doi.org/10.26398/IJAS.0032-006

A PATH-MODELING APPROACH TO PREFERENCE DATA ANALYSIS

Simona Balzano[1]

Department of Economics and Law, University of Cassino, Cassino, Italy

Giuseppe Giordano

Department of Social and Political Studies, University of Salerno, Fisciano, Italy

Natale Carlo Lauro

Department of Economics and Statistics, University of Naples Federico II, Naples, Italy

Abstract. *The paper introduces the preference data analysis, as settled in the scope of Conjoint Analysis, into the general framework of the Partial Least Squares approach to Structural Equation Models. The aim is to define a common background for interpreting the Conjoint Analysis results in terms of a path model. Once established this correspondence, we discuss how the proposed approach may represent a tool to enrich the data collection, model specification and results interpretation phases.*

Keywords: *Conjoint Analysis, Market Segmentation, Partial Least Squares, Path model, Structural Equation Model.*

1. INTRODUCTION

Metric Conjoint Analysis (CA) represents one of the main methods for the analysis of preference data, aiming at estimating the importance of some selected characteristics of a set of potential products or services, called *stimuli*, as a function of the global preference expressed by a set of judges. It is a *decompositional* method, mainly based on Design of Experiments and OLS Regression model (Green & Rao, 1971; Green & Srinivasan, 1978; Green & Krieger, 1991).

The peculiar data structure of CA is based on the combination of *i)* a *design matrix* holding dummy variables, describing a set of *stimuli* in terms of the presence/absence of specific *levels* of some fixed *attributes* and *ii)* a *preference matrix*, where each column includes the ratings expressed by a single *judge* on the set of stimuli. The method aims at estimating the *partial utility* coefficients (*part-worths*), i.e. the weight of each level in composing the *individual preference*

[1] Simona Balzano, s.balzano@unicas.it

expressed by each judges. Individual preference models are usually aggregated by averaging the part-worth coefficients to obtain the *aggregate preference model*.

Different variants of the traditional OLS regression model have been considered in the specialized literature (for a discussion see Furlan & Corradetti, 2005).

We refer to Structural Equation Model (SEM, Jöreskog, Sörbom, 1979) and to the Partial Least Squares approach to SEM (PLS-SEM or PLS-PM, Wold, 1975; Tenenhaus et al., 2005) for its estimation. Based on the block structure of the design matrix and on the hypothesis that stated preference depends on the characteristics of stimuli, we propose to interpret the Conjoint Analysis in terms of a path model in a multivariate multiple regression framework.

This model specification allows to estimate, at the same time, the aggregate preference as a multivariate synthesis of the individual preferences and the part-worth coefficients.

A reading of the Conjoint Analysis model in the PLS framework was proposed by Tenhenaus in 1998 using PLS regression (Wold et al., 1983). The use of PLS path-modeling in the analysis of consumer behavior was introduced by Pagès and Tenenhaus in 2001 for dealing with groups (*blocks*) of variables observed on the same set of stimuli. They used Multiple Factor Analysis (Escofier, Pagès, 1994) as an exploratory tool to define blocks, combined with PLS Regression and Path-Modeling, for predicting hedonic judgements on the basis of sensory and physicochemical characteristics of a set of products. Later, Tenenhaus et al. (2005) used the same strategy to find clusters of homogeneous consumers. Some more recent examples of application of PLS-PM to multiblock preference data structures in sensometrics can be found in Menichelli, 2013; Cariou et al., 2018; Llobella et al., 2020.

In the model we propose, referred to as *PLS-PM preference analysis*, we look at rating data collected at individual level as observed expression (manifest variables) of a latent construct representing the aggregate preference. More specifically, in the path model we specify the aggregate preference is defined as an "endogenous" latent variable that is affected by a set of "exogenous" latent variables, representing the attributes.

In this setting, we use a simulated dataset to establish the correspondence between the basic elements and estimates provided by the metric CA and the PLS-PM preference analysis. Then we show how the proposed model allows to detect possible heterogeneity among the set of respondents in order to identify different market segments.

Many authors proposed segmentation strategies in the multiblock data analysis and PLS-PM context (Ringle et al., 2013; Llobella et al., 2020). However, it is worth to note that the stimuli are the statistical units, while respondents' ratings play the role of variables, thus within the path model finding market segments is a clustering of variables issue.

The paper is organized as follows: Sections 2 and 3 give a general recall to the CA and to the PLS-PM models respectively; Section 4 introduces the PLS-PM preference analysis and describes the details of the model specification; in Section 5 we refer to a simulated dataset for assessing the correspondence and coherence between the CA and PLS-PM results; then, in Section 6 we apply our findings to a case-study, showing how an enhanced interpretation of PLS-PM results is possible.

2. THE METRIC CONJOINT ANALYSIS MODEL

In this context we refer to the metric CA approach in which the multiple linear regression model is used to estimate the part-worth coefficients for each judge.

