Conjoint analysis with textual external information

Simona Balbi[1], Giorgio Infante[1], Michelangelo Misuraca[2]

[1]Dip. di Matematica e Statistica – Univ. Federico II – 80127 Napoli – Italy

[2]Dip. di Economia e Statistica – Univ. della Calabria – 87036 Arcavacata di Rende – Italy

Abstract

In market researche we often need statistical methods able to jointly deal with numerical and textual variables. Particularly in market segmentation, several tools have been proposed for profiling customers, by surveying a sample of potential buyers. One of the most common statistical analysis aiming at identifying the "ideal" product for a target market is Conjoint Analysis (Green et Srinivasan, 1990). This technique considers a good/service as a proper combination of jointly evaluated characteristics. Each customer has to compare those combinations, i.e. the different potential (or actual) goods. The so-called full profile method often used in Conjoint Analysis sometimes seems too rigid in its experimental structure. Moreover, we are not always sure about the descriptions of the different goods, as we can consider only a small number of variables. We propose a procedure aimed at integrating Conjoint Analysis with textual information achieved by answers to open-ended questions. From a methodological viewpoint, this means introducing external information in the Takane and Shibayama (1991) sense. The proposed procedure improves the understanding of the preference structure and the ideal product definition. An empirical analysis on the watch market shows the effectiveness of this proposal.

Keywords: ideal product, preference structure, textual external information.

1. Introduction

The "ideal product" description obtained by Conjoint Analysis (C.A. in the following) is strongly dependent on how preferences are collected. The experimental design on which it relies seems to be too rigid for dealing with customer priorities. A free description of the ideal product written in natural language can supply new perspectives in marketing policies. Nevertheless it is difficult in sample surveys to deal with hundreds of interviewees (Lebart, 2004).

In the present paper we propose an integrated strategy based on C.A. together with textual data analysis, in the frame of a statistical approach to market segmentation. The first step of our strategy consists in introducing an open-ended question asking the free description of the ideal product, in a questionnaire typically designed for Conjoint Analysis. The individual answers to this question enable to verify and improve the segmentation process.

The theoretical challenge is to find the proper way of combining classical C.A. data structure with the unstructured information obtained by using natural language.

The methodological references are the multidimensional approach to C.A. proposed by Lauro et al. (1998), the factorial approach to regression coefficients in C.A. due to Giordano and Scepi (1999), and its application to textual data in Balbi and Giordano (2001).

Particularly we propose to consider the textual description as external information (Takane et Shibayama, 1991) in the analysis of the preference structure, in order to enrich the description of preferences and behaviours by means of textual information.

2. Methodological framework

Conjoint Analysis (Green et Srinivasan, 1990) deals with preference judgements about a set of *S stimuli* (goods or services). The *stimuli* are characterized by different categories (*levels*) of some variables (*factors*) according to a design matrix **X**. The preference judgements are expressed by rankings or ratings. Particularly C.A. aims at estimating the importance of each attribute/level in the evaluation of the preference of a *stimulus*. The estimates are computed for each judge individually.

2.1. Conjoint Analysis

Let **X** denote the experimental design matrix (*stimuli x levels*) partitioned in *H* juxtaposed indicator matrices $\mathbf{X_h}$ (h = 1, …, *H*), each referred to an attribute, with *L* equal to the sum of all the attribute levels. Let **Y** (*S*,*G*) denote the matrix having in each column the judgements expressed by the *g*-th judge (g = 1, …, *G*) on the *S stimuli*.

The classical C.A. model can be written as a multiple regression model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \qquad (1)$$

where **E** (*S*,*G*) is the error matrix, which has in columns the errors of each individual model and **B** (*L*,*G*) is the matrix of the *G* individuals part-worth coefficients associated to the *L* attribute levels:

$$\hat{\mathbf{B}}=\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \qquad (2)$$

Being $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ a singular matrix, Moore-Penrose generalized inverse $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}$ is usually performed.

2.2. Non symmetrical preferences analysis

A multidimensional approach to C.A. has been proposed in Lauro et al. (1998) in order to synthesize and visualize part-worth coefficients in a low-dimensional space.

