG $\mathscr{G}$. We say that $\boldsymbol{A}$ is *e-separated* from $\boldsymbol{B}$ by $\boldsymbol{C}$, after deletion of $\boldsymbol{D}$, denoted

$$\boldsymbol{A} \perp_e \boldsymbol{B} \mid \boldsymbol{C} \not\downarrow \boldsymbol{D},$$

if $\boldsymbol{A}$ is d-separated from $\boldsymbol{B}$ by $\boldsymbol{C}$ in the induced subgraph obtained by deleting the vertices in $\boldsymbol{D}$.

Theorem 1 ([11], Theorem 4.2) *Let $\mathcal{G}$ be a DAG and $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ be sets of observable vertices such that $\boldsymbol{A} \perp_e \boldsymbol{B} \mid \boldsymbol{C} \not| \boldsymbol{D}$ in $\mathcal{G}$. Let P be a discrete distribution which obeys the global Markov property for $\mathcal{G}$; then for any fixed value of $\boldsymbol{D} = \boldsymbol{\delta}$, there must exist a distribution P^* such that $P^*(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\delta} \mid \boldsymbol{C}) = P(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\delta} \mid \boldsymbol{C})$ for all $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, and under which $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C}$ $[P^*]$.*

The proof of this result is partially constructive, and some examples are given in Evans [11]. Applied to the IV model in Fig. 1, we can see that $Z \perp_e Y \not| X$, i.e. that Z and Y are separated (unconditionally) after deleting X. Applying a slight extension of the Theorem (not given for brevity) yields precisely Pearl's instrumental inequality (1). The theorem also has two appealing and easy corollaries.

Corollary 1 *Any e-separation implies a testable constraint on a joint distribution over discrete random variables.*

Corollary 2 *If any two vertices are not joined by an edge, nor share a latent parent, then a testable constraint exists for discrete random variables.*

The corollaries demonstrate the simplicity of finding inequality constraints with the e-separation criterion, especially in comparison to more direct computational approaches. We remark that testing inequality constraints in finite samples is still no trivial matter; see Ramsahai and Lauritzen [18] for an approach in the instrumental variables case.

2.4 Other Inequalities

The three methods given above for finding constraints (d-separation, nested constraints and e-separation) are all fully *graphical*; consequently they are, in our view, relatively easy to understand and work with. In contrast, other existing methods for deriving such inequalities involve complex algorithms [15] or are computationally infeasible for large graphs [5].

The full collection of inequalities associated with a marginalized DAG is, in general, extremely complicated. Even for instrumental variables, the simplest model containing a non-trivial inequality, if the instrument Z takes three states then the instrumental inequality (and equivalently Theorem 1) no longer suffices to describe the observed margin; instead we obtain additional inequalities such as

$$p_{01|1} + p_{00|2} + p_{01|0} + p_{11|1} + p_{10|0} \leq 2,$$

where $p_{ij|k} = P(X = i, Y = j \mid Z = k)$ [5]. Although the graphical approach we have outlined does not yield a complete set of constraints, we contend that

the simplicity and greater computational feasibility of this approach makes it more suitable for practical use.

3 Phylogenetic Trees

A phylogenetic tree model is an idealized mathematical representation of an evolutionary tree. It takes the form of a directed acyclic graph, $\mathcal{G}$, in which each vertex has precisely one parent, except for a single node known as the *root*; such a graph is called a *rooted tree*. The vertices in $\mathcal{G}$ without children are known as *leaves*, and are observed; the internal vertices are all unobserved. An example is given in Fig. 4.

All the variables, both latent and observed, are assumed to have the same state-space; we take the binary case for illustration. In order to distinguish between different graph topologies, there has been much focus in the literature on finding *phylogenetic invariants*; that is, polynomials which vanish under particular topologies [see, for example, 1,3,6].

The consequent results are highly graphical. Any internal edge h in the tree splits the graph into exactly two pieces; let the observed vertices in each piece be denoted $\boldsymbol{A}, \boldsymbol{B}$. Then define M_h to be the matrix whose (i, j)th entry corresponds to $P(\boldsymbol{A} = i, \boldsymbol{B} = j)$. It is not hard to show that because the vertices adjacent to this edge have only two states, the matrix M_h has rank at most two, and hence all its 3×3 minors must vanish. Note that this result acts as a form of weak independence, since if $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B}$ the matrix would have rank 1.

Consider the tree in Fig. 4, and let $p_{ijkl} = P(X_1 = i, X_2 = j, X_3 = k, X_4 = l)$ for $i, j, k, l \in \{0, 1\}$; define

$$M_{\{H_1,H_2\}} = \begin{pmatrix} p_{0000} & p_{1000} & p_{0100} & p_{1100} \\ p_{0010} & p_{1010} & p_{0110} & p_{1110} \\ p_{0001} & p_{1001} & p_{0101} & p_{1101} \\ p_{0011} & p_{1011} & p_{0111} & p_{1111} \end{pmatrix};$$

this is the matrix corresponding to the only internal edge: $H_1 \to H_2$. If the model holds, then this matrix has rank at most 2.

In fact Allman and Rhodes [1], Theorem 4, show that a binary phylogenetic model is satisfied, up to inequalities, if and only if the above condition holds. In

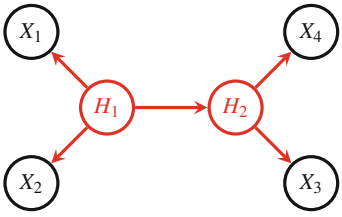

Fig. 4 A phylogenetic tree model with four leaves

particular, this gives us a way to distinguish between any two tree topologies. The beauty and utility of this result stems from the fact that we can test the plausibility of a phylogenetic model using our data before ever having to actually fit any latent variables; we need not worry about identifiability, local maxima, or other difficulties. In addition, although the mathematics behind the results in this area are complex, the condition is simple and easy to understand (as, hopefully, we have demonstrated above).

3.1 Phylogenetic Inequalities

The model described by the constraints above is typically much larger than the image of the latent variable model, because it fails to account for inequality constraints. There has been much progress in determining such constraints, although a full description is not available in general [3]. The main result along these lines is also a graphical one: the removal of any internal vertex splits the leaves into three groups, and the three-way tensor defined analogously to M_e must satisfy certain positivity conditions.

4 Other Models

There are other forms of graphical models with latent structure which induce testable constraints. One example is the *latent class model*, which consists of a single hidden variable of r states, and m observable children taking respectively $k_1, \ldots, k_m$ states. An example with $m = 5$ is given in Fig. 5.

Similar results are available to those mentioned for phylogenetic models [12], which are closely related objects. Though these equations are not exhaustive, they are very easy to compute and fairly intuitive. It is therefore disappointing that such constraints do not appear to be widely mentioned in the applied literature [see, for example, 7, 16]. The focus instead seems to be on fitting models with varying numbers of latent classes and comparing likelihood ratio statistics, even though model singularities mean that in many examples the asymptotic null distribution of such statistics is unknown.

One property which has been quite widely studied in the literature is the MTP_2 *constraint*, which applies when binary observed variables are assumed to be monotonically related to a univariate latent variable [14]. A likelihood ratio test

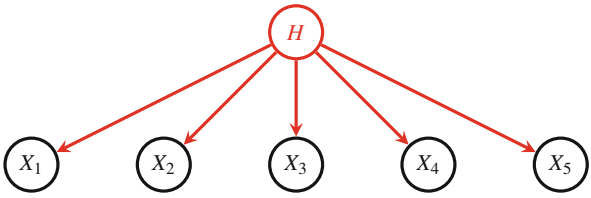

Fig. 5 A latent class model on five responses

was developed by Bartolucci and Forcina [4], but does not seem to be widely used. Recently Allman et al. [2] have derived a complete semi-algebraic characterization of the latent class model in the case of binary observed and latent variables; the inequalities are closely related to MTP_2 constraints.