Let $\mathbf{X}$ be the design matrix of size $S \times L$, whose rows refer to the administered stimuli and columns to the $L = \sum_{k=1}^{K} l_k$ levels of K attributes (i.e. the k^{th} attribute has l_k levels). The design matrix $\mathbf{X}$ is a partitioned matrix consisting of K blocks of indicator matrices $\mathbf{X}_k$ $(k = 1, \ldots K)$:

$$\mathbf{X} = [\mathbf{X}_1 | \ldots | \mathbf{X}_k | \ldots | \mathbf{X}_K] \tag{1}$$

The matrix $\mathbf{Y}$:

$$\mathbf{Y} = [\mathbf{y}_1 | \ldots | \mathbf{y}_g | \ldots | \mathbf{y}_G] \tag{2}$$

has dimensions $S \times G$ and holds the responses given by G judges to the S stimuli.

The explanatory variables $\mathbf{x}_j$ $(j = 1, \ldots, L)$ are binary indicators associated with the levels of each attribute based on experimental design. The basic individual CA model for the judge g is expressed as:

$$\mathbf{y}_g = \beta_{g1}\mathbf{x}_1 + \ldots + \beta_{gL}\mathbf{x}_L + \varepsilon_g \qquad g = (1, \ldots, G) \tag{3}$$

where ε_g is the vector of error terms. The regression coefficients $\beta_{gj}, (j = 1, \ldots, L)$ can be interpreted as individual part-worth coefficient of the level j for judge g.

According to the above notation, the CA model in Equation 3 can be written as the multivariate regression model:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E} \tag{4}$$

where the matrix **B** holds the OLS estimation, interpreted as part-worth coefficients associated to the attribute-levels for each judge, while **E** is the error term matrix. Due to rank deficiency of the design matrix, coefficients are identified by imposing usual constraints, e.g. by dropping one column for each factor and posing the relative coefficient equal to zero. Let us note that the coefficients in **B** are obtained as in separate OLS regressions. The individual part-worth coefficients are then used to predict market segments and product positioning (Green, Krieger, 1991).

An aggregate utility function requires a suitable synthesis of individual utilities. Among many possible solutions, the average preference model is commonly used, lying on the hypothesis that respondents belong to a homogeneous set. The homogeneity can be either defined based on some stratification (typically socio-demographical) variables (i.e. *ex-ante* segmentation) or deduced from the detection of clusters (i.e. *ex-post* segmentation).

3. THE PLS APPROACH TO STRUCTURAL EQUATION MODEL

In reading the CA model in terms of a Structural Equation Model we refer to the PLS approach (Wold, 1975; Tenehaus et al., 2005; Hair et al., 2014), which is one of the most used estimation method for SEM and is coherent with the predictive feature of the multivariate regression model underlying Conjoint Analysis.

According to the traditional PLS-PM notation, let **Z** be a data matrix partitioned by column in H blocks:

$$\mathbf{Z} = [\mathbf{Z}_1|\dots|\mathbf{Z}_h|\dots|\mathbf{Z}_H] \tag{5}$$

of order $n \times p$, where each block $\mathbf{Z}_h$ $(h = 1,\dots,H)$ has dimensions $n \times p_h$ and $\sum_{h=1}^{H} p_h = p$, n is the number of stimuli and p the number of all attributes' levels.

A *path diagram* (Figure 1) is the typical representation of a causal model where each block $\mathbf{Z}_h$ is a set of manifest variables and is conceptually connected to a latent variable ξ_h. In such a diagram, rectangles represent Manifest Variables (MV i.e. $\mathbf{z}_{hj}$, $j = 1,\dots,p_h$) ellipses are the Latent Variables (LV i.e. ξ_h) and arrows describe the relations between them, supposed to be linear.

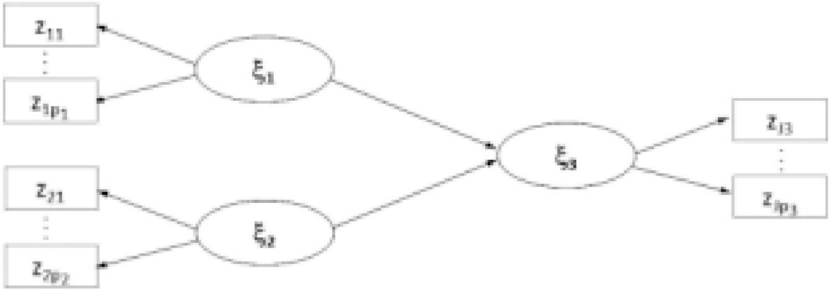

Fig. 1: Path diagram with two exogenous latent variables (ξ1 and ξ2), one endogenous latent variable (ξ3). All indicators are reflective.

Two sub-models are combined in the path diagram: the *structural model* (also called *inner* or *path* model) including the relations among latent variables, and the *measurement model* (also called *outer* model), including the relations between each manifest variable and the corresponding latent variable.

The directions of the arrows describe different model specifications of both sub-models: the structural model includes *exogenous* latent variables, i.e. latent variables which do not depend on other latent variables, and *endogenous* latent variables, i.e. latent variables which depend on other latent variables; the measurement model can be *reflective* (referred as mode A), when the manifest variables are a reflection of the latent variable (i.e. dealing with an independent LV and dependent MVs) or *formative* (mode B), when the manifest variables affect the latent variable (i.e. dealing with independent MVs and a dependent LV). It is worth to note that block's unidimensionality is a necessary condition for a reflective measurement model: lacking of unidimensionality requires a formative model.

