Available online at www.sciencedirect.com

ScienceDirect

Procedia Economics and Finance 17 (2014) 10 – 19

Procedia Economics and Finance

www.elsevier.com/locate/procedia

Innovation and Society 2013 Conference, IES 2013

A new scaling proposal for handling ordinal categorical variables in Co-Inertia (-PLS) Analysis

Antonio Lucadamo[a,*], Pietro Amenta[a]

[a]*Department of Law, Economics, Management and Quantitative Methods, University of Sannio, 82100, Benevento, Italy*

Abstract

In order to investigate the symmetrical relationships between several sets of variables, or regress one or more quantitative response variables on a set of variables of different nature, it is well known that it is necessary to transform non-quantitative variables in such a way that they can be analyzed together with the others measured on an interval scale.

This paper suggests a proposal to cope with the problem of the treatment of ordinal qualitative variables in Co-Inertia(-PLS) Analysis. In the literature there are different proposals based on the application of known statistical techniques to quantify ordinal variables. The approach consists in quantifying each non-quantitative variable according to the empirical distributions of the variables involved in the analysis assuming the presence of a continuous underlying variable for each ordinal indicator.

Keywords: Co-Inertia Analysis; Partial Least Squares Regression; Ordinal categorical variables; Quantification; Sensorial analysis.

1. Introduction

In applied or theoretical contexts we have often to deal with the study of numerical data tables obtained in experimental applications. The study of the complex structure of these data often requires the use of multivariate analyses in order to investigate the relationships between several sets of variables. We have to take moreover into account that in experimental data analysis a subset of variables can play a non symmetrical role towards the others. An example of this dependence framework can be found in sensory evaluations. The traditional way of evaluating sensory influences on overall liking consists in studying statistical links between explanatory sensory descriptive variables (sensory attributes evaluated for each product by trained judges) and criterion hedonic variables (scores given by consumers to the same products). The commonly used multivariate method is the External Preference Mapping (MacFie and Thomson, 1988; Schlich, 1995) which consists of two separate steps: a principal component analysis (or a generalized procrustes analysis) of sensory data gives some new synthetic sensory variables (sensory components) which summarize the main sensory differences among products; afterwards, the hedonic response is regressed on these main sensory components using a linear or quadratic model. Other approaches (Kvalheim, 1988; Huon de Kermadec F. et al., 1996, 1997) are based on the Partial Least Squares Regression (Wold et al. 1993). In the same context, in order

[*] Corresponding author. Tel.: +39-0824-305763; fax: +39-0824-305703.
E-mail address: antonio.lucadamo@unisannio.it

doi:10.1016/S2212-5671(14)00873-9

to study symmetrical relationships between subjective evaluations and measures of foodstuffs, Co-Inertia Analysis (Chessel and Mercier, 1993) (hereafter COA) has been proposed (Sabatier et al., 1992)

It is, anyway, often necessary, before developing the analysis, to consider the nature of the data because we can have the concomitant presence of quantitative and categorical variables measured at different scale levels (interval, nominal or ordinal). It is necessary then to transform non-quantitative variables in such a way that they can be analyzed together with the other variables measured on an interval scale. Moreover, data relative to the quality perceived by a service or a product have often an ordinal scale. The problem is that all these kind of data are not directly comparable (Green, Tull, 1988). This aspect is often ignored, but, if there is not a transformation in the data, also the use of a simple index, as the mean, is not applicable, because the ordinal scale is only a preference ranking. In these circumstances it is then necessary to determine a criterion to convert on a metric scale ordinal measurements. A simple technique is the so-called ”direct quantification” that hypothesizes that the modalities of a qualitative character are at the same distance (Likert, 1932). This hypothesis is not respected in many cases and furthermore it may lead to strange results. In fact, Nishisato (2005) showed as if we consider two ordered categorical variables that, looking at contingency table, have a strong relation, they may result uncorrelated if we consider Likert scores. For this reason, it is necessary to transform the data in linear and quantitative measures, graduated on the whole space of real numbers (Wright, Linacre, 1989). There are different techniques that allow to obtain new scores for ordinal measurements. The most used are the monotone regression of Kruskal (1965), the Rating Scale Models in the Rasch Analysis (Wright, Masters, 1982) and the psycometric approach of Thurstone (Zanella, 1999). This last approach assumes the presence of a continuous underlying variable for each ordinal indicator. Other approaches are those proposed by Guttman (1941), Fisher (1946) and Hayashi (1952). According to these methods, scores are assigned optimally in some objective and operational sense to categories of qualitative variables (Tanaka, 1979). Hayashi first quantification method is the core of the proposal by Russolillo and Lauro (2011) for handling categorical predictor variables in Partial Least Squares Regression. The approach consists in quantifying each non-quantitative predictor according to this quantification method, using the dependent variable (or, in the multivariate case, a linear combination of the response variables) as an external criterion. This proposal seems to not assure the ordinal property of the new scores if we are in presence of an ordinal categorical variable. They consider in addition only predictive (not response) categorical variables with respect to only quantitative response variables. In the OLS framework, MORALS (Young et al., 1976) and ACE (Breiman et al, 1985) algorithms are the most largely used in literature to optimize the transformation functions according to the multiple or canonical correlation criterion.

According to psychometric approach of Thurstone, we consider the empirical distributions of the variables involved in the analysis. If we have high frequencies on the first or the last modalities (e.g. customer satisfaction survey) then it could be more efficient using the standardized exponential distribution instead than normal one, fitting better the data in this way. This leads us to suggest a mixed quantification based on all these theoretical distributions. Finally, after applying the suggested mixed quantification, we perform two examples of Co-Inertia Analysis (-PLS) on real datasets.