Analogous results are available for factor analysis models, which concern Gaussian latent and observed variables [9]; again, their adoption in the applied literature seems limited.

5 Discussion

We end with a plea for the further development and dissemination of graphical (or otherwise relatively simple) methods for determining hidden structure. Latent variable models are widely used by researchers in myriad disciplines, both because they may fit with current scientific theory, and because they often yield highly interpretable results. It seems clear, however, that the mathematical theory underlying these models, which has advanced rapidly in recent years, is not matched by applied statistical practice.

Until and unless our structural tests and associated fitting methods are shown to match the simplicity and intuitive appeal of the models themselves, the practice will continue to lag behind the theory. It seems incumbent upon statisticians to bridge this gap by presenting these methods as naturally as is possible, and developing software which derives constraints and performs appropriate tests efficiently, even at the expense of statistical power.

References

1. Allman, E.S., Rhodes, J.A.: Phylogenetic ideals and varieties for the general Markov model. Adv. Appl. Math. **40**(2), 127–148 (2008)
2. Allman, E.S., Rhodes, J.A., Sturmfels, B., Zwiernik, P.: Tensors of nonnegative rank two (2013). arXiv:1305.0539. arXiv preprint
3. Allman, E.S., Rhodes, J.A., Taylor, A.: A semialgebraic description of the general Markov model on phylogenetic trees (2012). arXiv:1212.1200. arXiv preprint
4. Bartolucci, F., Forcina, A.: A likelihood ratio test for MTP_2 within binary variables. Ann. Stat. **28**(4), 1206–1218 (2000)
5. Bonet, B.: Instrumentality tests revisited. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, pp. 48–55 (2001)
6. Cavender, J.A., Felsenstein, J.: Invariants of phylogenies in a simple case with discrete states. J. Classif. **4**(1), 57–71 (1987)
7. Crow, S.J., Swanson, S.A., Peterson, C.B., Crosby, R.D., Wonderlich, S.A., Mitchell, J.E.: Latent class analysis of eating disorders: relationship to mortality. J. Abnorm. Psychol. **121**(1), 225–231 (2012)
8. Didelez, V., Sheehan, N.: Mendelian randomization as an instrumental variable approach to causal inference. Stat. Methods Med. Res. **16**(4), 309–330 (2007)
9. Drton, M., Sturmfels, B., Sullivant, S.: Algebraic factor analysis: tetrads, pentads and beyond. Probab. Theory Relat. Fields **138**(3), 463–493 (2007)
10. Drton, M.: Likelihood ratio tests and singularities. Ann. Stat. **37**(2), 979–1012 (2009)

11. Evans, R.J.: Graphical methods for inequality constraints in marginalized DAGs. In: IEEE International Workshop on Machine Learning for Signal Processing (2012)
12. Fienberg, S.E., Hersh, P., Rinaldo, A., Zhou, Y.: Maximum Likelihood Estimation in Latent Class Models for Contingency Table Data, Chap. 2, pp. 27–62. Cambridge University Press, Cambridge (2009)
13. Glymour, M.M., Tchetgen, E.J.T., Robins, J.M.: Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. Am. J. Epidemiol. **175**(4), 332–339 (2012)
14. Holland, P.W., Rosenbaum, P.R.: Conditional association and unidimensionality in monotone latent variable models. Ann. Stat. **14**(4), 1523–1543 (1986)
15. Kang, C., Tian, J.: Inequality constraints in causal models with hidden variables. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 233–240, Cambridge, AUAI Press (2006)
16. Mezuk, B., Kendler, K.S.: Examining variation in depressive symptoms over the life course: a latent class analysis. Psychol. Med. **42**, 2037–2046 (2012)
17. Pearl, J.: On the testability of causal models with latent and instrumental variables. In: UAI-95, pp. 435–443 (1995)
18. Ramsahai, R.R., Lauritzen, S.L.: Likelihood analysis of the binary instrumental variable model. Biometrika **98**(4), 987–994 (2011)
19. Shpitser, I., Evans, R.J., Richardson, T.S., Robins, J.M.: Introduction to nested Markov models. Behaviormetrika, **41**(1) 3–39 (2014)
20. Silva, R., Ghahramani, Z.: The hidden life of latent variables: Bayesian learning with mixed graph models. J. Mach. Learn. Res. **10**, 1187–1238 (2009)
21. Tian, J.: Studies in causal reasoning and learning. Ph.D. thesis, UCLA (2002)
22. Ver Steeg, G., Galstyan, A.: A sequence of relaxations constraining hidden variable models. In: Proceedings of the Twenty-seventh Conference on Uncertainty in Artificial Intelligence, pp. 717–727 (2011)
23. Verma, T., Pearl, J.: Causal networks: semantics and expressiveness. In: Schachter R., Levitt T.S., Kanal L.N., (eds.) Uncertainty in Artificial Intelligence 4. New York: Elsevier, pp. 69–76 (1990)

Understanding Equity in Work Through Job Quality: A Comparative Analysis Between Disabled and Non-Disabled Graduates Using a New Composite Indicator

Giovanna Boccuzzo and Licia Maron

Abstract

This paper compares the job quality of disabled and non-disabled graduates. Equity in work is measured based on not only whether a graduate has a job but also whether he/she has a good job. The Italian law favours disabled people by often providing them a job; however, the quality of the job is not considered. In this study, job quality is measured using a composite indicator (CI) comprising three dimensions: economic, professional and work-life balance. The proposal of the CI structure is original because the variables that compose the indicator can be quantitative, ordinal or dichotomous. The results of our study show that there is no difference in the job quality of disabled and non-disabled graduates; however, there are differences within the two groups in terms of the greatest dimension of the CI: economic dimension for the disabled group and professional dimension for the non-disabled group. Disabled people have the guarantee of a stable contract; however, their jobs are not consistent with their educational qualifications.

Keywords

Job quality • Composite indicator • Graduates • Mixed data • Generalised distance

Abbreviations

CI	Composite Indicator
CATI	Computer Assisted Telephone Interview

G. Boccuzzo (✉) • L. Maron
Department of Statistical Science, University of Padua, Padua, Italy
e-mail: giovanna.boccuzzo@unipd.it; licia.maron@gmail.com

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_5, Published online: 12 November 2014

1 Job Quality and Equity in Work

A respectable job that meets workers' competences and expectations is fundamental for their dignity and the realization of their ambitions. Personal realization in a working environment involves two steps: (1) to have a job and (2) to have a good job.

This paper aims to analyse the job quality of students 3 years after graduating from Padua University by focusing on the differences between the job quality of the disabled and non-disabled graduates.

Our study concentrates on job quality rather than on having a job owing to the following reasons: Firstly, this study refers to graduates 3 years after graduating and only 4 % of them are still looking for a job. Secondly, Italian regulations focus on improving the employment situation of disabled people. In fact, Law 68/1999 provides for the labour rights of the disabled, both as employees in social cooperatives and companies and as self-employed people. Thirdly, public competitions reserve a part of the available positions for disabled people or assign priority to the disabled (*ceteris paribus*). Unfortunately, although the law can promote the employment of disabled people, it cannot ensure the quality of their jobs. Finally, this research focuses on graduates who obtained a university degree, consequently have higher expectations from their jobs.

The way job quality is measured varies considerably across studies. In terms of *what* characteristics are considered when measuring job quality, some studies summarize job quality with a single variable, either objective, such as salary [14, 16], or subjective, such as job satisfaction [20, 22], while others suggest to consider several constitutive dimensions (*multi-faceted approach*) [7]. Concerning *how* the characteristics of job quality are measured, objective job attributes [18], subjective job perceptions [10] or a mix of the two have been adopted [19].

In this study, job quality is considered a multifaceted concept, based on a limited number of dimensions that can be described by objective and subjective indicators.