PLS path modeling is defined to deal with metric data for the dependent MV. However, quasi-metric data stemming from multi-point scales, such as discrete ratings or rankings, are also acceptable as long as the scale points can be assumed to be equidistant and there are five or more scale points (Rhemtulla et al., 2012). It is also possible to include categorical variables in a model. If a categorical variable has only two levels (i.e., it is dichotomous), it can immediately serve as a construct indicator. As in the case of the conjoint analysis model, categorical variables with more than two levels should be transformed into as many dummy variables as levels minus one (the reference level). Preferably, categorical variables should only play the role of exogenous variables in a structural model (Henseler, 2017). A non-metric approach to PLS-PM has been addressed in (Russolillo, 2012) to handle both metric and non-metric variables at once, based on optimal scaling features.

3.1. THE PLS-PM ALGORITHM

The PLS approach to SEM is an iterative algorithm aimed at estimating latent variables scores through alternated simple and multiple linear regressions. On a first instance it provides an external estimation of the latent variable, then an internal estimation is obtained.

The PLS-PM algorithm estimates 3 sets of parameters:

- Latent Variable scores u_h,
- Path coefficients (or inner weights) $d_{hh'}$ of exogenous LV ξ_h on the endogenous $\xi_{h'}$,
- Outer weights w_{hj} of the manifest variables z_{hj} on ξ_h,

and is based on alternating, until convergence, an *external* and *internal* estimate of the LV's, based on OLS regressions, according to the following steps:

1. Outer estimation: the estimate v_h of the LV ξ_h is obtained as:

$$v_h \propto \pm \left(\sum_{j=1}^{p_h} w_{hj} z_{hj} \right) \tag{6}$$

   that is a linear combination of the MV z_{hj} $(h = 1, \ldots, H;\ j = 1, \ldots, p_h)$, using arbitrary weights w_{hj} on the first iteration.

   In a reflective block, each MV depending on the LV, the outer estimation is based on a set of p_h simple regression models; in a formative block, the LV depending on its MVs, the estimation comes through a multiple regression model.

2. Inner estimation: based on v_h, a new estimate u_h is obtained for each LV ξ_h relating them to one another according to the structural scheme as follows:

$$u_h \propto \sum_{h'} d_{hh'} v_{h'} \tag{7}$$

where the $d_{hh'}$ are the inner weights and can be set equal either to the sign of the correlation coefficient between the outer estimates v_h and $v_{h'}$ of the h-th and the h'-th LVs (Centroid scheme), or to the correlation coefficient between them (Factor scheme), or to their regression coefficient (Path scheme).

3. Computation of the *outer weights* w_{hj} based on the covariance between observed z_{hj} and the inner estimate of the LV u_h:

$$w_{hj} \propto cov\left(u_h, z_{hj}\right). \tag{8}$$

At each iteration, the partial results for outer weights from Equation 8 are used in Equation 6 for the next outer estimation step. Once convergence between inner and outer estimates is reached, ξ_h estimates are used in a set of OLS regressions for determining the *path-coefficients* (or *inner weights*) i.e. the coefficients of the structural relations. For more details about the algorithm see (Tenehaus et al., 2005).

After the iterative procedure, some indicators are computed for interpreting results and model fitting. Among them, we will focus on *loadings* ρ_{z_{hj},u_h}, i.e. the correlations between each manifest variable and the final estimate of the corresponding latent, and *cross-loadings* $\tau_{z_{hj},u_{h'}}$, i.e. the correlations between each manifest variable and the other latent variables.

4. THE PLS-PM PREFERENCE ANALYSIS

In this section we will link the Conjoint Analysis data structure to the PLS-PM specification by adapting the path-model to the metric CA. We then restate the matrix $\mathbf{Z}$ in Equation 5 as the juxtaposition of matrices $\mathbf{X}$ and $\mathbf{Y}$ in Equations 1 and 2. $\mathbf{Z}$ is partitioned in $H = K + 1$ blocks and has dimension $S \times (G + L)$.

As shown in Figure 2, in this specification each block of $\mathbf{X}$, that is each of the K attributes of the design matrix, is related to an exogenous latent variable ξ, while the preference matrix $\mathbf{Y}$ is related to an endogenous latent variable η.

In order to represent the dependency relationships stated by the CA model, we adopt all formative measurement models for exogenous blocks (attribute-levels) and the reflective measurement model for the endogenous latent variable (preference data).