2. Introducing Co-Inertia (-PLS) Analysis: basic definitions and notations

In order to study symmetrical interdependence relationships, several techniques, originated from Canonical Correlation Analysis or from Co-Inertia Analysis and their generalizations, have been proposed. However, CCA could create highly correlated linear combinations but not necessarily the most explicative ones. COA, based on the covariance criterion, has been proposed to improve the Correlation Analysis. Co-Inertia (-PLS) analysis (hereafter COA-PLS) borns instead in order to study the asymmetrical dependence relationships between two groups of variables and it is based on the covariance criterion like COA. It can be easily regarded as the asymmetrical extension of COA and, in meantime, as a metric generalization of Partial Least Squares Regression (Cazes,1997) using the statistical study notation (Escoufier, 1987).

COA is known, even if not by that name, and practiced in several fields. It is very famous in ecology by the papers of Chessel and Mercier (1993) and Doledec and Chessel (1994). In the atmospheric sciences, where it is well known as Singular Value Decomposition Analysis (SVD), has been popularized by Bretherton *et al.* (1992) and Wallace *et al.* (1992) in order to detect temporally synchronous spatial patterns even if its first use in climatology was apparently by Prohaska (1976). In this context, Predictor Analysis (Thacker, 1999) can be regarded as a metric-based SVD of the cross-correlation matrix. It is very popular also in the social sciences where it belongs to a class of methods of matching matrices. Van de Geer (1984) referred to it as the MAXDIFF criterion. Previously, Tucker (1958) introduced this method, with the name Inter-Battery Factor Analysis, in order to find common factors in two batteries of tests presented to the same group of statistical units. It was also applied to sensorial data (e.g.: Vivien *et al.*, 2001, Blackman *et al.*, 2010). Finally, it is well known also in the study of behavioral teratology with the name Partial Least Squares–SVD where Sampson *et al.* (1989) introduced it in a study of the relationship between fetal alcohol exposure and neurobehavioral deficits.

Let $(\mathbf{X}, \mathbf{Q_X}, \mathbf{D})$ be a statistical study (Escoufier, 1987) associated with the matrix $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}^T$ of order $n \times p$, collecting a set of p quantitative/qualitative variables observed on n statistical units where $\mathbf{D} = diag(d_1, ..., d_n)$ specifies the (diagonal) weights metric in the vectorial space $\Re^n$ of variables with $\sum_{i=1}^{n} d_i = 1$ and ^{T} is the transpose symbol. Without loss of generality we can assume uniform weights ($d_i = 1/n$) for $\mathbf{D}$ in this paper. $\mathbf{Q_X}$ is a $(p \times p)$ non negative definite (hereafter *nnd*) matrix defining the metric measuring the distance between the data vectors $\mathbf{x}_j$, $\mathbf{x}_k$ of two statistical units j, k in $\Re^p$ given by $(\mathbf{x}_j - \mathbf{x}_k)^T \mathbf{Q_X} (\mathbf{x}_j - \mathbf{x}_k)$. Let $(\mathbf{Y}, \mathbf{Q_Y}, \mathbf{D})$ be the statistical study associated with the matrix $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}^T$ of order $(n \times q)$, collecting a second set of q (quantitative/qualitative) variables observed on the same n statistical units where $\mathbf{Q_Y}$ is the $(q \times q)$ *nnd* metric of the statistical units in $\Re^q$. We assume that both sets of variables are mean centred with respect to $\mathbf{D}$, i.e. the weighted mean value of each column in $\mathbf{X}$ and $\mathbf{Y}$ is set to zero ($\mathbf{1}_n^T \mathbf{D} \mathbf{X} = \mathbf{0}$ with $\mathbf{1}_n$ unitary column vector).

In order to study the common geometry of the statistical triplets $(\mathbf{X}, \mathbf{Q_X}, \mathbf{D})$ and $(\mathbf{Y}, \mathbf{Q_Y}, \mathbf{D})$, Co-Inertia Analysis (Chessel and Mercier, 1993) seeks linear combinations of the data $\mathbf{t}_i = \mathbf{X}\mathbf{Q_X}\mathbf{w}_i$ and $\mathbf{u}_i = \mathbf{Y}\mathbf{Q_Y}\mathbf{c}_i$ ($i = 1, ..., s$, $s = \min(p, q)$) with the maximum covariance

$$\begin{cases} \max\limits_{\mathbf{w}_i\mathbf{c}_i} \operatorname{cov}^2(\mathbf{t}_i, \mathbf{u}_i)_\mathbf{D} = (\mathbf{w}_i^T \mathbf{Q_X} \mathbf{X}^T \mathbf{D} \mathbf{Y} \mathbf{Q_Y} \mathbf{c}_i)^2 \\ ||\mathbf{w}_i||^2_{\mathbf{Q_X}} = 1 \\ ||\mathbf{c}_i||^2_{\mathbf{Q_Y}} = 1 \end{cases} \tag{1}$$

such that the unknown weight vectors $\mathbf{w}_i$ and $\mathbf{c}_j$ satisfy the constraints $\mathbf{w}_i^T \mathbf{Q_X} \mathbf{w}_i = \mathbf{c}_i^T \mathbf{Q_Y} \mathbf{c}_i = 1$ and $\mathbf{w}_i^T \mathbf{Q_X} \mathbf{w}_j = \mathbf{c}_i^T \mathbf{Q_Y} \mathbf{c}_j = 0$ for $i \neq j$ (that is $\mathbf{W}^T \mathbf{Q_X} \mathbf{W} = \mathbf{I}$ and $\mathbf{C}^T \mathbf{Q_Y} \mathbf{C} = \mathbf{I}$). The COA criterion can also be written as $cov^2(\mathbf{t}_i, \mathbf{u}_i) = cor^2(\mathbf{t}_i, \mathbf{u}_i) \times var(\mathbf{t}_i) \times var(\mathbf{u}_i)$ where $cor^2(\mathbf{t}_i, \mathbf{u}_i)$ is the square cosinus of the angle between $\mathbf{t}_i$ and $\mathbf{u}_i$ with $var(\mathbf{t}_i) = \mathbf{w}_i^T \mathbf{Q_X} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q_X} \mathbf{w}_i$, $var(\mathbf{u}_i) = \mathbf{c}_i^T \mathbf{Q_Y} \mathbf{Y}^T \mathbf{D} \mathbf{Y} \mathbf{Q_Y} \mathbf{c}_i$. So maximizing this criterion we also maximize the correlation between the components $(\mathbf{t}_i, \mathbf{u}_i)$ and their respective variances, simultaneously. Note that the square of the entity cor(.) is maximized via Canonical Correlation Analysis (CCA) while a co-inertia axis maximizes $\operatorname{cov}^2(.)$. COA makes then a compromise between a CCA of the two sets $\mathbf{X}$ and $\mathbf{Y}$ and the Principal Component Analyses of matrices $\mathbf{X}$ and $\mathbf{Y}$.