2 Data

This study was conducted on a sample of Italian students who graduated in 2007 and 2008. The data were obtained from the Agorà longitudinal survey on the career outcomes of graduates from University of Padua [8]. Respondents were interviewed after 6, 12 and 36 months from graduation using a Computer Assisted Telephone Interview (CATI) tool. Workers were required to provide considerable information regarding their current job, activities conducted by them when they were searching for a job, their perception of skill and educational mismatch and evaluation of their educational program.

A total of 2,885 people were interviewed 36 months after their graduation. Among them, 2,436 people were employed and therefore only these people were considered in this research.

Furthermore, we conducted a survey to collect the same data that were collected in the Agorà survey from all disabled graduates in 2004–2008. There were 307 disabled graduates during that period; however, many of them enrolled in a master degree or moved to another location. Of the 108 disabled people contacted, 74 were employed [2]. The disabled people were interviewed from 6 months to 5 years after their graduation; however, most of them were interviewed 2 or 3 years after their graduation.

The following analyses are relevant only when disabled graduates are compared with non-disabled ones. These two groups show different structures, especially with regard to their age and academic discipline; therefore, comparisons between the two groups are not appropriate from a methodological perspective. To make the two groups comparable, each disabled person was matched with one or two[1] non-disabled person of the same gender, age (5 year classes), disciplinary area (humanities, socio-economic, technical-scientific and life sciences) and type of degree (bachelor or other). The non-disabled sample comprised 130 graduates selected from the sample of Agorà-survey graduates, who were interviewed 36 months after graduating.

3 The Job Quality Indicator

Job quality is a multidimensional concept, which has been tentatively summarized using a composite indicator (CI).

The structure of the CI is hierarchical: the multidimensional concept of job quality comprises various dimensions. Each of these dimensions comprises several elementary indicators, represented by variables that can be directly measured (Fig. 1).

In our proposal, the aggregation method admits compensability: the CI is a weighted mean of dimensions and each dimension is a weighted mean of elementary indicators. This implies that poor performance in some indicators can

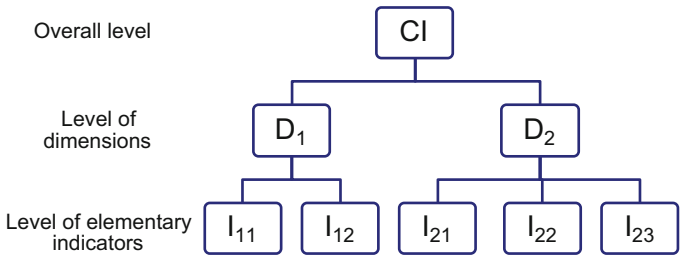

Fig. 1 Hierarchical structure of the composite indicator

[1] Occasionally, only one control can be found.

be compensated for by sufficiently high values in other indicators. It is necessary to discuss whether compensability among indicators should be permitted in the specific concept for which the CI is designed [21]. To this end, we consider the mainstream economic approach to job quality, which defends the existence of compensating differentials in the labour market: workers with the same skills will be offered different bundles of wage and disamenities, leading to the same job quality. Then, workers will choose whatever combination best suits their preferences [19]; for example, working in another location away from home in exchange for a better wage or accepting a low wage in exchange of an interesting and professional job. Given this approach, compensability is admissible.

The CI proposed herein comprises three dimensions [3]: economic, professional and work-life balance. The economic dimension concerns the aspects related to the economic exchange between worker and employer, which is generally included in the formal employment contract. The professional dimension is related to the characteristics of the job, which influences workers' human capital accumulation by enhancing their employability. The work-life balance dimension involves the aspects that affect the workers' personal life and work relationships.

Each dimension comprises several elementary indicators. Table 1 shows the original variables used for the construction of the elementary indicators.

As evident from Table 1, the nature of variables can differ: dichotomous, ordinal and quantitative. The solutions often adopted in aggregating variables are either to convert all the variables to the same scale (such as dichotomous variables) or to attribute a numerical value to ordinal variables.

On the other hand, we propose to formulate every elementary indicator composing the CI as distance of the original variable from its minimum. The concept of distance is usually used when referring to quantitative variables; however, it can be extended to qualitative variables (dissimilarity). Therefore, the problem of different units of measurement can be engaged by considering a distance/dissimilarity as the new elementary indicator from which the CI can be built. A new concept of CI is conceived: the CI is relative to the minimum from which it is desirable to move away; this idea seems to particularly fit the field of CI as development measures of a subject, conceiving the worst possible subject as 'a calamity to leave as far behind as possible' [24].

As compensability is acceptable for our job quality CI, we must consider that the elementary indicators (and, even if to a small extent, also the dimensions) used in the construction of the CI could be correlated, so it is necessary to avoid the problem of 'double counting' the same phenomenon: for instance, if two variables composing a dimension are correlated and we ignore such correlation, the phenomenon described by the dimension weighs too much in the CI. To this aim, we consider also the correlation between indicators/dimensions in our CI formulation.

Table 1 Variables used in the construction of the composite indicator of job quality

Dimension	Variable	Description	Elementary indicator
Economic	Hourly wage	Quantitative: monthly wage/# of monthly working hours	$(X - X_{min})/R$
	Contractual stability	Ordinal: permanent job; open-ended job and self-employment; other (e.g. temporary work)	$(r(X))/(R- 1)$
Professional	Coherence degree-work	Ordinal: 0 (not at all) to 9 (a lot)	$(r(X))/(R- 1)$
	Usefulness of the degree	Ordinal: for performing your current job, (i) the university degree that you hold is specifically required, (ii) a graduate from a different major could obtain similar results, (iii) university degrees are not necessary, a high school degree could suffice, (iv) a qualification lower than high school could suffice	$(r(X))/(R- 1)$
	Enhancing skills	Ordinal: to what degree can you exploit your professional skills at work? (i) not at all, (ii) Not much, (iii) Quite, (iv) Very much	$(r(X))/(R- 1)$
	Career perspectives	Dichotomous: Yes, No	yes=1, no=0
	Team work	Dichotomous: Yes, No	yes=1, no=0
	Supervision of team work	Dichotomous: Yes, No	yes $= 1$, no $= 0$
Work-life balance	Working hours	Quantitative: 1-(weekly working hours-normalized 0–1)	$(X - X_{min})/R$
	Workplace distance	Ordinal: the residence province, the residence region, abroad or in an Italian region (different from the residence region)	$(r(X))/(R- 1)$

Note: The coding of some variables has been reversed because the CI formulation needs to express all the variables in the same direction: high values correspond to a high job quality.
$r(X) =$ rank(X); R = Range; $X_{min} =$ sample minimum

The final formulation of the CI is an extension of Gower's generalized distance [4, 9, 12, 17]:

$$CI_C = \frac{\sum_j^J \frac{\lambda_j \nu_j}{N_j} \left(\sum_{i=1}^{N_j} l_i w_i d_{cmi}\right)}{\sum_j^J \frac{\lambda_j \nu_j}{N_j} \left(\sum_{i=1}^{N_j} l_i w_i\right)} \quad (1)$$

where j is the index of the dimensions and i is the index of the elementary indicators composing each dimension; N_j is the number of indicators forming the j-th dimension (this is necessary because in this way, the importance of a

dimension does not depend on the number of indicators describing it); λ_j and v_j are the importance and correlation weights of the upper level of the structural hierarchy, respectively, that is, the importance and correlation of each dimension, respectively; l_i and w_i are the importance and correlation weights at the lower level for the elementary indicators; and d_{cmi} is the dissimilarity measure of subject c from minimum m with respect to variable i.