Some considerations led to this choice (see Section 3): *i*) the blocks of $\mathbf{X}$ do not specify a unidimensional concept (i.e., the reflective model does not comply);

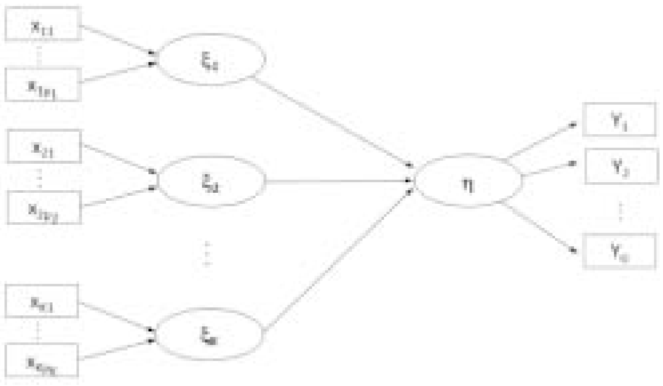

Fig. 2: The path-model for preference data analysis

ii) the model as a whole should respect the CA basic hypothesis that the characteristics of the stimuli in terms of attribute-levels (blocks of the matrix **X**) have an effect on the revealed preferences (block **Y**); *iii*) the aggregate preference latent construct determines the individual rating expression, then the block **Y** is set as reflective.

In the next section we show that in such a model the cross-loading $\tau_{x_{hj},\eta}$, between the j^{th} level of the k^{th} factor x_{hj} and η, can be interpreted as the part-worth coefficient of x_{hj}, while the outer loadings of the latent variable η, $\rho_{\mathbf{Y},\eta}$, measuring the correlations between the observed judges' ratings and the underlying preference construct η, can be used to detect possible heterogeneity among judges.

Due to the peculiar structure of the design matrix **X** and, thus, to the independence between blocks (factors are orthogonal each other), each $\tau_{x_{hj},\eta}$ is equal to the product between the outer loading of x_{hj} and the inner loading linking the block ξ_j with the latent preference η, i.e. $\tau_{x_{hj},\eta} = \rho_{x_{hj},\xi_j} \times \rho_{\xi_j,\eta}$.

Results from the two approaches overlap when dealing with a balanced design (and both the CA and PLS-PM estimates are provided at individual level).

We show these properties through a basic case-study with simulated preference data.

5. A TOY-STUDY TO DESCRIBE THE PLS-PM PREFERENCE ANALYSIS

In this section we apply the proposed approach to simulated data in order to check the expected relationships and to give a clear interpretation of the PLS-PM estimates in terms of the traditional metric CA results. Afterwards, we will be able to interpret the PLS-PM estimates applied to more general preference data structures.

Let $\mathbf{X}$ be the matrix holding the 2^4 full factorial experiment related to four factors $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4)$ described by two levels each (low level = -1; high level = $+1$), for a total of 16 profiles.

Let $\mathbf{Y}$ be the matrix holding the preference ratings assigned by 50 simulated judges to the 16 stimuli profiles.

The preference ratings in $\mathbf{Y}$ have been simulated by imposing the presence of two homogeneous groups of respondents (25 judges each). Data were generated from two multivariate Wallenius's non-central hypergeometric distribution (Wallenius, 1963), with suitable parameters to obtain the 16 ratings of 50 categorical random variables (the judgments) interpreted as preference data.

The multivariate Wallenius distribution rules the extraction of n balls of k different colors from a *biased urn*. In our case the k colors represent stimuli with given probability of being extracted $p(S_1), ..., p(S_k)$; the urn is biased since the probabilities for each stimulus to be extracted are different. The process simulates the distribution of a total amount of score (n) among the stimuli. It is equivalent to assign one point to each stimulus each time it is extracted, so that the value assumed by the random variable represents the final rating to each stimulus. Simulated data can be obtained acting on the probabilities $p(S_1), ..., p(S_k)$ in order to reflect two groups of judges, each group having strong internal coherence but showing reverse preference patterns when compared to the other.

5.1. The analysis at individual level

Let us first consider the analysis of an individual model (data in Table 1), the OLS results are reported in Table 2.

The metric Conjoint Analysis results show evidence of significant effects of the factors, specifically $\mathbf{X}_1$ (p-value: < 0.001) and $\mathbf{X}_2$ (p-value: 0.003); the adjusted R^2 is equal to 0.773. Since contrasts in the design matrix are coded so that effects sum up to zero, the part-worth coefficients of the missing levels in each factor can be easily derived by changing the sign of the estimated coefficients, that is $\beta_{k2} = -\beta_{k1}$, $k = 1 \ldots, 4$, (i.e., if $\beta_{11} = +2.875$ then $\beta_{12} = -2.875$; and so on). In this trivial case, dealing with a balanced design of four factors at two levels and only one judge, the PLS-PM results are identical to the OLS estimates of the metric Conjoint Analysis model. Specifically, looking at Figure 3, the outer model is trivially determined being all loadings equal to 1 (as well as weights when y is

Tab.1: *y*: individual simulated preference scores, ties allowed; X: 2^4 Full Design Matrix; Contrasts: Sum = 0. Last row: Mean and St.dev. of *y*; Cor- relation coefficients between *Y* and each factor.

y	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$
3	-1	-1	-1	-1
1	-1	-1	-1	1
1	-1	-1	1	-1
3	-1	-1	1	1
6	-1	1	-1	-1
5	-1	1	-1	1
5	-1	1	1	-1
3	-1	1	1	1
5	1	-1	-1	-1
8	1	-1	-1	1
8	1	-1	1	-1
7	1	-1	1	1
10	1	1	-1	-1
15	1	1	-1	1
8	1	1	1	-1
12	1	1	1	1
$\mu_y = 6.250$ $S_y = 3.872$	$r_{y,X_1} = 0.767$	$r_{y,X_2} = 0.467$	$r_{y,X_3} = -0.100$	$r_{y,X_4} = 0.133$

Tab. 2: Individual Conjoint Analysis Model.