Computationally $\mathrm{COA}(\mathbf{X}, \mathbf{Y})_{\mathbf{Q_X}, \mathbf{Q_Y}}$ amounts to the GSVD of the matrix $\mathbf{X}^T\mathbf{D}\mathbf{Y}$with the row metric $\mathbf{Q_X}$ and the column metric $\mathbf{Q_Y}$ and it is denoted $\mathrm{GSVD}(\mathbf{X}^T\mathbf{D}\mathbf{Y})_{\mathbf{Q_X}, \mathbf{Q_Y}}$. This method is also defined by the analysis of the statistical studies $(\mathbf{Y}^T\mathbf{D}\mathbf{X}, \mathbf{Q_X}, \mathbf{Q_Y})$. The pairs of axes $\mathbf{w}_i$ and $\mathbf{c}_j$ are obtained by the eigenvectors $\mathbf{g_X}$ and $\mathbf{g_Y}$ associated to the eigen-decomposition

$$\mathbf{Q}_X^{1/2} \mathbf{X}^T \mathbf{D} \mathbf{Y} \mathbf{Q_Y} \mathbf{Y}^T \mathbf{D} \mathbf{X} \mathbf{Q}_X^{1/2} \mathbf{g_X} = \lambda \mathbf{g_X} \tag{2}$$

(with $p < q$) or $\mathbf{Q}_Y^{1/2} \mathbf{Y}^T \mathbf{D} \mathbf{X} \mathbf{Q_X} \mathbf{X}^T \mathbf{D} \mathbf{Y} \mathbf{Q}_Y^{1/2} \mathbf{g_Y} = \lambda \mathbf{g_Y}$ ($q < p$), respectively, linked to the same maximum eigenvalue $\lambda = (\mathbf{w}_i^T \mathbf{Q_X} \mathbf{X}^T \mathbf{D} \mathbf{Y} \mathbf{Q_Y} \mathbf{c}_i)^2$ where $\sqrt{\lambda}$ is the covariance between $\mathbf{t}_i$ and $\mathbf{u}_i$. After diagonalization s principal axes are preserved. Finally, weight vectors $\mathbf{w}_i$ and $\mathbf{c}_i$ are given by $\mathbf{w}_i = \mathbf{Q_X}^{-1/2} \mathbf{g}_{\mathbf{X}_i}$ and $\mathbf{c}_i = \mathbf{Q_Y}^{-1/2} \mathbf{g}_{\mathbf{Y}_i}$, respectively.

Co-Inertia Analysis is then a symmetric coupling metric based method that provides a decomposition of the co-inertia criterion $tr(\mathbf{Q}_X^{1/2} \mathbf{X}^T \mathbf{D} \mathbf{Y} \mathbf{Q_Y} \mathbf{Y}^T \mathbf{D} \mathbf{X} \mathbf{Q}_X^{1/2}) = \sum_{s=1}^{p} \lambda_s$ on a set of orthogonal vectors where $tr()$ is the trace operator of a square matrix (sum of the elements on the main diagonal). It is easy to show that if we set $\mathbf{Q_X} = \mathbf{I}_p$, $\mathbf{Q_Y} = \mathbf{I}_q$ and $\mathbf{D} = \mathbf{I}_n$ then Tucker's approach, $\mathrm{COA}(\mathbf{X}, \mathbf{Y})_{\mathbf{I}_p, \mathbf{I}_q}$ and Undeflated PLS lead to the same results. First solutions of $\mathrm{COA}(\mathbf{X}, \mathbf{Y})_{\mathbf{I}_p, \mathbf{I}_q}$ and PLS Regression (Höskuldsson, 1988) are also equal.

For deeper COA features and its links with other multivariate coupling methods see Dolédec & Chessel (1994) and Dray et al. (2003), respectively.

COA-PLS is instead a "component-wise" method. "Component-wise" methods at first compute the "first order component and weighting vectors" and appropriate deflations of the original matrices are then developed before to compute another set of components. These deflations are performed at each step.

The COA and COA-PLS weight vectors solutions ($s = 1$) $\mathbf{w}_1$ and $\mathbf{c}_1$ are equal for both methods and differ for the remainders. To obtain the COA-PLS solutions of order $s > 1$ we start from the remark that the COA components scores result to be not $\mathbf{D}$-orthogonal in $\Re^n$. It is possible to overcome this remark by adding the orthogonality constraints for the COA components scores to the COA criteria (1). This leads to solve the following problem. Set $\mathbf{X}^{(0)} = \mathbf{X}$ and $\mathbf{Y}^{(0)} = \mathbf{Y}$ with $\mathbf{t}_1 = \mathbf{X}^{(0)}\mathbf{Q}_\mathbf{X}\mathbf{w}_1$ and $\mathbf{u}_1 = \mathbf{Y}^{(0)}\mathbf{Q}_\mathbf{Y}\mathbf{c}_1$. We look then for S pairs of components scores $\mathbf{t}_s = \mathbf{X}^{(s-1)}\mathbf{Q}_\mathbf{X}\mathbf{w}_s$ and $\mathbf{u}_s = \mathbf{Y}^{(s-1)}\mathbf{Q}_\mathbf{Y}\mathbf{c}_s$ ($s = 1, ..., S$) such that

