Stat. Meth. & Appl. (2007) 16:263–278
DOI 10.1007/s10260-006-0028-2

ORIGINAL ARTICLE

Two-step PLS regression for L-structured data: an application in the cosmetic industry

Vincenzo Esposito Vinzi · Christiane Guinot · Silvia Squillacciotti

Accepted: 25 September 2006 / Published online: 18 November 2006

Abstract The present paper proposes a PLS-based methodology for the study of so called "L" data-structures, where external information on both the rows and the columns of a dependent variable matrix is available. L-structures are frequently encountered in consumer preference analysis. In this domain it may be desirable to study the influence of both product *and* consumer descriptors on consumer preferences. The proposed methodology has been applied on data from the cosmetic industry. The preference scores from 142 consumers on 9 products were explained with respect to the products' physico-chemical and sensory descriptors, and the consumers' socio-demographic and behavioural characteristics.

Keywords Partial least squares (PLS) regression · Preference data · External information · L-structures

V. Esposito Vinzi · S. Squillacciotti
University of Naples "Federico II", Monte S. Angelo, 80126 Naples, Italy

V. Esposito Vinzi
ESSEC Business School, Avenue Bernard Hirsch,
B.P. 50105, 95021, Cergy-Pontoise, France
e-mail: vincenzo.espositovinzi@unina.it

S. Squillacciotti (✉)
EDF R&D – Département ICAME, 1, avenue du Général de Gaulle,
BP 408, 92141, Clamart, France
e-mail: silvia.squillacciotti@edf.fr

C. Guinot · S. Squillacciotti
CE.R.I.E.S., Biometrics and Epidemiology Unit, 20 rue Victor Noir,
92521 Neuilly-sur-Seine, France
e-mail: christiane.guinot@ceries-lab.com

1 Introduction

In "L-structures", a central matrix containing the dependent variables is explained by the interaction between a row-descriptor table and a column-descriptor table. This data structure is frequently encountered in the study of consumer preferences. In this framework, the central matrix (**Y**), containing the preference scores given by q consumers to n products, is explained by a row-descriptor matrix (**X**), containing p descriptors of the n products, and by a column-descriptor matrix (**Z**), containing the m descriptors of the q consumers. The name "L-structures" derives from the shape they assume when the matrices are placed one next to the other, as shown in Fig. 1.

The aim of the analysis of such a structure is to study how the interaction between column descriptors and row descriptors influences the dependent variables. Problems in modeling the interaction effect of the two descriptor matrices on **Y** are mainly due to the different dimensions of the three tables: as it can be observed in Fig. 1, matrices **X** and **Z** share no dimension, while each of them shares one common dimension with **Y**.

Consumer studies in market analysis represent a typical application field for L-shaped data tables: often, preferences concerning a certain number of products have been assessed by consumers and arranged in a two-way table **Y**. External information on both products (physical or chemical description or sensory characteristics) and consumers (demographic information, purchase behaviour) may be available respectively in row descriptor matrix **X** and in column descriptor matrix **Z**. Apart from problems related to the particular data structure, statistical studies in marketing must often take into account other specific problems, such as multicollinearity (typical in consumers' preference assessment, and in data expressing physical or chemical composition), missing values and noisy data. A model able to individualise the main underlying structure, and where results can be validated by cross validation or re-sampling techniques, is therefore required.

After a description of some of the techniques existing in literature for L-structure models, a different method is proposed here, based on a double PLS regression. The aim of this method is to reveal the structural relationships among the three matrices, i.e. the description of elements in **Y**, taking into account both the row and the column descriptors. The obtained model will

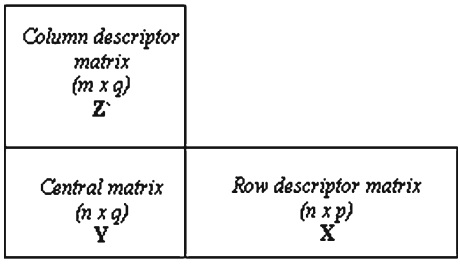

Fig. 1 Representation of a L-structure

therefore allow prediction of **Y** values for individuals whose values in **X** are known, given the corresponding **Z** values, or how new statistical units in **Z** will score on **Y** given the corresponding values in **X**. Furthermore, estimated relationships will allow the definition of groups among individuals, homogeneous with regard to the variables in **Y**, **Z** and **X**, which is the secondary aim of this study. An application of the said method to data regarding consumer preferences on cosmetic products is shown.

In the present work, matrices will be referred to in bold capital letters (e.g. **Y**, **X**), and vectors in bold lower case (e.g. **x, y**). Scalars will be indicated in italics (e.g. k, n).

2 State of the art

In recent years, many models have been proposed in literature to help in the estimation of relationships among matrices in a L-shaped structure: most of these methods are based on simultaneous or alternated singular value decompositions (SVD's). The first part of the present sub-chapter will be devoted to Principal Component Analysis with external information on both subjects and variables (Takane and Shibayama 1991) and successive extensions of this methodology, while the second part will devoted to PLS Regression for L-structured data (L-PLSR) proposed by Martens (Martens et al. 2005).