Referring to Gower's proposal [12], for dichotomous variables, $d_{cmi} = 0$ if the c-th subject shares the same categorization as its 'minimum' for variable i and $d_{cmi} = 1$ if it does not. For quantitative variables, the distance is calculated as the absolute value of the difference between the variable observed for subject c and its minimum m, standardized by the range R_i : $d_{cmi} = (x_{ci} - x_{mi})/R_i$. For ordinal variables, the formulation is similar to the quantitative case: R_i : $d_{cmi} = (rk(x_{ci}))/(Rk_i - 1)$, where $rk(x_{ci})$ is the rank of the c-th observation for the i-th ordinal variable (the first rank is 0) and Rk_i is the number of categories of the variable i. The distance d_{cmi} varies between 0 and 1 so that it is standardized. The overall CI has the same property.

The minimum to be considered for every indicator depends on the choice between the theoretical or sample minimum. In our case, this choice is relevant only for quantitative variables, given that theoretical and sample minimum for dichotomous and ordinal variables are the same (zero). For the two quantitative variables considered (working hours and hourly wage), we decided to use the sample minimum (i.e. 1 h for working hours and 2 euros for hourly wage) because no theoretical minimum has been universally established for these.

Correlation weight of each i-th indicator is a function of the correlation coefficients r_{il} between that indicator and all the other indicators in the dimension (indexed by l): $w_i = \Sigma_{l \neq i}(1 - |r_{il}|)$. Correlation or cograduation measures have been used for the computation of correlation weights, depending on the nature of the variables:

- two quantitative variables: the correlation coefficient is the common Bravais-Pearson coefficient.
- two ordinal variables: the Spearman correlation is considered.
- two dichotomous variables: the Phi correlation coefficient is regarded in this case [9].
- quantitative and dichotomous variable: the point-biserial coefficient [15] can be applied.
- quantitative and ordinal variable: the multiserial coefficient of Jaspen [13] is considered.
- ordinal and dichotomous variable: the rank biserial correlation coefficient [11] suits this case.

Correlation weights are considered only for elementary indicators.

With regard to importance weights, we address whether unit weights or differential weights are more appropriate. Several authors demonstrate that unit weights are preferable when the sample size is not large and/or a criterion measure is not available. For instance, Bobko et al. [1] demonstrate with a meta-analysis that unit

weights perform better than regression weights when the sample size is 75 or fewer and/or when R^2 is moderate or low. Given the size of our sample, unit weights do not necessarily perform better than differential weights. Furthermore, given the object of our study (job quality) and the population of interest (young graduates), we may expect that the weights assigned to the three dimensions will differ significantly one from the other. As a consequence, we compute differential weights. In order to verify whether our differential weights differ from unitary weights, we calculate the Spearman correlation between the rankings of the graduates sorted by the CI obtained using the two weighting methods—differential and unitary weights. In their meta-analysis based on fourteen studies and 3,182 participants, Bobko et al. [1] show that the correlation between the expert and the unit-weighted composite score is 0.99, with a credibility interval ranging from 0.94 to 0.99. In our case the Spearman correlation is 0.87, a value that is noticeably lower than that obtained by Bobko et al.

The definition of importance weights (calculated only for dimensions; equal weights are used for elementary indicators) is based on a hybrid approach [6], assuming that poor job quality may be a reason for job dissatisfaction. Respondents were required to express the level of job satisfaction (on a range between 1 and 10) with their job as a whole and by referring to a set of job characteristics. Weights are calculated from the standardized regression coefficients obtained through the ordinal logistic regression model, where the dependent variable is the overall job satisfaction score and the explanatory variables are the satisfaction scores for the job attributes considered in the dimensions. Each dimension is weighted using the arithmetic mean of standardized regression coefficients that refer to the proper job attributes. For example, for the economic dimension, we calculate the arithmetic mean of the standardized regression coefficients referring to satisfaction for wage and contractual stability. The weight for every dimension is then calculated by dividing each arithmetic mean by the sum of the three means related to the three dimensions.

The importance weights for the economic, professional and work-life balance dimensions are 0.241, 0.602 and 0.157.

We have tested the stability of the CI with two trials: initially, the original sample has been divided into two random sub-samples of equal size (without replacement) and subsequently, the sample has been divided into three random sub-samples of equal size and without replacement. The value of the CI and its dimension has been calculated separately for each sub-sample. With regard to the weights, both the importance and the correlation weights have been re-calculated for each random sub-sample. The results have shown that the three dimensions are stable in the sub-samples.

Moreover, the CI seems to actually measure what it was intended to measure (content validity). It can be assumed that, in general, the job quality of job seekers is lower than the job quality of the people who do not feel the need to change their job. In fact, those who are not satisfied with their jobs will probably try to change it in order to find a job that meets their expectations [5, 23]. We used the information (obtained from responses to the questionnaire of the Agorà survey) regarding the

intention to change job and the possible reasons of this choice. The question has the following possible answers: 1. I have never thought about leaving my job.; 2. I would leave to improve my compensation and the contractual arrangement; 3. I would leave to improve the work content and to have more opportunities to use my skills; 4. I would leave to work closer to home; 5. Another reason. We associated some of the answers from those who had thought about leaving their jobs to the job quality dimensions: in particular, the second answer corresponds to the economic dimension, the third answer to the professional dimension, and the fourth answer to the work-life balance dimension. The rationale for this validation procedure was that people who have thought of leaving their jobs for a specific reason (e.g., to improve their compensation) are likely to have a low score on the corresponding dimension of the CI. Starting from the answers to the question presented above, we calculated three dichotomous (dependent) variables, each of which is associated with one dimension of the JQCI: economic reasons to leave the job (1 = yes, 0 = no), professional reasons to leave the job (1 = yes, 0 = no), work-life balance reasons to leave the job (1 = yes, 0 = no). We performed a logistic regression for each dichotomous variable using as explanatory variables the scores on the three dimensions of the CI. We verified the content validity if, for instance, in the regression where the dependent variable is the dummy "Economic reasons to leave the job," only the economic dimension (as the explanatory variable) is statistically significant with a negative coefficient; that is, individuals considers leaving their jobs for economic reasons because the economic dimension of their job quality is low. Actually, the application of the three models show that only the dimension associated with the dependent variable has a significant and negative coefficient.

4 Job Quality of Disabled and Non-Disabled Graduates

Table 2 shows the mean scores of the CI and of the dimensions and elementary indicators for non-disabled and disabled graduates.

The most important result is that, on average, the CI is equal for the disabled and non-disabled graduates; however, this result comes from different dynamics. The economic dimension mean score is significantly higher for disabled graduates (0.55 vs. 0.44, $p = 0.0001$); this is probably because of the more favourable employment contracts that are offered to disabled graduates thanks to the Italian law. We found that 59 % of the disabled graduates work in the public sector (vs.10.8 % of the non-disabled graduates) and 45.8 % of the disabled graduates passed a public competition (vs.17.6 % of the non-disabled graduates) [2].

On the contrary, the professional dimension mean score is higher among the non-disabled graduates and the difference is mostly owing to the lack of consistency between the educational qualifications (university degree) of disabled graduates and the nature of work done by them (0.18 among disabled vs. 0.58 among non-disabled graduates, $p < 0.0001$).