$$\begin{cases} \max\limits_{\mathbf{w}_s\mathbf{c}_s} \operatorname{cov}^2(\mathbf{t}_s, \mathbf{u}_s)_\mathbf{D} \\ \|\mathbf{w}_s\|^2_{\mathbf{Q}_\mathbf{X}} = 1 \\ \|\mathbf{c}_s\|^2_{\mathbf{Q}_\mathbf{Y}} = 1 \\ \mathbf{t}_s^T\mathbf{D}\mathbf{t}_{s'} = 0 \\ \mathbf{u}_s^T\mathbf{D}\mathbf{u}_{s'} = 0 \end{cases} \tag{3}$$

with $s \neq s'$. Solutions of order $s = 1$ of the problem (3) are equal of the problem (1). COA-PLS solutions of order $s > 1$ of the problem (3) are then obtained according to the residuals of $\mathbf{D}$-projections:

Definition: *Generic step s* ($s > 1$) *of* COA-PLS *is defined by the* COA *solution applied to* $\mathbf{X}^{(s-1)}$, $\mathbf{Y}^{(s-1)}$ *and* $\mathbf{D}$, where $\mathbf{X}^{(s)} = \mathbf{P}^{\perp}_{\mathbf{T}^{(s-1)}/\mathbf{D}}\mathbf{X}^{(s-1)}$ and $\mathbf{Y}^{(s)} = \mathbf{P}^{\perp}_{\mathbf{T}^{(s-1)}/\mathbf{D}}\mathbf{Y}^{(s-1)}$ are then the residuals of the $\mathbf{D}$-projections of $\mathbf{X}^{(s-1)}$ and $\mathbf{Y}^{(s-1)}$ onto the subspace $\mathbf{T}^{(s-1)}$ spanned by $\{\mathbf{t}_1, ..., \mathbf{t}_{s-1}\}$. According to this definition, COA-PLS solutions of order s ($s > 1$) are then given by the eigenvector $\mathbf{w}_s = \mathbf{Q}_\mathbf{X}^{-1/2}\tilde{\mathbf{w}}_s$ where $\tilde{\mathbf{w}}_s$ is linked to the higher eigenvalue λ of the eigen-system

$$\mathbf{Q}_\mathbf{X}^{-\frac{1}{2}}\mathbf{X}^{(s-1)^T}\mathbf{D}\mathbf{Y}^{(s-1)}\mathbf{Y}^{(s-1)^T}\mathbf{D}\mathbf{X}^{(s-1)}\mathbf{Q}_\mathbf{X}^{-\frac{1}{2}}\tilde{\mathbf{w}}_s = \lambda\tilde{\mathbf{w}}_s \tag{4}$$

We remark that at each step s of the COA-PLS alghorithm, the squared $\mathbf{D}$-covariance $\operatorname{cov}^2(\mathbf{t}_s, \mathbf{u}_s)_\mathbf{D}$ is optimized. We highlight, moreover, that COA-PLS analysis shares the same properties and results of the Partial Least Square Regression (Höskuldsson, 1988). In fact, analogously to the PLSR, the properties of orthogonality of the S COA-PLS components imply the following additive decompositions of the matrices $\mathbf{X}$ and $\mathbf{Y}$

$$\mathbf{X} = \sum_{s=1}^{S}\hat{\mathbf{X}}^S + \mathbf{X}^{(S)} \text{ and } \mathbf{Y} = \sum_{s=1}^{S}\hat{\mathbf{Y}}^S + \mathbf{Y}^{(S)}$$

where $\hat{\mathbf{X}}^S = \mathbf{P}_{\mathbf{T}^{(s)}/\mathbf{D}}\mathbf{X}$ and $\hat{\mathbf{Y}}^S = \mathbf{P}_{\mathbf{T}^{(s)}/\mathbf{D}}\mathbf{Y}$ with $\mathbf{X}^{(S)}$ and $\mathbf{Y}^{(S)}$ error matrices not correlated with $\hat{\mathbf{X}}^S$ and $\hat{\mathbf{Y}}^S$, respectively. In addition, the explained variance of each matrix is splitted into additive parts.

In the same way, the components scores $\mathbf{t}_s = \mathbf{X}^{(s-1)}\mathbf{Q}_\mathbf{X}\mathbf{w}_s$ ($s = 1, ..., S$) (all gathered in the matrix $\mathbf{T}_s$) can be also written with respect to the original set of variables of $\mathbf{X}$ (Vivien, Sabatier, 2001), that is $\mathbf{t}_s = \mathbf{X}\mathbf{r}_s$, where the weight vectors $\mathbf{r}_s$ are given by

$$\begin{aligned} \mathbf{r}_1 &= \mathbf{Q}_\mathbf{X}\mathbf{w}_1 \\ \mathbf{r}_s &= \left(\mathbf{I}_p - \sum_{l=1}^{s-1}\frac{\mathbf{r}_l\mathbf{r}_l^T}{\|\mathbf{t}_l\|^2_\mathbf{D}}\mathbf{X}^T\mathbf{D}\mathbf{X}\right)\mathbf{Q}_\mathbf{X}\mathbf{w}_s \ (s \geq 2) \end{aligned}$$

where $\mathbf{w}_s$ is the COA-PLS weight vector computed with respect to the residual matrix $\mathbf{X}^{(s-1)}$. It is possible to show that the matrix $\mathbf{R}_s$ collecting the weight vectors $\mathbf{r}_s$ can be also directly computed by $\mathbf{R}_s = \mathbf{Q}_\mathbf{X}\mathbf{W}_s(\mathbf{P}_s^T\mathbf{Q}_\mathbf{X}\mathbf{W}_s)^{-1}$ where $\mathbf{W}_s$ is the matrix collecting the first s of the sequence of corresponding vectors $\mathbf{w}_s$ and $\mathbf{P}_s = \mathbf{X}^T\mathbf{D}\mathbf{T}_s(\mathbf{T}_s^T\mathbf{D}\mathbf{T}_s)^{-1}$ is the loadings matrix, such that $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$ and $\mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{F}$. According to the main characteristics of $\mathbf{Q}_\mathbf{X}$, matrix $\mathbf{P}_s^T\mathbf{Q}_\mathbf{X}\mathbf{W}_s$ is usually upper triangular and thus non singular and invertible.