	Estimate		Std. Estimate	S.E.	t-value	$Pr(>\|t\|)$
(Intercept)	6.250	$***$	0.000	0.462	13.540	0.000
$\mathbf{X}_1$	2.875	$***$	0.767	0.462	6.228	0.000
$\mathbf{X}_2$	1.750	$**$	0.467	0.462	3.791	0.003
$\mathbf{X}_3$	-0.375		-0.100	0.462	-0.812	0.434
$\mathbf{X}_4$	0.500		0.133	0.462	1.083	0.302
R^2 : 0.833;	$Adj.R^2$: 0.773;		$N.obs.$: 16;	$RMSE$: 1.85 on 11 *dof*		

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

standardized). As a consequence, for the latent preference, the *Redundancy*, that measures the quality of the structural model for each endogenous block, corresponds to the R^2 of the CA model.

Then, the PLS-PM cross-loadings, representing the linear correlation coefficients between each manifest factor block X and the latent preference η, here coincide with the inner weights. In a general setting, they are equal to the product between the inner weights and the outer loadings, that is $\tau_{x_{k1},y} = \rho_{x_1,X_k} \times \beta_{X_k,y}$.

Note that the cross-loadings are equal to the linear correlation coefficients between the dependent variable and the X factors (see Tables 1 and 2) and if the dependent variable were standardized they would be equal to the part-worth coefficients.

This results established a proper specification of the PLS-PM in terms of the CA model. In the following we extend the PLS-PM specification to the case of multiple responses.

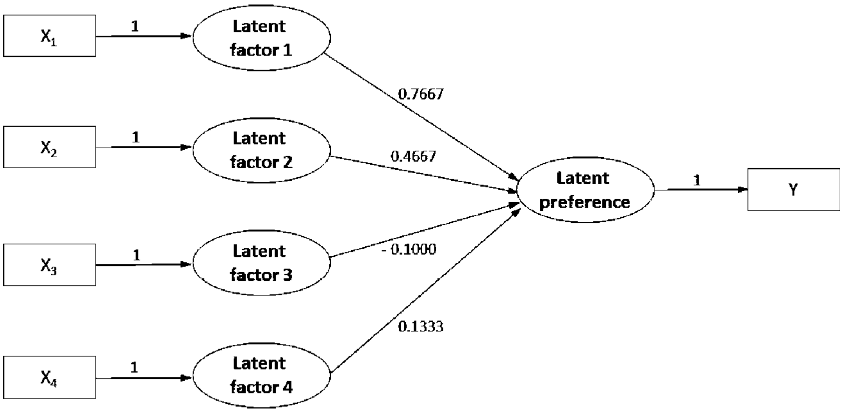

Fig. 3: The PLS-PM estimations: one-judge case

Let us now consider the whole set of simulated preference data. They are related to two homogeneous subgroups of judges revealing two opposite preference structures. In the classical CA framework it is possible to derive an aggregate model for the whole set of respondents. It is usually estimated by using as dependent variable the average of individual preference scores; the coefficients estimated for the aggregate, average model, are reported in Table 3.

The PLS-PM specification including the whole set of judges leads to the results shown in Figure 4. In this case the cross-loadings are still equal to the inner

Tab. 3: Aggregate Conjoint Analysis Model.

	Estimate		Std. Estimate	S.E.	t value	$Pr(>\|t\|)$
(Intercept)	6.250	***	0.000	0.072	87.198	0.000
$\mathbf{X}_1$	0.020		0.006	0.072	0.279	0.785
$\mathbf{X}_2$	0.005		0.001	0.072	0.070	0.946
$\mathbf{X}_3$	-0.188	*	-0.005	0.072	-2.616	0.024
$\mathbf{X}_4$	-0.168	*	-0.005	0.072	-2.337	0.040
R^2 : 0.530;	$Adj.R^2$: 0.360;		$N.obs.$: 16;	$RMSE$: 0.29 on 11 dof		

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

weights, being the outer weights equal to 1, but they are no longer identical to the standardized partial utility coefficients obtained by the aggregate CA model. This is expected, since the two approaches optimize two different criteria.

The main drawback of estimating CA model on the average preference is that the presence of two respondents' patterns is not considered and cannot be revealed by standard results. Further investigation is needed before carrying out the aggregated model by exploring the individual utility models for all respondents.

Conversely, PLS-PM takes into account the covariance structure of response variables and allows to discriminate among the response patterns.

In our case, the presence of two groups in matrix **Y** emerges from the individual outer loadings (as well as from outer weights) related to the latent aggregate preference (Vigneau and Qannary, 2003). In fact, in such cases judges belonging to different groups show opposite signs in their correlations with the latent aggregate preference (see Figure 4).

Generally speaking, the outer loadings can be seen as a tool to identify possible preference sub-models: they, in fact, measure judges' individual coherence with the aggregate preference, so that their different signs reveal different attitudes toward the aggregate preference, (i.e., heterogeneity among judges).