Although the difference in the professional dimension mean score between the disabled and non-disabled graduates is lower than the difference in the economic

Table 2 Values of the mean scores of composite indicator, dimensions and elementary indicators for non-disabled and disabled graduates

Indicator	Non-disabled	Disabled	*p*-value
Composite indicator	**0.54**	**0.54**	**0.934**
Economic dimension	**0.44**	**0.55**	**0.000**
Elementary indicator of wage	0.26	0.25	0.760
Elementary indicator of contract	0.63	0.87	< 0.0001
Professional dimension	**0.56**	**0.51**	**0.091**
Elem. Ind. 'Degree of specialization'	0.68	0.68	0.894
Elem. Ind. 'Coherence degree-work'	0.58	0.18	< 0.0001
Elem. Ind. 'Supervision of team work'	0.34	0.25	0.215
Elem. Ind. 'Career perspectives'	0.41	0.44	0.693
Elem. Ind. 'Being in a team work'	0.68	0.86	0.003
Elem. Ind. 'Enhancing skills'	0.68	0.64	0.232
Work-life balance dimension	**0.62**	**0.61**	**0.710**
Elem. Ind. 'Distance home-work'	0.59	0.63	0.001
Elem. Ind. 'Working hours'	0.65	0.58	0.167

Note: *p*-value refers to the difference between non-disabled and disabled graduates

Table 3 Value of mean scores of composite indicator and its dimensions for disabled and non-disabled graduates by gender

Disabled	CI	Economic	Professional	Work-life balance
Men	0.54	0.54	0.47	0.62
Women	0.55	0.55	0.54	0.60
p-value	0.178	0.732	0.052	0.837
Non-disabled				
Men	0.57	0.46	0.61	0.60
Women	0.51	0.42	0.52	0.63
p-value	0.0173	0.222	0.0141	0.243

Note: *p*-value refers to the difference between men and women

dimension mean score, the CI results for the two groups are equal owing to the higher weight of the professional dimension.

Note (Table 3) that job quality is significantly higher for non-disabled men than non-disabled women, whereas this difference is negligible between disabled men and women. This difference can be attributed to the professional dimension; this is significantly higher among non-disabled men with respect to non-disabled women, whereas the contrary can be observed among disabled men and women. The majority of disabled women work in the public sector (71.4 % vs. 28.6 % among disabled men) and more than a quarter of the disabled women work in the life sciences area as nurses, physiotherapists or educators. These jobs are characterized by good coherence with employees' educational qualifications and consequently higher job quality from the professional perspective.

Table 4 Estimates of linear regression parameters on the composite indicator of job quality and its dimensions by presence of disability

Disabled	CI	Economic	Professional	Work-life balance
Intercept	−0.0963	−1.8562	−0.1527	0.7700*
Gender (*Ref: Men*)				
Women	0.1534	0.5963	0.2904	−0.2125
Type of degree (*Ref: 5 years single cycle*)				
Bachelor	−0.1230	0.4270	−0.3135	−0.1422
Master degree	−0.5514***	−2.1287**	−0.6213*	−0.8087**
Disciplinary area (*Ref: Humanities*)				
Life sciences	0.3617***	0.6358	0.6846**	−0.1282
Socio-economic	0.3669**	1.5148	0.3261	0.0677
Techn.-scientific	0.0763	2.3908**	0.2429	−0.8621**
Degree grade (*Ref: 91–100*)				
Grade <=90	−0.0453	−0.0937	−0.1393	0.3752
Grade>100	0.1707	0.8965	0.1552	0.3435
Working sector (*Ref: Private*)				
Public sector	0.1307	0.8738	0.0382	−0.0012
R^2	***0.282***	***0.174***	***0.172***	***0.224***
Non-disabled	CI	Economic	Professional	Work-life balance
Intercept	0.3421	0.1813	0.3069	0.8230*
Women	−0.2125**	−0.2033	−0.3678**	0.1425
Bachelor	−0.2938	−0.8676	−0.2754	−0.0983
Master degree	−0.2022	−0.5670	−0.1080	−0.3454
Life sciences	0.0869	0.4889	0.0568	−0.2128
Socio-economic	0.2314*	0.2098	0.4600**	−0.4407**
Techn.-scientific	0.3535**	0.5773*	0.5206*	−0.0962**
Grade <= 90	0.0994	−0.0641	0.4853*	−0.3038
Grade >100	0.0100	0.0660	0.1025	−0.3278**
Public sector	0.1999	0.1137	0.2203	0.4267**
R^2	***0.116***	***0.079***	***0.132***	***0.133***

Note: Dependent variable in regression is given by log (y/(1-y)), where y represents the composite indicator or one of its dimensions. * =< 0.1;** =< 0.05;*** =< 0.01.

These results are confirmed, ceteris paribus. In Table 4 the results of four linear regression models of the logit of composite indicator and the logit of its dimensions are shown. The results show that a degree in life sciences positively influences job quality for disabled graduates. Moreover, no significant differences are evident between disabled women and men in terms of job quality; however, significant differences exist between non-disabled women and men in this regard.

The linear regression models also highlight the negative effect of a master degree among disabled graduates both on the CI and its dimensions. Even if negative, these effects are not significant among non-disabled graduates. The results indicate that a higher specialization is not recognised, especially for disabled graduates.

Conclusions

The conclusions of our study consider both the methodological approach and the results.

A new formulation of Composite Indicator has been proposed to measure complex phenomena at micro-level, that is dealing with individuals or small groups of subjects (such as families). Information on single individuals can be collected as quantitative, ordinal and dichotomous variables. Thus, one of the main characteristics of this proposal is to consider variables of different nature. Other features of our Composite Indicator are to take into account the correlation among variables, and to express the overall measure in the form of a distance from a minimum, maintaining at the same time the hierarchical form of the Composite Indicator.

The usefulness of this approach is that it can be used both at macro and at micro-level for measuring a wide range of complex phenomena, for example to express development measures for various subjects (job quality, environmental development,. . .). It can be used not only for ranking aims, but also for rating, to state where a subject is positioned in the range (0–1) of definition of the Composite Indicator. Attention is paid on maintaining as much as possible both the original nature of the variables and of its relations, and the original nature of the multidimensional phenomenon. Finally, the formulation of our Composite Indicator can be considered as a form of weighted mean, easily understandable concept also for non-technical people, and this further expands its applicability.

From a methodological point of view, we are aware that our proposal does not completely solve the problem of variables of different nature, because the generalized distance used for the construction of the CI makes a sort of quantification (i.e. ranks) of ordinal and dichotomous variables. Some fields of research are approaching this problem (i.e.: some Multicriteria approaches, Benefit of the Doubt method, Partial Order Theory) in ways that depart from the classical approach. The proposed solutions are very interesting, but, at the same time, not easily understandable for non-technical people. We prefer to maintain the classical and simpler approach, also because more feasible in case of compensability.

Our study considered graduates from the University of Padua and we compared the job quality of disabled and non-disabled graduates. The job quality of disabled and non-disabled graduates originates from different job dimensions; although disabled graduates have the guarantee of contractual stability, they experience a lack of consistency between the jobs offered to them and their educational qualification (university education). Since the disabled and non-disabled graduates were also matched in terms of their disciplinary areas, this difference of consistency is not due to the different specializations between the two groups.

With regard to the study and labour rights of the disabled, the Italian legislation is one of the best in the world. However, it can only control objective aspects of the job, such as the contract. It cannot control the consistency between

jobs and educational qualifications, use of competences, and professionalism. These intangible aspects cannot be imposed by law; they need to be instituted by means of a culture pervading the whole society.