After S dimensions have been extracted, the models $\hat{\mathbf{Y}}^S$ of rank S can be then written with respect to the original set of variables of $\mathbf{X}$: $\hat{\mathbf{Y}}^S = \mathbf{X}\hat{\mathbf{B}}^S_\mathbf{X}$ where $\hat{\mathbf{B}}^S_\mathbf{X} = \sum_{s=1}^{S}\frac{\mathbf{r}_s\mathbf{r}_s^T}{\|\mathbf{t}_s\|^2_\mathbf{D}}\mathbf{X}^T\mathbf{D}\mathbf{Y}$ is the matrix of COA-PLS regression coefficients. Finally, the matrix of fitted values $\hat{\mathbf{Y}}^S$ from COA-PLS is also simply the $\mathbf{D}$-projection of the observed responses $\mathbf{Y}$ onto the first S COA-PLS components $\mathbf{T}_s = \mathbf{X}\mathbf{R}_s$: $\hat{\mathbf{Y}}^S = \mathbf{X}\mathbf{P}_{\mathbf{T}_s/\mathbf{D}}\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}}^S_\mathbf{X}$, where the COA-PLS regression coefficients $\hat{\mathbf{B}}^S_\mathbf{X}$ can be alternatively rewritten as usual with respect to the ordinary least squares regression coefficients $\hat{\mathbf{B}}^{OLS}_\mathbf{X}$: $\hat{\mathbf{B}}^S_\mathbf{X} = [\mathbf{Q}_\mathbf{X}\mathbf{W}_s(\mathbf{P}_s^T\mathbf{Q}_\mathbf{X}\mathbf{W}_s)^{-1}\mathbf{P}_s^T]\hat{\mathbf{B}}^{OLS}_\mathbf{X}$. The knowledge of $\hat{\mathbf{b}}^S_\mathbf{X}$ provides an easy and fast way to compute the predictions errors and perform cross-validation to choose the number S of components for a good predictive model.

We highlight, in addition, that if we want to consider an "Orthogonalized" version of COA (we named it OCOA) then it is sufficient to change only the ways to obtain the residual in the previous definition of generic step s of COA-PLS, obtaining the following

Definition: *Generic step s ($s > 1$) of* OCOA *is defined by the* COA *solution applied to* $\mathbf{X}^{(s-1)}$, $\mathbf{Y}^{(s-1)}$ *and* $\mathbf{D}$, where $\mathbf{X}^{(s)} = \mathbf{P}^{\perp}_{\mathbf{T}^{(s-1)}/\mathbf{D}}\mathbf{X}^{(s-1)}$ and $\mathbf{Y}^{(s)} = \mathbf{P}^{\perp}_{\mathbf{U}^{(s-1)}/\mathbf{D}}\mathbf{Y}^{(s-1)}$ are then the residuals of the $\mathbf{D}$-projections of $\mathbf{X}^{(s-1)}$ and $\mathbf{Y}^{(s-1)}$ onto the subspaces $\mathbf{T}^{(s-1)}$ and $\mathbf{U}^{(s-1)}$ spanned by $\{\mathbf{t}_1, ..., \mathbf{t}_{s-1}\}$ and $\{\mathbf{u}_1, ..., \mathbf{u}_{s-1}\}$, respectively.

It is evident that COA-PLS can be also considered as a statistical framework because if we consider the different nature and coding of $\mathbf{X}$ and $\mathbf{Y}$, diverse choices of $\mathbf{Q_X}$ and $\mathbf{Q_Y}$, then a variety of existing coupling approaches are realized. As example, COA subsumes, among the others, CCA while PLS Regression and PLS-Discriminant Analysis are special cases of COA-PLS.

3. The scaling of ordinal measurement

As stated in section 2, COA deal with qualitative and quantitative variables observed on n statistical units. Generally, the qualitative variables are judgements about characteristics of services or products. In these circumstances the fundamental aspect is that measurement reflects only an order relation and it is then necessary to determine a criterion to convert on a metric scale ordinal measurements. In section 1 we underlined the importance of use an "indirect quantification", that consists in assigning real numbers to the categories of the qualitative variable. In this type of quantification the numbers are not equidistant but they depend on a latent variable. The well-known Thurstone scaling procedure is based on the hypothesis that the model is normally distributed.