Beyond this trivial case, whereas different signs among judges loadings appear, we propose to use Cluster Analysis in order to reveal aggregate patterns of preference models present in the data and to understand how many clusters of judges are there. The dendrogram obtained by Complete Linkage method on the outer loadings is represented in Figure 5, where the presence of two groups (and their composition) clearly emerges.

Thus, we use these results to specify a new path model in order to consider

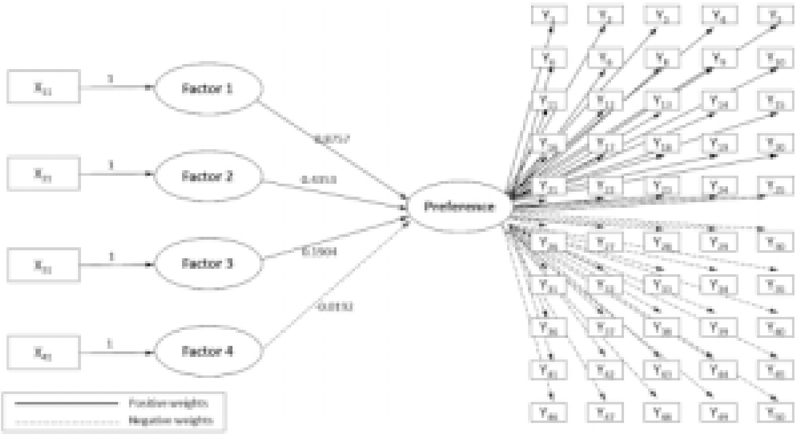

Fig. 4: The PLS-PM weights estimations on the whole judges set.

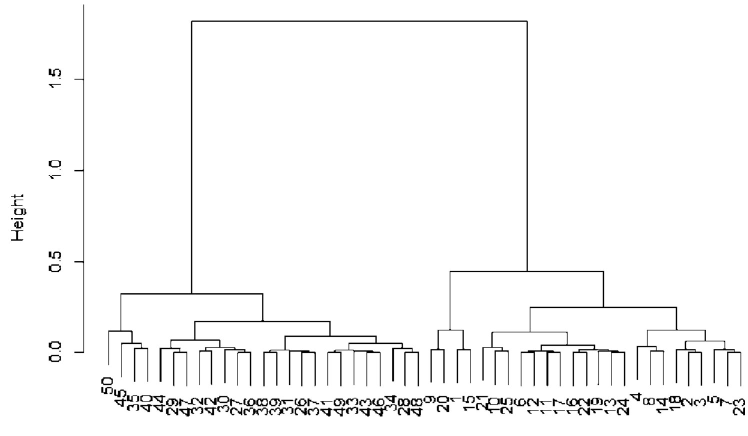

Fig. 5: Cluster of the 50 PLS-PM outer loadings. Euclidean distance, Complete Linkage method

two latent preference models. The new model specification and the results are shown in Figure 6. The two latent aggregated preference variables show a correlation of -0.96. The two preference models have been identified as follows:

$$Pref_1 = 0.8828X_1 + 0.4387X_2 + 0.1149X_3 - 0.0850X_4 \quad (9)$$

$$Pref_2 = -0.8512X_1 - 0.4226X_2 - 0.2617X_3 - 0.0461X_4 \quad (10)$$

where $Pref_1$ and $Pref_2$ are the scores corresponding to the two latent preference sub-models.

In conclusion, this toy study shows how it is possible to derive suitable model description from raw data and to assess the correspondence between the CA model and PLS-PM approach to preference data. In the following section we apply the PLS-PM preference analysis to a real dataset.

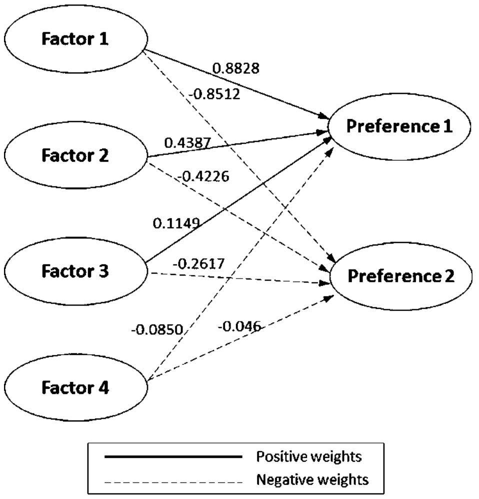

Fig. 6: The estimates of the PLS-PM inner model with two sub-models.

6. THE PLS-PM PREFERENCE ANALYSIS: A CASE STUDY

In this section an application of PLS-PM preference analysis is shown, using survey data included in the R package Conjoint (Bak, Bartlomowicz, 2012). The survey aimed at estimating the importance of 5 attributes describing different kinds of chocolates, namely Type (levels: Milky, Stuffed, Delicacies, Sour), Price (levels: Low, Medium, High), Packaging (levels: Soft pack, Hard pack), Size (levels: Small, Medium, Large) and Calories (levels: Low calories, High calories).

In this example, the fractional design $\mathbf{X}$ includes 16 profiles of chocolates and $\mathbf{Y}$ holds the preference ratings expressed by 87 judges.