Bibliography

1. Bobko, P., Roth, P.L., Buster, M.A.: The uselfulness of unit weights in creating composite scores: a literature review, application to content validity, and meta-analysis. Organ. Res. Meth. **10**, 689–709 (2007)
2. Boccuzzo, G., Fabbris, L.: How do the disabled graduates achieve and spend their human capital gained at university? In: Fabbris, L. (ed.) Indicators of University Education Effectiveness, pp. 105–118. McGraw-hill, Milano (2012)
3. Boccuzzo, G., Gianecchini M.: Measuring Young Graduates' Job Quality through a Composite Indicator. Social Indicator Research. DOI: 10.1007/s11205-014-0695-6 (2014)
4. Cox, T., Cox, M.: A general weighted two-way dissimilarity coefficient. J Classif. **17**, 101–121 (2000)
5. de Bustillo Llorente, R.M., Macías, E.F.: Job satisfaction as an indicator of the quality of work. J. Soc. Econ. **34**, 656–673 (2005)
6. Decancq, K., Lugo, M.A.: Weights in multidimensional indices of wellbeing: an overview. Economet. Rev. **32**, 7–34 (2013)
7. Erhel, C., Guergoat-Larivière, M.: Job quality and labour market performance. Economic Policy CEPS Working Document, vol. 330 (2010)
8. Fabbris, L.: Il progetto Agorà dell'Università di Padova. In: Fabbris, L. (ed.) Dal Bo' all'Agorà. Il capitale umano investito nel lavoro, pp. III-XLV. Cleup, Padova (2010)
9. Fabbris, L.: Statistica multivariata, analisi esplorativa dei dati. Mc Graw-Hill, Milano (1997)
10. Foley, K., Schwartz, S.: Earnings supplements and job quality among former welfare recipients: evidence from the self-sufficiency project. Relat. Ind./Ind. Relat. **58**, 258–286 (2003)
11. Glass, G.: Note on rank-biserial correlation. Educ. Psychol. Meas. **26**, 623–631 (1966)
12. Gower, J.: A general coefficient of similarity and some of its properties. Biometrics **27**, 857–871 (1971)
13. Jaspen, N.: Serial correlation. Psychometrika **11**, 23–30 (1946)
14. Kalleberg, A.L., Reskin, B.F., Hudson, K.: Bad jobs in America: standard and nonstandard employment relations and job quality in the United States. Am. Sociol. Rev. **65**, 256–278 (2000)
15. Lev, J.: The point biserial coefficient of correlation. Ann. Math. Stat. **20**, 125–126 (1949).
16. Loveman, G.W., Tilly, C.: Good jobs or bad jobs-evaluating the American job creation experience. Int. Lab. Rev. **127**, 593–611 (1988)
17. Maron, L.: A Proposal for a new composite indicator based on a weighted measure of distance. Master degree thesis, In Statistical Science, University of Padua (2012)
18. McGovern, P., Smeaton, D., Hill, S.: Bad jobs in Britain nonstandard employment and job quality. Work Occupations **31**, 225–249 (2004)
19. Muñoz de Bustillo, R., Fernández-Macías, E., Antón, J., Esteve F.: Measuring More than Money: the Social Economics of Job Quality. Edward Elgar, Cheltenham (2011)
20. Nagy, M.: Using a single-item approach to measure facet job satisfaction. J. Occup. Organ. Psychol. **75**, 77–86 (2002)
21. OECD: Handbook on constructing composite indicators: Methodology and user guide www.oecd.org/std/42495745.pdf (2008)
22. Skalli, A., Theodossiou, I., Vasileiou, E.: Jobs as Lancaster goods: facets of job satisfaction and overall job satisfaction. J. Soc. Econ. **37**, 1906–1920 (2008)
23. Souza-Poza, A., Henneberger, F.: Analyzing job mobility with job turnover intentions: an international comparative study. J. Econ. Issues **XXXVIII**, 113–137 (2004)

24. UNESCO: Synchronic and Diachronic Approaches in the UNESCO project on Human Resources Indicators. Wroclaw Taxonomy and Bivariate Diachronic Analysis http://unesdoc.unesco.org/images/0000/000008/000801eb.pdf (1972)

Business Failure Prediction in Manufacturing: A Robust Bayesian Approach to Discriminant Scoring

Maurizio Baussola, Eleonora Bartoloni, and Aldo Corbellini

Abstract

This paper provides a methodological analysis of credit risk in manufacturing firms. By using a representative sample of both healthy and bankrupted firms during the period 2003–2009 we provide an in-depth comparison of the standard discriminant approach for bankruptcy prediction based on a logistic regression model and a Robust Bayesian Approach. We conclude that the use of a robust GLM regression methodology enables us to provide a more accurate separation between sound and unsound firms thus suggesting that this methodological framework may be used to achieve a more reliable measure of firms credit worthiness.

Keywords

Bankruptcy • Discriminant analysis • Forward search • Robust GLM regression

M. Baussola (✉)
Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore, Piacenza, Italy
e-mail: maurizio.baussola@unicatt.it

E. Bartoloni
ISTAT, Sede per la Lombardia, Milano, Italy
e-mail: bartolon@istat.it

A. Corbellini
Dipartimento di Economia, Università di Parma, Parma, Italy
e-mail: aldo.corbellini@unipr.it

M. Carpita et al. (eds.), *Advances in Latent Variables*, Studies in Theoretical and Applied Statistics, DOI 10.1007/10104_2014_8, Published online: 12 November 2014

1 Introduction

The study of a firm's financial performance is relevant in the context of the present economic downturn, as it allows us to understand whether significant threats to economic recovery do exist and whether investment decisions by firms may stimulate and sustain economic growth in the medium to long term. A firm's decision to invest may crucially be affected by its level of financial constraint. Thus, an understanding of the distribution of such financial constraints is particularly relevant with respect to innovative investment, which represents the key to business success.

Indeed, although several definitions of credit or financial constraint have been proposed by the relevant literature—[7] refer to a wedge between the internal and external cost of funds, while [6] refers to a situation in which there is a wedge, sometimes large, between the rate of return required by an entrepreneur investing his own funds and that required by external investors—there is currently no general agreement on how financially-constrained firms can be identified empirically. The debate concerning the measurement of financial friction at the firm level may gain interesting input from the field of business failure prediction. The main goal here is to predict bankruptcy risk, i.e. to develop models of financial failure at the firm level before this actually happens. Although business failure has long been debated in both economic and accountancy research [4], accurate credit risk analysis has become even more important today than it was in the past due to the recent global financial crisis, which has demonstrated how difficult it is to measure and manage business distress. In this contribution we provide a thorough analysis of credit risk in manufacturing firms during the period 2003–2009 by combining the standard discriminant analysis (DA) with the Forward Search [3] in a Bayesian perspective framework as described by [8].

In the field of business failure prediction discriminant analysis has been widely used. This approach is essentially based on the idea that a firm's probability of default may be estimated by using a set of key variables. These variables, appropriately combined together, produce a range of quantitative scores, which can be used as a classification tool when combined with an appropriate cut-off point. We refer to the seminal work by [1] and further developments [2, 5], which employ a linear discriminant model based on accounting data of failed and non-failed firms in order to determine a firm's bankruptcy risk. Ohlson [9] proposed a conditional logistic model that has the advantage of overcoming problems associated with the assumption of normality and equal covariances that, the linear discriminant model could require. The peculiar feature of this approach is the way a model's precision is tested for by considering both classification and future prediction accuracy. Classification accuracy is assessed on the original database, that is the data-set used in order to specify the model. Following this, prediction accuracy is tested for by using a new data set, in order to assess how well the model works for future predictions. The use of the Forward Search coupled with a Bayesian probit regression model allows us to detect multiple outliers more efficiently compared to

traditional exploratory techniques, from now on we will call this method Robust Bayesian Approach (RBA). In addition, the application of a Bayesian method to the probit specification, on one hand, helps us with the right prior to estimate default probabilities and, on the other hand, allows us to reduce interchange rates from one step to another of the Forward Search, driving out smooth confidence curves and potentially robustifying the algorithm.

2 Data Description

Our main sample of firms is derived from the tenth Unicredit Survey on Manufacturing Firms (2009). This sample is composed of more than five thousand firms representative of the manufacturing sector and extracted from the AIDA data base. A rich set of information is collected by this survey, including firm-specific characteristics, investment and innovative activities. This starting sample has been inflated with a rich set of accounting data. The economic and financial information derived from firms' balance sheets has allowed us to derive the financial indexes used in the credit scoring procedures which will be described in the following sections. Bankruptcy data have been collected from the AIDA data base. We extracted a sample of 150 firms which went bankrupt during the years 2005 and 2006. Balance sheet information refers to years 2003 and 2004 in order to have an adequate time span difference (not less than 1 year) between the last relevant balance sheet and the bankruptcy date.