It has propose a factorial decomposition of the matrix $(\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-}\mathbf{X}^{\mathrm{T}}\mathbf{Y})$, containing the variance of the preference judgements explained by the attributes.

From a geometrical viewpoint the method consists in a Principal Component Analysis performed on the images of the column vectors of the preference judgement matrix **Y** on the disjoint subspaces spanned by the column vectors of the design matrix partitions $\mathbf{X_h}$.

The maximum inertia axes are obtained as the eigen-solutions of the characteristic equation:

$$\left(\mathbf{Y'X}(\mathbf{X'X})^{-}\mathbf{X'Y}\right)\mathbf{u}_\alpha = \mathbf{Y'}\sum_{\mathbf{h=1}}^{\mathbf{H}}\mathbf{X_h}\left(\mathbf{X_h'X_h}\right)^{-}\mathbf{X_h'Yu}_\alpha = \boldsymbol{\lambda}_\alpha\mathbf{u}_\alpha \qquad (3)$$

for *α* = 1, …, *m* (where *m* is the rank of the matrix **X**), by imposing the ortho-normalizing constraints.

2.3. Factorial approach to regression coefficients

The geometrical approach to C.A. was extended by Giordano and Scepi (1999) by adding an additional matrix **Z**, as external information related to matrix **Y** rows, i.e. known characteristics of judges. In other words, they proposed a factorial approach to regression coefficient in C.A. in which classes of individuals with similar kinds of preferences and behaviours are projected into suitable reference subspaces.

Let consider two sets of regression coefficients: the set **B** defined as partial utilities explained by the product characteristics and the set **D** defined as the partial utilities, explained by socio-demographic variables.

Therefore, together with the classical solution for **B** coefficients in (2), they introduce another regression model:

$$\mathbf{Y}=\mathbf{D}\mathbf{Z}^{\mathrm{T}}+\mathbf{F} \tag{4}$$

where **Z** is an indicator matrix inherent to the auxiliary information about judges.

A singular value decomposition of the inter-relation matrix is proposed, considering:

$$\mathbf{\Theta}=(\mathbf{Z}\mathbf{Z}^{\mathrm{T}})^{-}\,\mathbf{Z}\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-} \tag{5}$$

where the generic value represents the estimated parameter of the relation through the socio-demographic characteristics and the conjoint analysis levels.

2.4. Textual Data Analysis

In Textual Statistics we need to transform documents we want to analyse in a similar fashion of native structured (numerical) data. In the parsing step the text is acquired, numerized and encoded by using a Bag-of-words scheme, with the aim of transforming each document into a vector.

According to this encoding strategy a document $\mathbf{d}_{\mathrm{j}}$ is represented as a vector (t_{1j}, …, t_{ij}, …, t_{Ij}), where t_{ij} is the “importance” of the *i*-th term listed in the vocabulary with respect to the *j*-th document.

This importance can be represented in several ways. In the following, we use the simplest way, i.e. a Boolean weighting scheme: 0/1 if the word is absent/present in the document. By juxtaposing the J document-vectors it is possible to obtain the *lexical table*, which cross-tabulate the documents with the vocabulary.

The coding step, in which we consider a document collection as input and a statistically analyzable matrix as output, is one of the most important in a textual statistics framework. In particular choosing the analyzed units and the weighting system assumes a main role, depending on the nature of the documents and the aim of the analysis.

In the proposed strategy the textual data are given by short open-ended questions and the aim is to deeply consider the judge’s preferences. In this frame it is convenient to consider the lemmas as units, so that the meanings are quite disambiguated, and a Boolean weighting system for avoiding a too high “importance” of the most frequent terms respect to the length of the documents.

3. The strategy

In order to obtain a better specification of Conjoint Analysis results, we propose an integrated strategy based on C.A. together with textual data analysis.

In the frame of our proposal the questionnaire is divided into three sections:

- in the first section each judge has to fill his own socio-economic and demographic data;

- in the second one the judges have to describe their behaviour in using the product or the service considered;
- in the last section the judges have to answer to an open-ended question like "*What's your ideal product/service?*"

Moreover we consider a set of cards, one for each *stimulus*, as in the classical C.A. approach.