According to this assumption the modalities x_i, ($i = 1, 2, \ldots, r$) of a qualitative variable X are associated to the values of a quantitative latent variable Z normally distributed. Let $F(i)$ be the cumulative relative frequency, corresponding to x_i and let $\phi^{-1}[F(i)]$ be the inverse of the cumulative distribution function, the quantile z_i associated to x_i can be expressed as:

$$z_i = \phi^{-1}[F(i)] \tag{5}$$

The procedure is repeated for all categorical observed variables and the expected values $E(Z_i)$ for each modality, calculated over all the X variables in the data-set, will be the new scores. Anyway when the categorical variables are not symmetric, the assumption that the latent variable is normally distributed can lead to strange results, almost when the asymmetry is very high (Portoso, 2003a). In many cases in fact, considering the association between the normal area and the empirical frequencies, if they are prevalently on the left side, it is easy to see that the quantile will be pushed towards right. If, instead the frequencies are on the right extreme the quantiles will slide towards the left side. In these situations, ordinal categorical variables with prevalently positive scores, will have new average scores and vice-versa. In this way the evaluations about products or services will be completely distorted. This incongruity leads to use a distribution that could better express, in a numerical way, the categorical variables characterized by this particular structure. For example, in this case, the (negative or positive) exponential distribution may be the right ones. To calculate the new scores it is necessary to consider the characteristics of these distributions. The negative exponential may be defined as follows:

$$\begin{cases} f(x) = exp^{-x} & \text{if} \quad 0 \leq x \leq +\infty \\ f(x) = 0 & \text{otherwise} \end{cases}$$

The mean and the variance of the distribution are:

$$E(X) = \int_0^{\infty} xf(x)d(x) = \int_0^{\infty} xexp^{-x}d(x) = 1 \tag{6}$$

$$Var(X) = \int_0^{\infty} x^2exp^{-x}d(x) = 1 \tag{7}$$

If we standardize the variable we obtain a new variable (S):

$$S = \frac{X-1}{1} \tag{8}$$

with

$$\begin{cases} f(s) = exp^{-s-1} \quad \text{if} \quad -1 \leq s \leq \infty \\ f(s) = 0 \quad \text{otherwise} \end{cases} \tag{9}$$

The cumulative distribution function is

$$\begin{cases} \Psi(s) = 1 - exp(-s-1) \quad \text{if} \quad -1 \leq s \leq \infty \\ \Psi(s) = 0 \quad \text{otherwise} \end{cases}$$

In the same way, if we consider the positive exponential distribution given by

$$\begin{cases} f(x) = exp^{x} \quad \text{if} \quad -\infty \leq x \leq 0 \\ f(x) = 0 \quad \text{otherwise} \end{cases}$$

the mean and the variance are defined as follows:

$$E(X) = \int_{-\infty}^{0} xf(x)d(x) = \int_{-\infty}^{0} xexp^{x}d(x) = -1 \tag{10}$$

$$Var(X) = \int_{-\infty}^{0} x^2 exp^{x} d(x) = 1 \tag{11}$$

and the standardized variable is

$$P = \frac{Y+1}{1} \tag{12}$$

In this case the cumulative distribution function of P is:

$$\begin{cases} \Psi(p) = exp(p-1) \quad \text{if} \quad -\infty \leq p \leq 1 \\ \Psi(p) = 0 \quad \text{otherwise} \end{cases}$$

To build the scores, both for the negative exponential distribution and for the positive one, it is necessary to consider the relative frequencies $f(i)$ and the cumulative relative ones $F(i)$. In this way the empirical distribution of cumulative frequencies is: $G(i) = F(i-1) + f(i)/2$ with $i = 1, 2, \ldots, r$ and comparing it with the two formulas of the cumulative theoretical distribution we can obtain the standardized quantiles:

$$s_i = -1 - ln[1 - G(i)] \tag{13}$$

$$p_i = 1 + ln[G(i)] \tag{14}$$

for the negative and positive exponential distribution respectively. The choice of the right distribution is a fundamental aspect in the quantification and an instrument to evaluate which latent variable could be assumed is the EN index (Portoso, 2003b). The EN index is an indicator that assumes values between -1 and +1. The value -1 is assumed when all the frequencies are associated to the first modality (in this case we have maximum negative concentration), while when there is maximum positive concentration the value assumed by the index will be +1. If the frequencies are balanced in a symmetric way then the EN index will be equal to 0. The index can be written as follows:

$$EN = \sum_{i}^{r/2} (f_{r-i+1} - f_i)(r - 2i + 1)/(r-1) \tag{15}$$

where r is the number of modalities and if they are odd the value $r/2$ is round off to the smaller integer while f_i are, as already stated, the relative frequencies associated to the modality i, f_{r-i+1} are the relative frequencies associated to the opposite modality and $(r-2i+1)$ is the difference between the position of the two opposite modalities. An alternative formulation of the index can be:

$$EN = 1 - 2\sum_{i}^{r-1} F(i)/(r-1) \tag{16}$$

that shows some similarities with the Gini index and where (F_i) have already been defined as cumulative relative frequency. If the value of the index EN is close to 0, the use of normal distribution does not generate any problems, but if the absolute value of this index grows then the use of exponential distribution can lead to better results. The problem is to define a threshold to decide which distribution is better to apply. Lucadamo and Portoso (2011) underline that if the value of the EN index is between −0.30 and 0.30 the normal distribution may be a good latent variable; if the index is between −0.90 and −0.50 the negative exponential distribution is the right one, while if the value is between 0.50 and 0.90 the positive exponential distribution can be used.

4. Applications to real datasets

During the last twenty years firms' strategy has gradually shifted from marketing to Total Quality Management to Customer Satisfaction. Particularly, for a company, the knowledge of the customer evaluation of a given product represents an important starting point for every business strategy. The need of marketing teams is, then, to define which mixes of technical parameters mostly influence the acceptability and the liking of a product. In this way it is possible to obtain directives to correct the defects in their products. The selection of the optimal mix of parameters may be not a simple question.

To deal with this problem, in this section we apply COA-PLS1 (for a single dependent variable) and COA-PLS2 (for several dependent variable) to two different datasets about sensory analysis. In both cases we deal with some ordinal categorical variables that have been quantified with Thurstone procedure and with a mixed procedure. Latter procedure considers negative/positive exponential or normal distribution as latent variable, according to the empirical distribution of the data. We apply the regression models to the new data and we analyze the effects of both quantification methods by comparing explicative and predictive indexes of models. We highlight that the evaluation of the full COA-PLS results for both examples is not the main goal of this paper.