2.1 Principal component analysis with external information on both subjects and variables

Principal Component Analysis (PCA) with external information on both subjects and variables (Takane and Shibayama 1991) combines features of regression analysis and PCA in a unique two-step method. In the first step (external analysis), external information from row and column descriptors is taken into account: variability of **Y** is decomposed according to a linear regression model in four sources of variability (variability due to both **X** and **Z**, variability due to **Z**, variability due to **X** and variability due to error):

$$\mathbf{Y} = \mathbf{XMZ}' + \mathbf{BZ}' + \mathbf{XC} + \mathbf{E}. \tag{1}$$

Although the complete model in (1) also includes the single effect of **X** and **Z**, the interest here is especially in the estimation of the interaction effect, hence on the first term in equation (1). Coefficient matrices **M**, **B**, and **C** are subsequently estimated (either simultaneously or sequentially), following a sum of squares minimisation criterion. Estimate of **M** will be therefore matrix $\hat{\mathbf{M}}$ minimising $tr(\mathbf{E}'\mathbf{E})$, where

$$\hat{\mathbf{M}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{YZ}(\mathbf{Z}'\mathbf{Z})^{-}. \tag{2}$$

By replacing equation (2) into (1) it can be observed that the effect of the interaction of the two descriptor tables on $\mathbf{Y}$ is obtained by orthogonal projection of $\mathbf{Y}$ onto the subspaces spanned by the columns of $\mathbf{X}$ and $\mathbf{Z}$, as shown in (3):

$$\hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}\mathbf{P}_{\mathbf{Z}}. \tag{3}$$

Thus the non-uniqueness of the solution in (2) due to the generalized inverse is solved by using the orthogonal projectors.

The second step of the method (internal analysis) consists of a PCA on each of the isolated components of $\mathbf{Y}$ variability: particularly interesting may be the PCA performed on the interaction effect in (3), or on $\hat{\mathbf{M}}$ itself.

Successive extensions of such methodology have been proposed by Amenta and D'Ambra (1997) in the framework of Principal Component Analysis onto a Reference subspace (PCAR) (see Lauro and D'Ambra 1992 for a review), and by Giordano and Scepi (1999) in the framework of Conjoint Analysis.

2.2 PLS regression for L-structured data (L-PLSR)

In the framework of consumer preference studies, L-structures combine sensory analysis (in the relationship between consumer preferences and product descriptors) and consumer studies (in the relationship between preferences and consumer characteristics). Both domains show some typical problematic issues. Sensory analysis studies are generally characterised by a low number of statistical units compared to the number of variables. Moreover, both the product descriptors and the preferences are often strongly correlated. On the other hand, data concerning consumers contains in many cases a high degree of structural noise. PLS seems therefore well adapted to these particular data structures. PLS methods allow the estimation of the model also in case of multicollinearity, missing data, low number of statistical units with respect to the variables, and noisy data.

Under this point of view PLS Regression for L-structured data (L-PLSR) has been developed (Martens et al. 2005).

The PLS multivariate regression (PLS2) can be seen either under an algorithmic point of view (NIPALS) or it can be connected to classical multivariate theory according to Wold (Wold et al. 1983) and Stone and Brooks (Höskuldsson 1988, Stone and Brooks 1990). These works show how parameter estimation in PLS2, and consequentially the extraction of the A relevant components, can be solved through classical eigen-problems, by a single SVD of the cross product matrix $(\mathbf{X}'\mathbf{Y})$ or by repeating A times the SVD of the cross-product of the deflated matrix $\left(\mathbf{X}'_{a-1}\mathbf{Y}\right)$, with $a = 1, \ldots, A$. The deflated matrix $\mathbf{X}_{a-1}$ is the residual after reconstruction of matrix $\mathbf{X}$ from the components obtained in the previous step ($\mathbf{X}_a = \mathbf{X}_{a-1} - \mathbf{t}_a\mathbf{p}'_a$). This solution leads to the same results as the traditional algorithmic approach, when SVD's are feasible (no missing data, no landscape tables and no multicollinearity).

In L-structures the dependence of **Y** must be studied with respect to the row descriptors **X** and to the column descriptors **Z**. The model proposed by Martens is given in Eq. (4).

$$\mathbf{Y} = \mathbf{T}_X^{(M)}\mathbf{D}_A^{(M)}\mathbf{T}_Z'^{(M)} + \mathbf{G}_A^{(M)}, \tag{4}$$

where the superscript (M) indicates that components, loadings, weights and coefficients are obtained according to Martens' method.

$\mathbf{T}_X^{(M)}$ and $\mathbf{T}_Z^{(M)}$ are, respectively, the matrices containing the components of **X**, defined as the linear combinations of its columns, and of **Z**. $\mathbf{G}_A^{(M)}$ is the error term matrix. Equation (4) focuses on the analysis of the interaction effect but it can be extended so as to take into account the single effects of **X** and **Z** on **Y** by adding two new additive terms $\mathbf{T}_X^{(M)}\mathbf{Q}_X^{(M)}$ and $\mathbf{Q}_Z^{(M)}\mathbf{T}_Z'^{(M)}$. The information concerning the interaction effect is contained in the parameter matrix $\mathbf{D}_A^{(M)}$ (where A is chosen by means of cross-validation procedures). Following the eigen-solution proposed for PLS2, in L-PLSR parameter estimation is implemented either through simultaneous extraction of components, by means of a unique SVD of matrix $(X'YZ)$, or through sequential extraction of components by means of sequential SVD's on A deflated matrices $\left(\mathbf{X}'_{a-1}\mathbf{Y}\mathbf{Z}_{a-1}\right)$. In both cases matrices **X** and **Z** are supposed to be centred, while matrix **Y** is supposed to be double-centred with respect to both its rows and its columns. The matrices of the loading vectors for **X** and **Z**, respectively $\mathbf{W}_X^{(M)}(A_X xp)$ and $\mathbf{W}_Z^{(M)}(A_Z xm)$, are thus estimated as the left and right singular vectors corresponding to the largest singular value. The estimated value for $\mathbf{D}_\mathrm{A}$ is computed as:

$$\left(\mathbf{T}_X'^{(M)}\mathbf{T}_X^{(M)}\right)^{-1}\mathbf{T}_X'^{(M)}\mathbf{Y}\mathbf{T}_Z^{(M)}\left(\mathbf{T}_Z'^{(M)}\mathbf{T}_Z^{(M)}\right)^{-1}. \tag{5}$$

The effect of the interaction between **X** and **Z** on **Y** is evaluated through a reduced number of components of each descriptor matrix, estimated according to PLS principles. Through the computation of data-model correlations, a graphical representation of all the elements in the analysis (row and column descriptors, dependant variables and statistical units) can then be obtained, as it will be shown in the application.