3 The Logistic Regression Model DA

We estimate the default probability of a firm by using a logistic discriminant function defined as follows:

$$\pi(x_i) = \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}} = \frac{1}{1 + e^{-\beta x_i}} \qquad i = 1, \ldots, n. \tag{1}$$

y_i is our binary dependent variable, which assumes the value of 1 if we observe a default event between years 2005 and 2006 and 0 otherwise and $x_1, \ldots, x_k$ is the vector of covariates, i.e. firm-specific characteristics and financial indexes which are observed in years 2003 and 2004.

We have included a set of variables which are commonly considered good predictors of the outcome event in the relevant literature:

- a measure of a firm's leverage (LEV), the ratio of total debts to net capital, which is expected to affect the default probability positively, as a highly-leveraged structure may worsen the perceived financial risk;
- a measure of short-term indebtedness (CL_S), the ratio of current liabilities to sales, whose expected sign is positive, given that a firm with a high short-term

debt may find it difficult to borrow additional resources to finance its short run activities and, thus, may be close to insolvency;
- another similar indicator, the ACID ratio; this measures the extent to which short-term debt is covered by short term liquidity. Creditors prefer a high ACID ratio as it reduces their risk. We thus expect a negative sign;
- firm operating profitability (ROS), proxied by the ratio of operating margins to sales. We expect a negative effect on the default risk, as the higher a firm's profitability the higher the flow of internal resources available to cover debt exposure should be;
- the firm's interest burden, proxied by the interest payment to sales (IR) ratio, which is expected to positively affect the default probability given that a high interest burden may worsen the financial risk associated with external finance. We have used a dummy variable which assumes the value of 1 when a firm shows an interest burden ratio higher than 5 %, which identifies the last 5 % of the IR distribution, and 0 otherwise, in order to capture the effect of those firms which are potentially financially constrained;
- finally, structural characteristics, captured by variables AGE (years) and SIZE, proxied by a firm's total assets (logarithmic values). We expect a negative effect of both these variables, as agency costs related to indebtedness are expected to be higher for those firms with a low reputation or contractual power, such as those which are smaller or less well established.

We estimate default probabilities within one and 2 years. In the first case the model is computed by using predictors observed in the year 2004, while in the second case we use information for the year 2003.[1] In both models our variables present the expected signs, although it is worth noting that the explanatory power is higher when information 2 years before bankruptcy is used. This evidence suggests that the choice of an adequate lead time span is a relevant point and needs to be taken into account. In our case, by using accounting information from 2 years prior the default event, we can build a more accurate prediction model.

4 Robust Bayesian Approach, RBA

In evaluating prediction accuracy with standard logistic regression model, from now on DA, there is no way of adjusting the cut-off point for the distribution in order to reduce simultaneously the two types of classification errors, that is the error of classifying a sound firm as unsound (Type I error) and the error of classify an unsound firm as sound (Type II error).

In practice, as there is a trade-off between the two types of error, a pragmatic rule is adopted depending on the specific aim of the classification and, therefore, on the

[1] Regression results available on request.

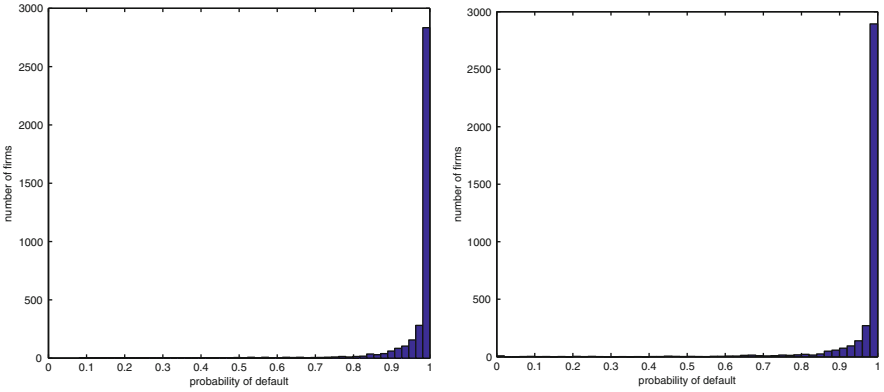

Fig. 1 Standard DA—histogram of the probability of belonging to the group of healthy firms year 2003 (*left panel*) and year 2009 (*right panel*)

characteristics of the users of such financial information.[2] Indeed, a bank which is evaluating a firm's financial position is probably more interested in minimizing the cost of making a bad investment (Type II error) due to lending funds to a potentially defaulting customer, whereas a shareholder in an innovative firm may be willing to reduce the cost of under-investment (Type I error) resulting from not taking advantage of an investment opportunity.

The application of standard logistic regression produces the frequency distribution of the estimated probabilities of belonging to the group of healthy firms given in Fig. 1. The left panel of this figure, shows that all firms, apart few exceptions, have an estimated probability of belonging to the group of healthy firms that is greater than 0.9. A similar effect takes place for the estimated probabilities referred to year 2009 (right panel).

Given these limitations of standard DA, we show that the use of a RBA can help us to better separate the two groups. As in the standard DA approach, we use the same data referred to year 2003. The RBA underlying model is a *probit* link function, which has some convenient statistical proprieties in the Bayesian framework. Here Φ is the standard normal cdf and the probability of default for firm i which we denote with $g(\mu_i)$ is linked to the set of explanatory variables previously described as follows: $g(\mu_i) = \Phi^{-1}(\mu_i)$ with $i = 1, \ldots, n$. The steps of the procedure are as follows: first we start with a robust subset of statistical units. In order to achieve this purpose we use Least Trimmed Squares to find the best

[2] Our results, available on request, show that if a cutoff point of 0.02 is fixed, a Type II Error of 0.5 is obtained with the 2003 model (84 % of bankruptcy cases correctly predicted). However, as at this cutoff point we also wrongly classify as unsound 16 % of healthy firms, we prefer to accept a small increase in Type I Error in order to reach a better classification for the group of healthy firms. Thus, a cutoff point of 0.04 seems to be a reasonable compromise (74 % of bankruptcy cases correctly predicted and 90 % of sound firms correctly classified).

statistical units belonging to group of healthy firms. We repeat the same procedure to find the subset of the defaulted firms. At this point, we start the Forward Search monitoring the deviance residuals and the misclassification index *(mis)* defined as follows:

$$mis = \sum_{i}^{n} (y_i - \hat{\pi}(x_i))^2 \tag{2}$$

where y_i is a dichotomic variable which is equal to 0 if the firm is healthy or is equal to 1 if the firm is defaulted, and $\hat{\pi}(x_i)$ is the estimated probability of default. Fig. 3 shows the trajectory of the misclassification index along the Forward Search. We pick the function minimum, corresponding to a specific subset, S^*_{min}, before the curve starts behaving chaotically, an anomalous behavior caused by lack of convergence of the probit regression, as outliers enter the search. Using the subset S^*_{min}, we calculate the corresponding $\hat{\beta}$ coefficient of the probit regression. Then we take $\hat{\beta}$ as a prior and we start the Forward Search using Bayesian probit regression with the following posterior distribution:

$$\pi(\beta|y, X) \propto |X'X|^{1/2} \Gamma((2k-1)/4)(\beta'(X'X)\beta)^{-(2k-1)/4} \pi^{-k/2} \times \\ \times \prod_{i=1}^{n} \Phi(x_i'\beta)^{y_i} [1 - \Phi(x_i'\beta)]^{1-y_i}. \tag{3}$$