3.1. Data structure

Let **Y** be the preference matrix, whose columns consist of the judgements given by *G* judges with respect to *S stimuli*. This matrix is centred in the rows and in the columns. Let **X** be the experimental design matrix, whose *L* columns are the levels of the attributes. Let **Z** be a presence/absence matrix relative to information concerning with the characteristics and behaviours of the *G* judges. We achieve this information from the first and the second part of the questionnaire.

Let **T** be a lexical table cross-tabulating the *G* descriptions, obtained by the open-ended question on the product/service, and the *V* words of our vocabulary, with general element t_{ij}, presence/absence of the *i*-th word in the *j*-th description.

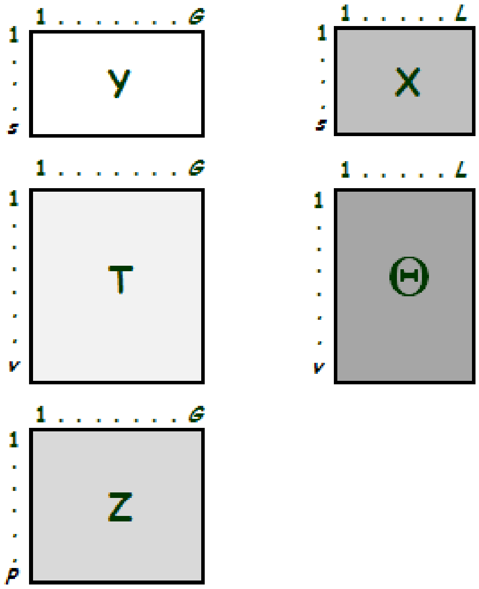

Figure 1 - Data structure

3.2. Conjoint Analysis with external information

In order to obtain a description of the ideal product by judges, we can consider textual information as external information, in Takane and Shibayama sense (1991).

We introduce **T** for linearly constraining the **B** matrix, as in Takane and Shibayama strategy. Thus we build the **Q** matrix of dimension (*V,L)*:

$$\mathbf{Q}=\left(\mathbf{T}\mathbf{T}^{\mathrm{T}}\right)^{-}\mathbf{T}\mathbf{B}^{\mathrm{T}}=\mathbf{W}\boldsymbol{\Xi}\mathbf{V}^{\mathrm{T}} \qquad (6)$$

where q_{jl} is the estimated parameter linking the attribute levels to the judges textual descriptions of the ideal product.

Now we perform the SVD of the **Q** matrix, with the ortho-normalizing constraints:

$$\mathbf{W}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{W}=\mathbf{I}_{\mathrm{L}}$$
$$\mathbf{V}\left(\mathbf{T}\mathbf{T}^{\mathrm{T}}\right)\mathbf{V}=\mathbf{I}_{\mathrm{V}} \tag{7}$$

In this way, we can visualise the terms in the verbal descriptions together with the C.A. levels on a common principal plane. The information about the judges can be projected as supplementary points (Lebart et al., 1984), so to enrich the global interpretation of C.A. results.

4. A case study: the watch market

To illustrate our proposal, an application on watch market has been carried out. A sample of 150 individuals, stratified respect to the age and the gender has been interviewed. The levels and the factors used in the C.A. have been selected considering expert knowledge. Individuals were asked on: the *number of watches* belonging to them; the interest on the *brand*; the main *reason for choosing* a watch (Section 2), together with their socio-demographic characteristics (Section 1). Then they were asked to answer to the question "*Describe your ideal watch*".

Conjoint Analysis section was the last one in the questionnaire, in order to avoid influences in the free description of the ideal product. In the experimental design six factors with two levels were considered (Figure 2).