4.1. COA-PLS1 example

The first set of data is relative to a survey on the preferences of consumers about different coffees. For this data we apply a COA-PLS1 because the aim is to study the relationship between a dependent variable (overall evaluation of the consumers) and the judgments about some characteristics of the coffees (sweetness, acidity, aftertaste, bitterness, smell, taste, consistency, aroma). The data are collected in a column vector with 23 observations and a matrix **X**, of order (36,8). In a first step of the analysis, all the judgments are quantified considering the psycometric approach and the COA-PLS1 is applied. Considering the EN index, introduced in section 3 we can see that only for two of the explicative variables it is possible to consider a normal distribution as latent variable, three variables seem to follow a positive exponential distribution and three a negative ones. For this reason, in a second step of the analysis the model is applied on the data quantified with the mixed procedure. To verify which quantification has an higher explicative ability we consider the R^2 indexes, while to verify the predictive abilities we can look at the Q^2 that is a cross validated R^2 index. In table 1 we can see the results we obtain.

Table 1. Thurstone vs Mixed Procedure: explicative and predictive power of the two COA-PLS1 models (2 components)

Index	Thurstone Procedure	Mixed Procedure
Q^2	0.556	0.651
R^2Y	0.589	0.656
R^2X	0.564	0.573

It is evident that the COA-PLS1 model, built on the data obtained with a mixed quantification, has better explicative and predictive abilities than the model based on Thurstone scores. In fact both R^2 and Q^2 indexes have higher values. It is interesting also to consider the ranking of the predictors on the basis of their importance in the prediction, according to the VIP (Variable Importance in the Prediction) (Wold et al. 1993).

The VIP index is an useful tool for variable selection: it provides a measure of the impact of all the independent variables. Moreover, since the average of squared VIP scores equals 1, "greater than one rule" is generally used as a

Table 2. Thurstone vs Mixed Procedure: explicative and predictive power of the two COA-PLS2 models (2 components)

Variable	VIP for Thurstone Procedure	VIP for Mixed Procedure
Consistency	1.509	1.447
Aroma	1.437	1.386
Taste	1.385	1.250
Sweetness	0.762	0.816
Smell	0.662	0.804
Aftertaste	0.560	0.616
Bitterness	0.455	0.564
Acidity	0.447	0.642

criterion for variable selection. In our analysis we note that variables with higher VIP in both analyses are Consistency, Aroma and Taste. Their values in Mixed procedure are lower than in Thurstone procedure, but anyway higher than 1. In the meantime we highlight that the new quantification leads an improvement of the values of all other variables and some of them drift to 1.

4.2. COA-PLS2 example

In the second data set we want to evaluate the influence of technique features on overall pasta liking by consumers. We analyze the statistical links between some chemical and physical descriptive variables and some hedonic variables for 36 different brands of pasta. These 36 brands were rated for technical descriptive variables (physical: color, temperature, humidity and ashes; chemical: gluten, protein and acidity) by a laboratory and these data are collected in a matrix, **X** of order (36, 7). The same brands were also rated by 11 trained consumers for the hedonic variables: pile, consistency and stickiness. In this case only the dependent variables are quantified according to the two procedures and average values over the judges is calculated. These hedonic data are so collected in a matrix **Y** of order (36,3). It is evident that pasta liking by consumers is affected by the technical structure and also for COA-PLS2 we compare the obtained results (table 3).

Table 3. Thurstone vs Mixed Procedure: explicative and predictive power of the two COA-PLS2 models (2 components)

Index	Thurstone Procedure	Mixed Procedure
Q^2	0.410	0.483
R^2Y	0.605	0.614
R^2X	0.760	0.779

Also for COA-PLS2, it is evident that mixed quantification shows better results for explicative and predictive power.

5. Conclusion

In several research or applied contexts, explorative multidimensional analysis are often applied also to ordinal variables. In these cases, ordinal scales are treated like being of interval type. This practice is justified and supported by the pragmatic approach to statistical measurement (Hand, 2009), for which when the researcher defines a construct then the measuring instrument is specified simultaneously. This aspect allows the researcher to select the kind of scale. Several approaches have been proposed to quantify ordinal measurements. The most used is the psycometric approach of Thurstone scaling procedure (Zanella, 1999). This approach assumes the presence of a continuous underlying variable for each ordinal indicator. According to Thurstone scaling procedure, we consider also the standardized exponential distribution, suggesting then a mixed quantification.

COA-PLS has been applied to two real data set, using the different quantification procedure. In both cases mixed procedure leads to better results in explicative and predictive sense. Obviously the method must be applied also on other kind of data and a simulation study too is necessary to verify the goodness of the quantification. Furthermore in section 3 we introduced the EN index and it is easy to see that there are some values for which neither the exponential nor the normal distribution may be appropriate. Anyway, according to the suggested approach, new theoretical distributions can be investigated (e.g. Beta or Logistic Weibull cumulative density functions) in order to better fit the empirical distributions of the variables and to obtain better results. All these remarks are actually under investigation.

References

Blackman J., Rutledge, D.N., Tesica, D., Saliba, A., Scollary, G.R., 2010. Examination of the potential for using chemical analysis as a surrogate for sensory analysis. Analytica chimica acta 660, 2–7.

Breiman, L., Friedman, J., 1985. Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc. 80, 580–598.

Bretherton, C.S., Smith, C.A., Wallace, J.M., 1992. An intercomparison of methods for finding coupled patterns in climate data. J. Clim. 5, 541–560.

Cazes, P., 1997. Adaptation de la régression PLS au cas de la régression aprés analyse des correspondances multiples, Revue de Statistique Appliquées XLV(2), 89–99.

Chessel, D., Mercier, P., 1993. Couplage de triplets statistiques et liaisons espces-environnement, in "*Biométrie et Environnement*". In: Lebreton, J.D., Asselain, B. (eds.) , 15–44.

Dolédec, S., Chessel, D., 1994. Co-Inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biol. 31, 277–294.