From the Bayesian perspective we use Metropolis Hastings algorithm with a Zellner's prior (Marin et al. 2012) to obtain a sampling distributions of $\hat{\beta}$. With this distribution we are able to perform a sensitivity analysis along the credibility interval of the β parameter, unlike using point estimate obtained applying maximum likelihood. Besides, we are also given the opportunity to introduce some (sensible) prior information, which could help us to better discriminate candidate default firms. In a way, exploring and analysing thoroughly the dataset in year 2003 we can model prior information to be applied to year 2009 proposal model. We can see that the algorithm is working properly, both in the predictive stage and in the evaluation phase, the separation between defaulted and healthy firms is quite high, as shown respectively in Fig. 2, left panel, year 2003 and Fig. 2, right panel, year 2009. Table 1 and Fig. 2 show us the frequency distribution of the estimated probabilities of default by the model using the classes $(0,\ 0.1], (0.1,\ 0.2], \ldots, (0.9,\ 1]$ respectively for years 2003 and 2009. In Fig. 3 we show the differences between the frequentist implementation of the Forward Search and the Bayesian version of Forward Search. In the first panel we see a chaotic behaviour on the right part of the graph: large oscillations of the misclassification index are caused by either lack of convergence or when the interchange rate of the units entering and exiting the Forward Search is

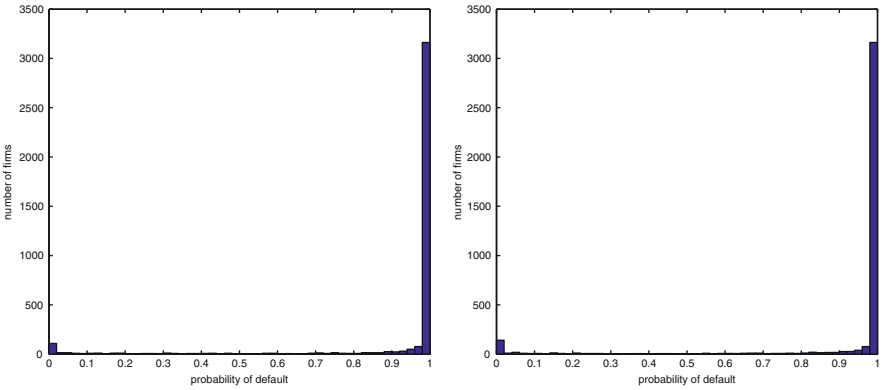

Fig. 2 RBA—histogram of the probability of belonging to the group of the healthy firms, evaluation stage—year 2003 (*left panel*), predictive stage—year 2009 (*right panel*). Note that the estimated probabilities are highly separated

Table 1 Distribution of estimated probabilities of default

Year	(0, 0.1](%)	(0.1, 0.2](%)	(0.2, 0.3](%)	(0.3, 0.4](%)	(0.4, 0.5](%)
2003	3.94	1.03	0.66	0.85	0.79
2009	4.82	0.89	0.87	0.55	0.58
Year	(0.5, 0.6](%)	(0.6, 0.7](%)	(0.7, 0.8](%)	(0.8, 0.9](%)	(0.9, 1](%)
2003	0.58	0.66	1.24	1.93	88.31
2009	0.58	1.08	0.92	2.05	87.65

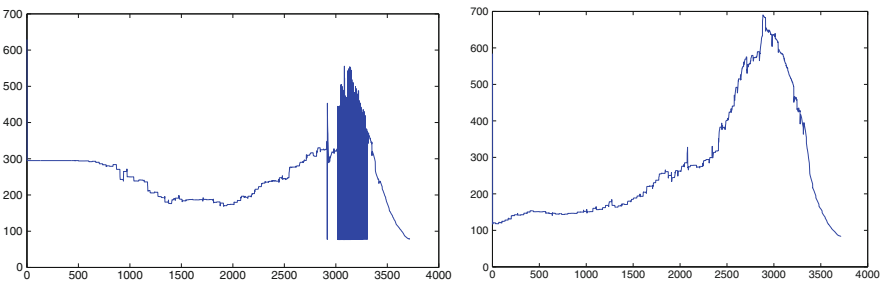

Fig. 3 Monitoring of misclassification index—please note the chaotic behavior in the first panel where algorithm convergence is missing and/or the interchange rate of units entering the Forward Search is high

set too high. In the second panel we show the results using the Bayesian Forward Search: we now see a smoother curve of the misclassification index which is both the result of two conditions: the first is that now the algorithm is converging at every step of the Forward Search, the second is that a low interchange rate between units belong or not to the Search is obtained.

Table 2 $\hat{\beta}$ Comparison between standard DA and RBA

Var.	$\hat{\beta}_{standardDA}$		$\hat{\beta}_{RBA}$	
	Estimate	*P*-value	Estimate	*P*-value
Constant	5.9096	0.0009	9.0086	8.78E−10
ACID	−1.2726	0.0172	−2.0078	7.47E−07
LEV	0.0225	0.0002	0.0736	1.10E−10
CL_S	0.6633	0.1473	0.0848	0.4930
ROS	−7.8080	0.0009	−12.3619	3.01E−10
dIR	1.0454	0.0085	3.0383	1.46E−10
L_TA	−0.1504	<0.0001	−0.8082	4.26E−05
AGE	−0.4106	0.0004	−0.1892	5.01E−15
dNW	−2.0611	<0.0001	−3.6294	1.33E−16
dNE	−1.2579	0.0002	−3.1568	1.79E−13
dC	−0.5496	0.1130	−1.7431	5.86E−07

Conclusive Remarks

Although business failure has long been debated in both economic and accountancy research, accurate credit risk analysis has become even more important today than it was in the past due to the recent global financial crisis, which has demonstrated how difficult it is to measure and manage business distress. This contribution represents a step forward with respect to standard discriminant approach DA to credit scoring.

We have set up an appropriate default probability model which has been tested by using both standard DA and RBA. Both methods provide good performances in terms of expected signs and significance of the selected regressors (Table 2), although it is worth noting that the *P*-values of the robust estimation show greater significance levels.

However, the use of the Forward Search technique combined with a robust GLM regression has allowed us to reach a separation between sound and unsound firms which is more accurate compared to standard DA. This result also allows us to overcome one of the main criticism of traditional logistic scoring applied to bankruptcy prediction, i.e. the fact that it is inappropriate for predicting a rare event, such as bankruptcy, as it requires the selection of an adequate proportion of failed firms in the final sample, in order not to underestimate bankruptcy probabilities.

It is worth stressing that the study of financial performance at the firm level could not be limited to a simple separation between sound and unsound firms. Indeed, a significant advance could be the attribution of specific credit worthiness judgments within such a robust technique, also using different specifications, e.g, multinomial logit.

References

1. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Finance **23**(4), 589–609 (1968)
2. Altman, E.I., Haldeman, R.G., Narayanan P.: A new model to identify bankruptcy risk of corporation. J. Bank. Finance **1**(1), 29–54 (1977)
3. Atkinson, A.C., Riani, M.: Forward search added variable *t* tests and the effect of masked outliers on model selection. Biometrika **89**, 939–946 (2002)
4. Bartoloni, E., Baussola, M.: Financial performance in manufacturing firms: a comparison between parametric and non-parametric approaches. Bus. Econ. **49**(1), 32–45 (2014)
5. Deakin, E.B.: A discriminant analysis of predictors of business failure. J. Acc. Res. **10**(1), 167–179 (1972)
6. Hall, B.: The financing of research and development. Oxf. Rev. Econ. Policy **18**, 35–51 (2002)
7. Kaplan, S.N., Zingales, L.: Do financing constraints explain why investment is correlated with cash flows. Q. J. Econ. **112**(1), 169–215 (1997)
8. Marin, J.-M., Pudlo, P., Robert, C.P., Ryder, R.: Approximate Bayesian computational methods. Stat. Comput. **22**(1), 1167–1180 (2012)
9. Ohlson J.A.: Financial ratios and the probabilistic prediction of bankruptcy. J. Acc. Res. **18**(1), 109–131 (1980)