3 Two-step L-PLSR

The use of an "eigen-problem" approach in PLS2 Regression is feasible only if the product matrix SVD is actually possible (no missing data, no landscape tables, low multicollinearity). In any other case, the original iterative NIPALS procedure must be adopted. SVD must be excluded in particular when one or both the matrices contain missing data. Hence, L-PLSR, being based essentially on matrix decompositions, encounters the same problem, and in case of missing

data, model parameters can only be estimated after imputation or deletion of the missing values.

Furthermore, in PLS2 regression, the decomposition of the product matrix $(\mathbf{X}'\mathbf{Y})$ can be interpreted and explained in relation to PLS optimisation criterion, i.e. the squared covariance among the components of each table. The extension to L-structures by adding the third term $\mathbf{Z}$ to the cross product matrix to be decomposed, however, does not find the same statistical justifications.

Hence, L-PLSR first of all is based on an optimisation criterion which is of obscure statistical interpretability, and, secondly, does not allow the model estimation when matrices contain missing data, although this is one of the major advantages of PLS methods. Finally, the model predictivity is often quite low.

The methodology proposed in the present paper aims at defining a technique for dealing with L-structured data in the framework of PLS. The defined technique shall lead to a statistically interpretable optimisation criterion, comparable with the squared covariance in traditional PLS Regression. Also, it will be possible to implement the defined technique in all situations where PLS is more appropriate than OLS regression: multi-collinearity, landscape tables and missing values.

Two-step PLS Regression for L-structured data (two-step L-PLSR) has been developed in order to define a model for tables in a L-structure when data are affected by multicollinearity and missing values and tables contain more variables than individuals.

Two-step L-PLSR is based on a double PLS Regression (Tenenhaus 1998): in step 1 the variables in $\mathbf{Y}$ are regressed on row descriptor columns in $\mathbf{X}$. The relationship between the two tables is expressed in the following equation:

$$\mathbf{Y} = \mathbf{T_X}\mathbf{C}'_{\mathbf{X}} + \mathbf{Y_E}, \tag{6}$$

where $\mathbf{T_X}$ is the $(n\,x\,A_X)$ matrix containing the components related to matrix $\mathbf{X}(\mathbf{T_X} = \mathbf{X}\mathbf{W}^*_{\mathbf{X}}, \mathbf{W}^*_{\mathbf{X}}$ being the $p\,x\,A_X$ matrix containing the loadings allowing to build the components in $\mathbf{T_X}$ from the p originary variables in $\mathbf{X}$) and $\mathbf{Y_E}$ is a residual-term matrix of dimensions $(n\,x\,q)$. Matrix $\mathbf{C_X}(q\,x\,A_X)$ contains the regression coefficients relating $\mathbf{Y}$ to $\mathbf{T_X}$. Such coefficients are the expression of the relationship existing between $\mathbf{X}$ and $\mathbf{Y}$ through components $\mathbf{T_X}$, which are built with the aim of maximising both the explained variability in $\mathbf{X}$ and in $\mathbf{Y}$.

In step 2, the coefficient-loading matrix $\mathbf{C_X}$ is regressed on the column descriptor matrix $\mathbf{Z}$. The following equation relates the two tables:

$$\mathbf{C_X} = \mathbf{T_Z}\mathbf{C}'_{\mathbf{Z}} + \mathbf{Y_{EZ}}, \tag{7}$$

where columns in $\mathbf{T_Z}$ are the components related to matrix $\mathbf{Z}$ $(\mathbf{T_Z} = \mathbf{Z}\mathbf{W}^*_{\mathbf{Z}})$, $\mathbf{Y_{EZ}}$ is the second step residual matrix of dimensions $(q\,x\,A_X)$ and $\mathbf{C_Z}$ is the coefficient loading matrix linking $\mathbf{C_X}$ to $\mathbf{Z}$. Equation (7) can thus easily be reformulated by expressing $\mathbf{C_X}$ as a function of $\mathbf{Z}$:

$$\mathbf{C_X} = \mathbf{Z}\mathbf{W}^*_{\mathbf{Z}}\mathbf{C}'_{\mathbf{Z}} + \mathbf{Y_{EZ}}.$$

The choice of $\mathbf{C_X}$ as the regressand in step 2 allows first of all to explain how variables in $\mathbf{Z}$ can influence the relationship between $\mathbf{X}$ and $\mathbf{Y}$, since $\mathbf{C_X}$, whose columns are the coefficients linking $\mathbf{T_X}$ to $\mathbf{Y}$, contains the relevant information on the relationship between $\mathbf{X}$ and $\mathbf{Y}$. Moreover, this choice allows to overcome the problem of $\mathbf{X}$ and $\mathbf{Z}$ having no common dimension.

The final global model is expressed in the following equation, which relates consumer preferences in $\mathbf{Y}$ to both $\mathbf{Z}$ and $\mathbf{X}$-components:

$$\mathbf{Y} = \mathbf{T_X C_Z T'_Z} + \mathbf{Y}_{\text{res}}. \tag{8}$$

The structural term ($\mathbf{T_Z C_Z T'_Z}$) in equation (8) contains the information on $\mathbf{Y}$ variability explained by $\mathbf{X}$-$\mathbf{Z}$ interaction, while $\mathbf{Y}_{\text{res}} = \mathbf{T_X Y'_{EZ}} + \mathbf{Y_E}$ is a residual term. Equations (4) and (8) are very similar to one another, though obtained by means of very different procedures: a SVD of the product matrix $\mathbf{X'YZ}$ in the former and, in the latter, a two-step regression which allows model parameters estimation despite missing data and landscape tables.