As shown in Figure 7, the path model specification for preference analysis on this data includes five exogenous latent variables, representing the five attributes, each described by the corresponding levels, affecting the endogenous aggregate preference.

We use the effect coding for the attribute-levels and standardized manifest variables. In this case, we obtain the results for dropped levels by setting the sum of the cross-loadings for each attribute equal to 0.

In Figure 7 the outer and inner estimates are displayed, while Table 4 shows both the partial utility coefficients provided by the aggregate CA model and the PLS-PM cross-loadings between latent preferences and attribute-levels.

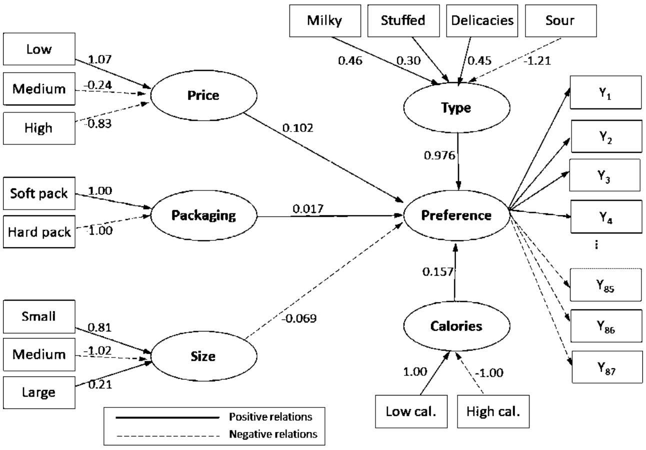

Fig. 7: Path-model estimations for preference analysis of chocolate data

Unless the CA aggregate part-worth coefficients and PLS-PM cross-loadings appear highly coherent, the presence of different signs of the outer loadings of the preference latent variable reveals heterogeneity among judges, suggesting to proceed with a Cluster Analysis. The Euclidean distance between judges is then derived and the *Complete linkage* criterion has been used.

Based on the dendrogram in Figure 8 we split the judges into three groups, each defining a latent preference sub-model.

Table 5 shows the three resulting preference models. They can be interpreted in terms of a classical Conjoint Analysis output, that is attribute importance, ideal product and maximum utilities can be derived.

The range of the cross-loadings values for each factor provides a measure of how important the factor is to overall preference. Factors with greater ranges play a more significant role than those with smaller ranges.

Attributes importance for the three preference submodels are shown in Table 6, where clearly appears that attributes *type* and *price* are discriminant between groups, while attributes *package* and *size* play a marginal role for all the three groups; *calories* is important for all groups. Specifically, groups 1 and 3 assign high importance to *type* and *calories*, while group 2 considers *price* and *calories*

Tab. 4: CA partial utility coefficients and PLS-PM cross-loadings

Attribute	Level	CA standardized coefficients	PLS-PM cross-loadings
Type	Milky	0.428	0.821
	Stuffed	0.288	0.744
	Delicacies	0.363	0.816
	Sour	-1.079	-2.381
Price	Low	0.263	0.100
	Medium	-0.047	-0.022
	High	-0.216	-0.122
Packaging	Soft	0.016	0.016
	Hard	-0.016	-0.016
Size	Small	0.078	0.026
	Medium	-0.068	-0.046
	Large	-0.010	0.020
Calories	Low	0.360	0.157
	High	-0.360	-0.157

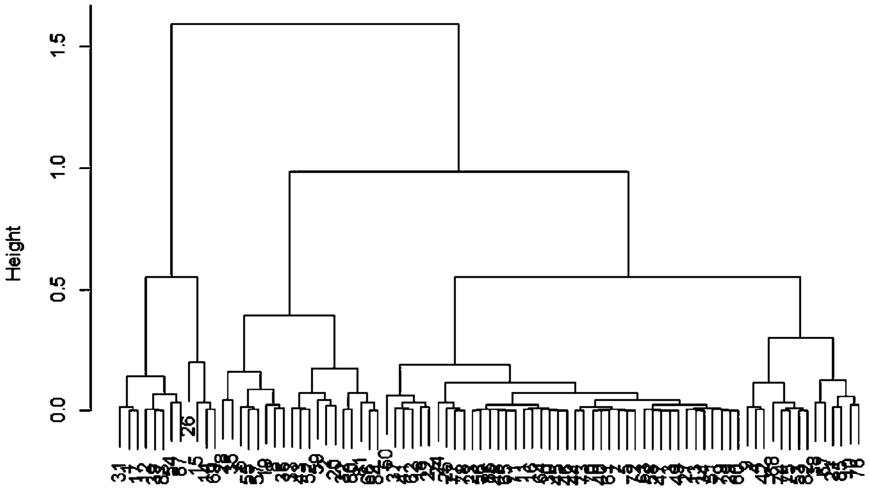

Fig. 8: Dendrogram of the cluster analysis on the judges' outer loadings.

as leading factors.

Table 7 summarizes the three ideal profiles. Provided that the attributes *type* is the most discriminant, it is evident how its levels differently characterize each group.