Dray, S., Chessel, D., Thioulouse, J., 2003. Co-Inertia analysis and the linking of ecological data tables, Ecology 84(11), 3078–3089.

Escoufier Y., 1987. The duality diagram: a means of better practical applications, in "*Development in numerical ecology*". In: Legendre P., Legendre L. (Eds.). NATO advanced Institute, Springer Verlag, Berlin.

Fisher, R.A., 1946. Statistical Methods for research workers. Oliver and Boynd, London.

Green, P.E., Tull, D.S., 1988. Research for Marketing Decisions. Prentice Hall, New York.

Guttman, L., 1941. The quantification of a class of attributes: a theory and method of scaling construction, in *The prediction of personal adjustment*". In: Horst, P. (ed.), Social Science Rresearch Council, New York.

Hand, D.,2009. Measurement theory and practice: the world through quantification. John Wiley, New York.

Hayashi, C., 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. Annals of the Institute of Statistical Mathematics 3(1), 69–98.

Hoskuldsson, A., 1988. PLS regression methods. Journal of chemometrics 2, 211–228.

Huon de Kermadec, F., Durand, J.F., Sabatier, R., 1996. Comparaison de méthodes de régression pour l'etude des liens entre données hédoniques, in Third Sensometrics Meeting, E.N.T.I.A.A., Nantes.

Huon de Kermadec, F., Durand, J.F., Sabatier, R. (1997). Comparison between linear and nonlinear PLS methods to explain overall liking from sensory characteristics. Food Quality and Preference 8, (5/6).

Kruskal, J.B., 1965. Analysis of Factorial Experiments by Estimating Monotone Transformations of the data. Journal of the Royal Statistical Society, B 27, 251–263.

Kvalheim, O.M., 1988. A partial least squares approach to interpretative analysis of multivariate analysis. Chemometrics and Intelligent Laboratory System, 3.

Likert, R., 1932. A technique for the measurement of attitudes. Archives of Psychology 140, 5–55.

Lucadamo, A., Portoso, G. 2011. Valori soglia dell'indice EN per la scelta della distribuzione normale o esponenziale nella quantificazione indiretta. Rivista Italiana di Economia, Demografia e Statistica LXV(1), 125–132.

MacFie, H.J.H, Thomson, D.M.H., 1988. Preference mapping and multidimensional scaling methods, in "*Sensory Analysis of Foods*" Elsevier Applied Science, London.

Nishisato, S., 2005. On the Scaling of Ordinal Measurement: A Dual-Scaling Perspective, in "*Contemporary Psychometrics*" In: Maydeu-Olivares, A., McArdle, J.J. (eds.). Lawrence Erlbaum Associates, Mahwah, pp. 479-507.

Portoso, G., 2003a. La quantificazione determinata indiretta nella customer satisfaction: un approccio basato sulluso alternativo della normale e dellesponenziale. Quaderni di dipartimento SEMeQ 53.

Portoso, G., 2003b. Un indicatore di addensamento codale di frequenze per variabili categoriche ordinali basate su giudizi, Quaderni di dipartimento SEMeQ, 66.

Prohaska, J., 1976. A technique for analyzing the linear relationships between two meteorological fields. Mon. Weather. Rev. 104, 1345–1353.

Russolillo, G., Lauro, C.N., 2011. A proposal for handling categorical predictors in PLS regression framework, in "*Classification and Multivariate analysis for complex data structures*". In: Fichet, B., Piccolo, D., Verde, R., Vichi, M. (Eds). Springer, pp 341-348.

Sabatier, R., Chessel, D., Maury L., 1992. Comment mesurer la concordance entre jugements subjectifs et observations multivariables pour des produits alimentaires. In: "Agro-Industrie et méthodes statistiques". Compte-rendu des 3èmes journées européennes. Montpellier 30/11-1/12 1992. Association pour la Statistique et ses Utilisations, Paris, pp.103-106.

Sampson, P.D., Streissguth, A.P., Barr, H.M., Bookstein, F.L., 1989. Neurobehavioral effects of prenatal alcohol: part II, Partial Least Squares analysis. Neurotoxicol. Teratol 11(5), 477–491.

Schlich, P., 1995. Preference mapping: relating consumer preferences to sensory or instrumental measurements, Bioflavour.

Tanaka, Y., 1979. Review of the methods of quantification. Environmental Health Perspectives 32, 113–123.

Thacker, W., 1999. Principal predictors. Int. J. Climatol. 19, 821–834.

Tucker, L.R., 1958. An inter-battery method of factor analysis. Psychometrika 23(2), 111–136.

Van De Geer, J.P.,1984. Linear relations among k sets of variables. Psychometrika 49, 79–94.

Vivien, M., Sabatier, R., 2001. Une extension Multi-tableaux de la régression PLS. Revue de Statistique Appliquée XLIV(1), 31–54.

Young, F.W., Jan de Leeuw, J., Takane, Y., 2976. Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. Psychometrika 41, 505–529.

Wallace, J.M., Smith, C.A., Bretherton, C.S., 1992. Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. J. Clim. 5, 561–576

Wold, S., Johansson, E., Cocchi, M., 1993. PLS: partial least squares projections to latent structures, in "*Drug design: Theory, Methods and Applications*". In: Kubinyi, H. (ed.). Escom science publishers, Leiden.

Wright, B.D., Linacrer, J.M., 1989. Observations are always ordinal: measures, however, must be interval. Archives of Physical Medicine and Rehabilitation 70, 857–860.

Wright, B.D., Masters, G.N., 1982. Rating Scale Analysis, Rasch Measurement. Mesa.

Zanella, A., 1999. Introduzione alla Misurazione della Customer Satisfaction. In: Valutazione della qualitá e Customer Satisfaction: il ruolo della statistica. - Aspetti oggettivi e soggettivi della Qualitá. pp. 217-231. Vita e Pensiero, Milano.