The interaction parameter matrix $\mathbf{C}_Z$ can be expressed as in equation (9):

$$\mathbf{C}_{\text{Z}} = \left(\mathbf{T}'_{\text{X}}\mathbf{T}_{\text{X}}\right)^{-1}\mathbf{T}'_{\text{X}}\mathbf{Y}\mathbf{T}_{\text{Z}}\left(\mathbf{T}'_{\text{Z}}\mathbf{T}_{\text{Z}}\right)^{-1}. \tag{9}$$

Hence, according to equations (9) and (5) in both L-PLSR and two-step L-PLSR the estimated value of $\mathbf{Y}$, $\hat{\mathbf{Y}}$, is obtained by means of a double orthogonal projection onto the subspaces spanned by the components of the two descriptor tables. Put aside the differences in the component scores, the way such projections are performed differ according to the chosen method. In L-PLSR the orthogonal projection is performed simultaneously, but the columns of $\mathbf{Y}$ are projected onto the subspace spanned by the columns of $\mathbf{T}_X$ while the rows are projected onto the subspace spanned by the columns of $\mathbf{T}_Z$. In two-step L-PLSR, instead, the columns of $\mathbf{Y}$ are first of all projected onto the subspace spanned by $\mathbf{T}_X$; this projection, $\hat{\mathbf{Y}}_X$, is then projected, in step 2, onto the subspace spanned by $\mathbf{T}_Z$, thus leading to $\hat{\mathbf{Y}}$. The double projection resulting from the two-step procedure is reflected in the two residual terms in $\mathbf{Y}_{\text{res}}$ in (8). The dimensionality reduction resulting from the two orthogonal projections determines a loss of information related to the quality (in terms of predictivity) of the two steps. A comparison among L-PLSR and two-step L-PLSR in terms of global predictivity will be performed in the application.

Following Martens (Martens et al. 2005), two-step LPLSR allows single graphical representations of the different elements in the analysis (statistical units, preference notes, row descriptors and column descriptors) by means of data-model correlations. Product descriptors ($\mathbf{X}$ columns) and consumer preferences ($\mathbf{Y}$columns) are represented through their correlations to $\mathbf{T}_X^{(M)}$ columns, while consumer descriptors ($\mathbf{Z}$ columns) and products ($\mathbf{Y}$ rows) are represented through their correlations to $\mathbf{T}_Z^{(M)}$. Four single representations or two double representations are possible, while a global graphical representation on the same subspace can be obtained by means of superimposition of the single

representations. Martens justifies the simultaneous representation by means of superimposition on the same plane by the scale identity of the co-ordinates. However the possibility of interpreting the relative positions of the points (in terms of distance or proximity) requires a more accurate investigation in order to obtain a statistical justification for the superimposition. It is important to provide a justification showing that a positive score on $\mathbf{t_X}$ corresponds to a high score on $\mathbf{t_Z}$.

According to Eq. (8), and supposing for simplicity sake that only two components have been retained, the predicted values for y_{ij} in two-step L-PLS Regression are obtained as follows:

$$y_{ji} = \left[\mathbf{t_{x1}}_{(i)}\ \mathbf{t_{x2}}_{(i)}\right]\begin{bmatrix} c_{z11} & c_{z21} \\ c_{z12} & c_{z22} \end{bmatrix}\begin{bmatrix} \mathbf{t_{z1}}_{(j)} \\ \mathbf{t_{z2}}_{(j)} \end{bmatrix}$$

where y_{ij} is the note given by the generic consumer j ($j = 1, \ldots, q$) to the generic product i ($i = 1, \ldots, n$).

Given to PLS properties, the dependent variables in step 2 ($\mathbf{C_X}$) are non correlated, hence $\mathbf{C_Z}$ results in an almost diagonal matrix with off-diagonal elements very close to 0, and the final equation for y_{ji}, taking into account the sole structural model, is therefore, according to Eq. (8), the following:

$$y_{ji} = c_{z11}\mathbf{t_{x1}}_{(i)}\mathbf{t_{z1}}_{(j)} + c_{z22}\mathbf{t_{x2}}_{(i)}\mathbf{t_{z2}}_{(j)} \tag{10}$$

According to Eq. (10), let us suppose that variables $\mathbf{x}_s$ and $\mathbf{z}_r$, representing respectively the generic column of $\mathbf{X}$($s = 1, \ldots, p$) and the generic column of $\mathbf{Z}$($r = 1, \ldots, m$), are positively and strongly correlated to, respectively, $\mathbf{t_x}$ and $\mathbf{t_z}$. Supposing c_{z11} and c_{z22} positive for simplicity sake, Eq. (10) shows that if a certain consumer, characterised by $\mathbf{z_r}$ is positioned on the plane close to product characteristic $\mathbf{x}_s$, the value of y_{ji} for that consumer will be high for products characterised by $\mathbf{x}_s$. An empirical demonstration will be given in the following. This enables us to graphically relate product descriptors and preferences to consumer descriptors and products, and subsequently to superimpose the representations and to interpret the relative positions of points coming from spaces spanned by components of different nature.

4 Application in the cosmetic industry

Two-step L-PLSR has been used in order to describe the global relationships among consumer preferences on cosmetic products, product physico-chemical and sensory descriptors and consumer socio-demographic characteristics and behavioural descriptors.