Tab. 5: PLS-PM cross-loadings between levels and the three latent preferences

Attribute	Level	Model 1	Model 2	Model 3
Type	Milky	-0.320	0.021	0.380
	Stuffed	-0.341	0.112	0.215
	Delicacies	-0.307	-0.079	0.359
	Sour	0.968	-0.054	-0.954
Price	Low	0.077	0.878	0.112
	Medium	-0.012	-0.220	-0.040
	High	-0.065	-0.658	-0.071
Package	Soft	-0.028	-0.037	0.017
	Hard	0.028	0.037	-0.017
Size	Small	0.022	0.046	0.061
	Medium	0.019	-0.034	-0.071
	Large	-0.042	-0.012	0.010
Calories	Low	0.185	0.236	0.217
	High	-0.185	-0.236	-0.217

Tab. 6: Attributes' importance in the three preferences submodels, relevant values are in bold.

	Type	Price	Package	Size	Calories	*Total*
Pref 1	**67.44**	7.32	2.89	3.30	**19.06**	*100*
Pref 2	8.12	**65.28**	3.14	3.40	**20.06**	*100*
Pref 3	**63.01**	8.64	1.61	6.24	**20.50**	*100*

Tab. 7: Ideal profiles for the three preferences submodels.

	Type	Price	Package	Size	Calories
Pref 1	Sour	Low	Hard	Small	Low
Pref 2	Stuffed	Low	Hard	Small	Low
Pref 3	Milky	Low	Soft	Small	Low

7. CONCLUSION

In this paper we propose to read the metric Conjoint Analysis model in the scope of PLS-PM approach.

The main contribution of this paper is to define a formal correspondence between the Conjoint Analysis results and the output of a path model estimation, where the model specification respects the Conjoint Analysis data structure and its decompositional nature.

We define a path model where each attribute is an exogenous latent variable, described by attributes' levels as manifest variables related to them, and the aggregate preference is an endogenous latent variable, described by all the preferences expressed by judges.

Then, a toy-study shows how *i)* CA partial utility coefficients correspond to the cross-loadings provided by the PLS-PM algorithm; *ii)* inner weights reproduce the importance of attributes on the aggregate preference; *iii)* outer loadings of preferences reveal the possible presence of clusters and/or outliers among judges.

The possibility to detect the presence of clusters makes the PLS-PM preference analysis a useful tool for detecting market segments.

Another strength of the PLS-PM preference analysis is that, unlike the classical CA model, it does not require attributes' levels to be constrained to a design matrix. Further research will be addressed to the generalization of this approach to *revealed preferences*, where observational data can be used for describing the set of stimuli, and a prior information on judges can be included.

REFERENCES

Bak, A., Bartlomowicz, T. (2012). Conjoint analysis method and its implementation in conjoint R package. In Pociecha J., Decker R., eds., *Data Analysis Methods and its Applications*, 239–248, C.H. Beck.

Cariou V., Qannari E.M., Rutledge D.N., Vigneau E. (2018). ComDim: From multiblock data analysis to path modeling. *Food Quality and Preference*, 67, 27–34.

Escofier B., Pagès J. (1994). Multiple factor analysis (AFMULT package).*Computational Statistics and Data Analysis*, 18, 121–140.

Furlan R., Corradetti R. (2005). An empirical comparison of conjoint analysis models on a same sample. *Rivista di Statistica Applicata*, 17(2), 141– 158.

Green, P.E., Krieger A.M. (1991). Segmenting Markets with Conjoint Analysis. *Journal of Marketing* 55(4), 20–31. doi:10.2307/1251954.

Green P.E., Srinivasan V. (1978). Conjoint analysis in consumer research: issues and outlook. *Journal of Consumer Research*, 5(2), 103–123.

Hair J. F., Hult T., Ringle C. M., and Sarstedt, M. (2014). *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Thousand Oaks, CA, Sage.

Henseler J. (2017). Partial Least Squares Path Modeling. In Leeflang P., Wieringa J., Bijmolt T., Pauwels K., eds., *Advanced Methods for Modeling Markets*. International Series in Quantitative Marketing. Springer, Cham.

Jöreskog K.G, Sörbom D. (1979). *Advances in Factor Analysis and Structural Equation Models, Abstract Books*. Cambridge, Massachusetts.

Llobella F., Cariou V., Vigneau E., Labenne A., Qannari E.M. (2020). Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. Application to sensometrics. *Food Quality and Preference* 79, 103520.

Menichelli E. (2013). *Multi-block methods for investigating consumer acceptance of food*. PhD thesis.

Pagès J., Tenenhaus M. (2001). Multiple factor analysis combined with PLS path modelling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements. *Chemometrics and Intelligent Laboratory Systems* 58, 261–273.

Ringle C.M., Sarstedt M., Schlittgen R., Taylor C.R. (2013). PLS path modeling and evolutionary segmentation. *Journal of Business Research* 66, 1318–1324.

Russolillo, G. (2012). Non-Metric Partial Least Squares. *Electronic Journal of Statistics* 6, 1641–1669. doi:10.1214/12-EJS724.

Rhemtulla, M., Brosseau-Liard, P.E., Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373.

Tenenhaus, M. (1998). *La Régression PLS: théorie et pratique*. Technip. Paris.