4.1 Material

Data in the present work come from two parallel studies: a sensory description of nine hydrating cosmetic products (in four different textures: milk, cream, gel and gel-cream) and a hedonic evaluation given by 142 female consumers, regularly using this type of products. On the same women, a number of behavioural and socio-demographic descriptors was collected. The resulting tables are therefore the following:

- Table **Y** (9×142) containing the preference scores, on a 0–20 scale, from the 142 women on the nine hydrating products;
- Table **X** (9×29) containing 15 physico-chemical descriptors [$\mathbf{pc_1} \ldots \mathbf{pc_{15}}$], coded as dummy variables, and the sensory evaluation on 14 relevant product characteristics given by a panel of expert judges. This table gives a complete description of the products, according to the characteristics, considered as relevant by the final user, of the products' texture, tactile aspects [$\mathbf{t_1}, \ldots, \mathbf{t_4}$], immediate application effects [$\mathbf{a_1}, \ldots, \mathbf{a_4}$] and delayed application effects [$\mathbf{d_1}, \ldots, \mathbf{d_6}$].
- Table **Z** (142×148) containing the consumer descriptors. Some of the descriptors (such as age, number of daily product applications) are on a continuous scale, but most of them have been coded as dummy variables. Such descriptors include socio-demographic variables (age, income, etc.), purchase behaviour and cosmetic habit descriptors [$\mathbf{hab_1}, \ldots, \mathbf{hab_{108}}$] and skin characteristics (such as sensitivity and skin age) [$\mathbf{skin_1}, \ldots, \mathbf{skin_{24}}$]. Socio-demographic variables have kept their original notation [$\mathbf{age}, \mathbf{reven_1} \ldots \mathbf{reven_6}, \mathbf{socioprof_1} \ldots \mathbf{socioprof_4}$].

The nine products are not the result of a sampling procedure over a larger population, but represent themselves the entire population. The aim of this analysis is hence the description of the joint effect of product descriptors and consumer characteristics in the consumers' preferences for the nine products.

The preliminary descriptive analyses performed on the data prove a strong multicollinearity in tables **X** and **Y**. As to table **Z**, Principal Component Analysis (PCA) results show a high degree of noise: 51 dimensions are needed according to eigenvalue 1 rule (40 are needed in order to explain 75% of total variability). Furthermore, table **Z** contains 160 missing values: in the present work, we have aimed at taking into account the effects of **Z** and **X** on **Y** without previously estimating the empty cases values.

Before performing the two-step L-PLS R, separate PLS regressions have been performed, of **Y**on **X** and of **Y** on $\mathbf{Z}'$. The regression of consumer preference scores over product descriptors is common in sensory analysis, and under an applicative point of view it allows to detect which products or product characteristics are most strongly appreciated by consumers, and, if possible, to define a new product as a combination of the most preferred descriptors. The regression of consumer preferences over consumer descriptors helps in understanding which typologies of consumers (described by age, profession, etc.) prefer which products.

As explained in the introduction, if both product and consumer descriptors are available, an analysis taking into account the three tables will lead to a richer information than the two separate analyses. Especially when searching for classes of consumers, the separate regressions allow to define classes leading to a limited information. In the first case (consumer preferences over product characteristics), the researcher will not know who the consumer preferring certain product characteristics are, which is a relevant information for the definition of effective marketing and communication strategies. In the second case (consumer preferences over consumer characteristics) the researcher will not know which product characteristics are preferred in the different consumer classes, which is a relevant information in the definition of a product strategy.

Consumer data contained in table $\mathbf{Z}$ is often noisy: this is the case in our data, as shown by the results of the PCA on $\mathbf{Z}$. PLS Regression of $\mathbf{Y}$ over $\mathbf{Z}'$ leads to only two significant components, explaining 19% of $\mathbf{Y}$ variability. Technically, two-step PLS R may start from the regression of $\mathbf{Y}$ on $\mathbf{X}$ as well as from the regression of $\mathbf{Y}$ on $\mathbf{Z}'$. Given however the low predictive power of $\mathbf{Z}'$ on $\mathbf{Y}$, we have chosen to begin by the regression of consumer preferences over product descriptors. The results of this first step are detailed in the following paragraph.

4.2 Results

A first PLS regression of consumer preferences on product characteristics is carried out. The analysis has been performed on SIMCA 10.0 (Umetrics 2002). In order to increase $R^2(\mathbf{Y})$ on the first components and consequently reduce the information dispersion on the last components, a selection of $\mathbf{Y}$ variables was taken: all $\mathbf{y}_j$ variables showing a weak R^2 sum (<0.50) on the first three components have been excluded from the model. This is equivalent to excluding consumers whose position on the $\mathbf{w}_a^*\mathbf{c}_a$ plane ($a = (1, 2, 3)$) is very close to the origin of axes, and whose behaviour, not being predictable on the basis of the explanatory variables, is therefore scarcely interesting for the aims of the study. According to this criterion, 50 consumers ($\mathbf{Y}$ variables) have been excluded, and a new model has been estimated. The model has then been re-estimated on the remaining 92 women, leading, on the basis of cross-validated $R^2(Q^2)$, to the choice of four components. Overall results for the latter model, which has been retained as step 1 model, are shown in Table 1.

In step 2, coefficient loadings from step 1 ($\mathbf{C_X}$) are regressed on consumer descriptors contained in $\mathbf{Z}$. In order to minimise the loss of information in passing from step 1 to step 2, six coefficient loadings were retained as dependent variables in step 2, even though step 1 Q^2 suggested only four components. A first model was then estimated on all the variables in $\mathbf{Z}$, leading to no significant component. The result was probably due to the high degree of noise contained in the explanatory variables: two components were nevertheless computed in order to perform a selection of predictors. Variable selection was carried out according to Variable Importance in Projection (VIP) and to the predictors' importance in strategy for the final user. Variables showing a VIP value <0.8,

Table 1 Two-step L-PLSR step 1 overall results

Comp. no.	$R^2(\mathbf{X})$	$R^2(\mathbf{X})$ cum	$R^2(\mathbf{Y})$	$R^2(\mathbf{Y})$ cum	Q^2	Q^2 cum
1	0.18	0.18	0.42	0.42	0.01	0.01
2	0.25	0.43	0.18	0.60	0.05	0.06
3	0.24	0.68	0.11	0.71	0.07	0.13
4	0.13	0.81	0.06	0.77	0.01	0.14
5	0.09	0.90	0.08	0.85	0.05	0.18
6	0.06	0.96	0.05	0.89	−0.02	0.17
7	0.03	0.99	0.07	0.96	0.47	0.56
8	0.011	1	0.040	1	1	1

Table 2 Two-step L-PLSR step 2 overall results

Comp.	$R^2(\mathbf{Z})$	$R^2(\mathbf{Z})$ cum	$R^2(\mathbf{Cx})$	$R^2(\mathbf{Cx})$ cum	Q^2	Q^2 cum
1	0.0572	0.0572	0.1168	0.1168	0.0357	0.0357
2	0.0543	0.1115	0.0768	0.1935	0.0098	0.0452

if not considered as relevant for the analysis results, were removed. Only 73 of the 148 independent variables were therefore retained in the analysis. Model overall results are given in Table 2.

The global graphical representation (Fig. 2) is obtained by means of data-model correlations and superposition of the four graphical representations of

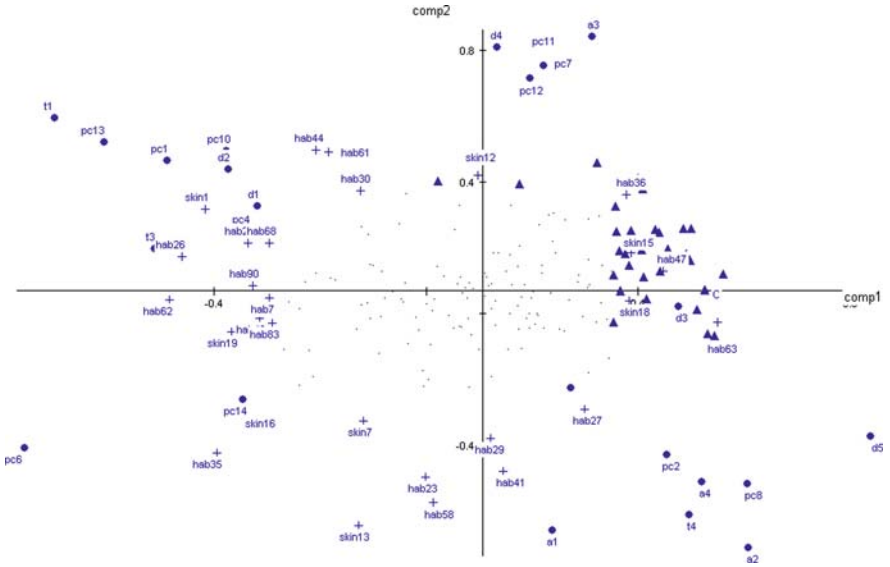

Fig. 2 Global graphical representation of two-step L-PLSR results. Consumer preferences (*filled triangles*) are related to product descriptors (*filled circles*) and consumer descriptors (crosses); products are represented as diamonds (*open diamonds*)

the elements in the analysis (products, product descriptors, consumer preferences and consumer descriptors). In order to improve the figure interpretability, elements close to the centre of the axes have not been labelled. Individuals (represented by triangles) are mostly disposed on the upper right section of the plane. Column descriptors in $\mathbf{Z}$ (represented by crosses) show rather low correlations to the model components, especially when compared to physico-chemical and sensory descriptors (dots), mainly distributed at the extremities of the plane. This result, also found in (Martens et al. 2005), is probably due to the high degree of structural noise contained in the consumer data. Finally, products are represented by diamonds. Figure 2 enables us to visualise which product characteristics have been appreciated by the consumers, and which consumer characteristics are most important in explaining the relationship among product characteristics and consumer preferences. Product descriptors $\mathbf{a_3}, \mathbf{d_3}, \mathbf{d_4}, \mathbf{pc_{11}}$, and $\mathbf{pc_{12}}$ seem to have been appreciated by most consumers, while consumer descriptors showing a strong correlation with preferences and product descriptors are skin characteristics **skin5, skin18, skin12** and **skin2** and habits **hab36 , hab47, hab63** and **hab80**.

The figure also allows a numerical justification to the superposition of the four graphical representations. According to Fig. 2, variables showing the strongest positive correlations with the first axis are product descriptor **d3** and consumer descriptors **hab63, hab47, skin18**. Consumer C (the only labelled consumer in Fig. 2, on the right side) is positioned at a short distance from the four variables: the graphical interpretation of the relationship among the variables is confirmed by the observation of the raw data. Consumer C shows the characteristics **hab63 hab47** and **skin18,** and has given a high score to product 6 (15). Product 6 has obtained the highest evaluation from the sensory panel with respect to descriptor **d3 (7.42)**.

In order to perform a comparison among models in terms of predictivity, L-PLSR has been implemented on the same data. Missing values in $\mathbf{Z}$ were replaced by the NIPALS (*Non Linear Iterative PArtial Least* Squares) estimates obtained through a PLS2 regression of $\mathbf{Y}$ on $\mathbf{Z}'$. Explained variability in each model has been computed according to the general formula:

$$\text{Expl Var} = \frac{\text{tr}\left(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}\right)}{\text{tr}(\mathbf{Y}'\mathbf{Y})}$$

The comparison between L-PLSR (simultaneous and sequential component extraction) and two-step L-PLSR is given in Table 3.

Two-step L-PLSR proposed in this paper clearly outperforms L-PLSR in terms of predictive capability.

Once the global model estimated, a secondary aim of the analysis was to define clusters of consumers showing homogeneous preference models and background characteristics. The need for the research of consumer clusters was first of all based on the assumption that the "ideal" product does not exist since consumers have different tastes. The knowledge of these different tastes *and*

Table 3 Comparison between L-PLSR and two-step L-PLSR in terms of model predictivity (explained variance values)

Martens L-PLSR (simultaneous SVD)	Martens L-PLSR (sequential SVD's)	Two-step L-PLSR (proposed method)
0.21	0.13	0.53

Table 4 Results from step 1 on class 2

Comp. no.	$R^2(\mathbf{X})$	$R^2(\mathbf{X})$ cum	$R^2(\mathbf{Y})$	$R^2(\mathbf{Y})$ cum	Q^2	Q^2 cum
1	0.16	0.16	0.51	0.51	0.05	0.05
2	0.28	0.44	0.14	0.65	0.03	0.08
3	0.23	0.67	0.10	0.75	0.05	0.13
4	0.13	0.80	0.06	0.81	−0.02	0.11
5	0.11	0.90	0.06	0.88	0.11	0.21
6	0.05	0.95	0.05	0.92	−0.06	0.16
7	0.04	0.99	0.04	0.97	0.49	0.57
8	0.01	1	0.03	1	1	1

Table 5 Results from step 2 on class 2

Comp. no.	$R^2(\mathbf{X})$	$R^2(\mathbf{X})$ cum	$R^2(\mathbf{Y})$	$R^2(\mathbf{Y})$ cum	Q^2	Q^2 cum
1	0.0706	0.0706	0.1924	0.1924	−0.0089	−0.0089
2	0.0572	0.1278	0.1588	0.3512	−0.0045	−0.0134

of consumer characteristics allows the definition of more targeted and effective strategies (Risvik et al. 2003). The low step 2 model predictivity and the position of products in Fig. 2 (spread around and close to the origin) also suggested the existence of different consumer groups. A classification was performed on the second step components $\mathbf{t_{z1}}$ and $\mathbf{t_{z2}}$: such components are supposed to take into account both the variability in consumer descriptors (in $\mathbf{Z}$) and the relationship among consumer preferences and product descriptors (in $\mathbf{C_X}$). In order to define groups of consumers homogeneous with respect to both consumer and product descriptors, first of all a hierarchical ascendant classification has been performed on the scores $\mathbf{t_{z1}}$ and $\mathbf{t_{z2}}$ in order to choose the number of clusters. The ascendant hierarchical classification (Ward criterion) has pointed out two consumers which are probably outliers and, tending to cluster together, may deform the results of the classification. The ascendant hierarchical classification has therefore been repeated on the 90 remaining women after removal of the two outliers. In order to obtain clusters not containing an excessively low number of units, a partition with three clusters has been chosen, and a k-means clustering has been then performed over the same components with $k = 3$. The three obtained clusters are respectively composed of 31 (group1), 33 (group2) and 26 (group3) individuals. Two-step L-PLS Regression has then been performed over each group in order to obtain local models and subsequently to

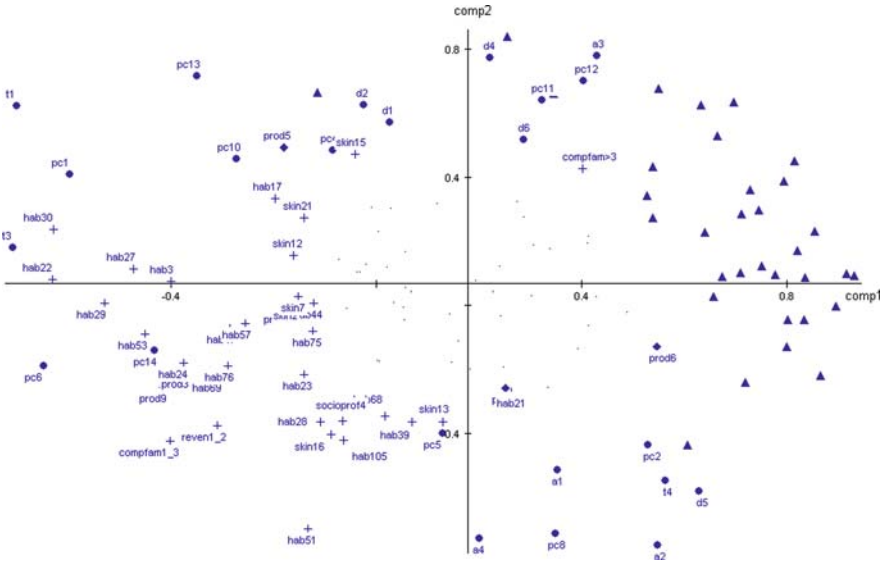

Fig. 3 Two-step L-PLSR local model for consumer class 2. Consumer preferences (*filled triangle*) are related to product descriptors (*filled circle*) and consumer descriptors (+); products are represented as diamonds (*open diamond*)

compare the groups in order to highlight eventual differences in the preference models. Results from group 2 will be discussed here in detail. Tables 4 and 5 show the main results from the two steps, while Fig. 3 shows the results for the local two-step L-PLSR model estimated on class 2. Preferences are now located far from the origin of axes. As to product descriptors, while some keep the importance they had in the global model in explaining preferences, others gain importance for this particular class (such as product descriptor t2). It is easier to characterise the consumers belonging to this group with respect to socio-demographic variables (compfam>3 and reven6), to their skin characteristics (skin5) and to their behaviours (hab55). Finally, products prod6 and prod8 seem to be the most appreciated products from the consumers in class 2.

5 Discussion and conclusions

Two-step PLS regression has allowed the establishment of relationships among consumers preferences, product sensory and physico-chemical descriptors and consumers' behaviour and cosmetic habits descriptors. PLS basic principles have allowed the construction of the model although data was characterised by a strong multicollinearity (especially consumers' preferences and physico-chemical descriptors). Moreover, differently from Martens L-PLSR, the optimised criterion is statistically interpretable and homogeneous in the entire analysis. A major advantage, finally, was that no imputation, external estimation, or deletion was required, although 110 values were missing in matrix **Z** (0.77% of all cases).

The application of this methodology to data from the cosmetic industry has allowed the definition of groups among consumers with homogeneous consumption and life habits, as well as showing preferences for similar product characteristics. The knowledge of the preference models of the consumers from each group as well as of their behavioural habits can lead to the definition of a more targeted communication strategy (Risvik et al. 2003).

Future research perspectives will aim at the definition of an iterative algorithm, allowing an "optimal" classification by taking into account PLS principles, mainly **Y** variables prediction, as PLS Typological Regression does (Esposito Vinzi et al. 2004). On-going research is also focused on the study of the double error-term in Eq. (8) due to the two-step procedure. The error term is evidently related to the quality of the model (in terms of predictivity) and to the loss of information in each step: an appropriate information selection in each step can help in reducing the loss of information and hence the error dimension. However, instead of performing information selection as variable selection by removing scarcely informative variables, current research is focused on information selection by means of partial analysis criteria, under a Orthogonal Projection to Latent Structures (O-PLS) point of view (Trygg and Wold 2002) or in a Oriented PLS (OR-PLS) optic (Rayens and Andersen 2004). Finally, the existence of two variable blocks of different nature in matrix **X**, i.e. physico-chemical and sensory variables influencing preferences should be considered: PLS Path Modelling (Wold 1975, 1982, 1985, Chatelin et al. 2002, Tenenhaus et al. 2005) could allow the estimation of local causal models (one for each consumers' group) linking preferences to the two row-descriptors' blocks.

Acknowledgments The present paper has been financially supported by short term mobility grant (*Borsa di Breve Mobilità*) from University of Naples *Federico II*. The authors wish to thank the sensory and product development laboratory teams of Bourjois for their important contribution to the data, and in particular Michele Vincent for her advice and encouragement as well as Michel Tenenhaus for his precious suggestions.

References

Amenta P, D'Ambra L (1997) L'analisi in componenti principali in relazione a un sottospazio di riferimento con informazioni esterne, Quaderni di Statistica, 18, DMQTE, Pescara

Chatelin YM, Esposito Vinzi V, Tenenhaus V (2002) State of Art on the PLS Path Modelling through the available software. HEC Research Papers, 764/2002

Esposito Vinzi V, Lauro C, Amato S (2004) PLS typological regression: algorithmic, classification and validation issues. In: Vichi M, Monari P, Mignani S, Montanari A (eds) New developments in classification and data analysis, Springer, Berlin Heidelberg New York, pp 133–140

Giordano G, Scepi G (1999) La progettazione della qualità attraverso l'analisi di strutture informative differenti. In: Proceedings of XV Riunione Scientifica SIS

Höskuldsson A (1988) PLS regression methods. J Chemometrics 2:211–288

D'Ambra L, Lauro, NC (1992): "Non Symmertical Exploratory Data Analysis", Statistica Applicata, 4(4), pp 511–529

Martens H, Anderssen E, Flatberg A, Gidskehaug LH, Høy M, Westad F, Thybo A, Martens M (2005) Regression of a data matrix on descriptors of both its rows and its column descriptors via latent variables: L-PLSR. Comput Stat Data Anal 48(1):103–123

Rayens W, Andersen A (2004) Oriented Partial Least Squares, Rivista di Statistica Applicata, Italian Journal of Applied Statistics, RCE Edizioni, Napoli, 15(3), pp 367–388
Risvik E, UelandØ, Westad F (2003) In: Segmentation strategy for a generic food product in Proceedings of the 5th Pangborn Sensory Science Symposium, Boston, MA, USA
Stone M, Brooks RJ (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares, and principal component regression. J R Stat Soc Ser B 52:237–269
Takane Y, Shibayama T (1991) Principal component analysis with external information on both subjects and variables. Psychometrika 1:97–120
Tenenhaus M (1998) La régression PLS: théorie et pratique, Editions Technip, Paris
Tenenhaus M, Esposito Vinzi V, Chatelin YM, Lauro C (2005) PLS path modelling. Comput Stat Data Anal 48(1):159–205
Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). J Chemometrics 16:119–128
Umetrics (2002) SIMCA-P and SIMCA-P+ 10 User Guide, UMETRICS AB, Umeå, Sweden
Wold H (1975) Modelling in complex situations with soft informations. In: Third World Congress of Econometric Society, Toronto, Canada, 21–26 August 1975
Wold H (1982) Soft modelling: the basic design and some extensions. In: Jöreskog KG, Wold H (eds) Systems under indirect observation, vol 2. Wiley, New York, pp 587–599
Wold H (1985) Partial least squares. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 8. Wiley, New York, pp 581–591
Wold S, Martens H, Wold H (1983) The multivariate calibration problem in chemistry solved by the PLS method. In: Paper presented at the Matrix Pencils, Heidelberg