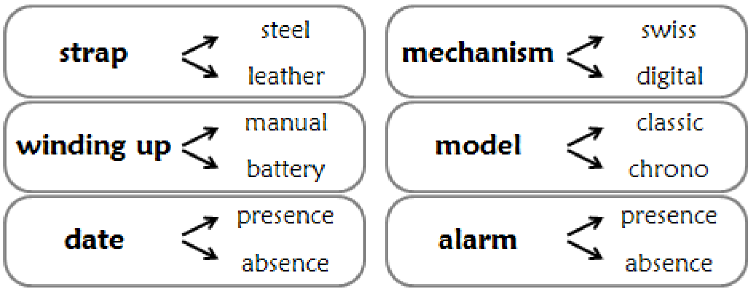

Figure 2 – The experimental design

The answers to the open-ended question have been normalized in order to reduce the possibility of *data splitting*. After carrying out a quite in-depth lexicalisation in order to avoid trivial cases of ambiguity, a vocabulary of 215 textual forms, partially marked with *Part of Speech* tags, has been made out. Furthermore, a stop list was considered to eliminating the instrumental terms and a special threshold was introduced for the infrequent terms.

The Figure 3 shows the C.A. levels on the first factorial plane. The first factorial plane represents the 44% of the total inertia. We can see that the first factorial axis opposes *mechanical* watches with *manual winding up* on the right and *digital* watches with *battery* on the left. In this axis we can see the contraposition between classical watches and modern watches, as *ideal product*.

On the second factorial axis, we can see the contraposition between informal watches (with date and alarm) and elegant watches (no date, no alarm and leather strap).

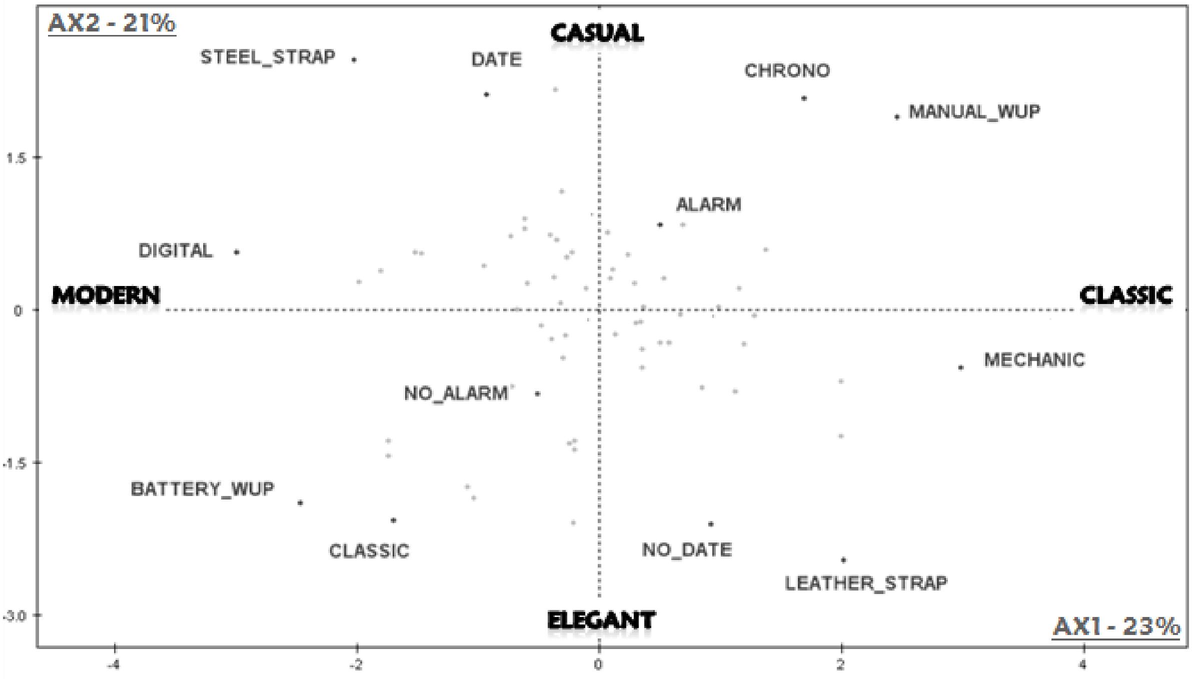

Figure 3 – Factorial representations of the C.A. levels

This coarse market segmentation based on the proposed stimuli, as in standard C.A., can be refined by introducing the textual information.

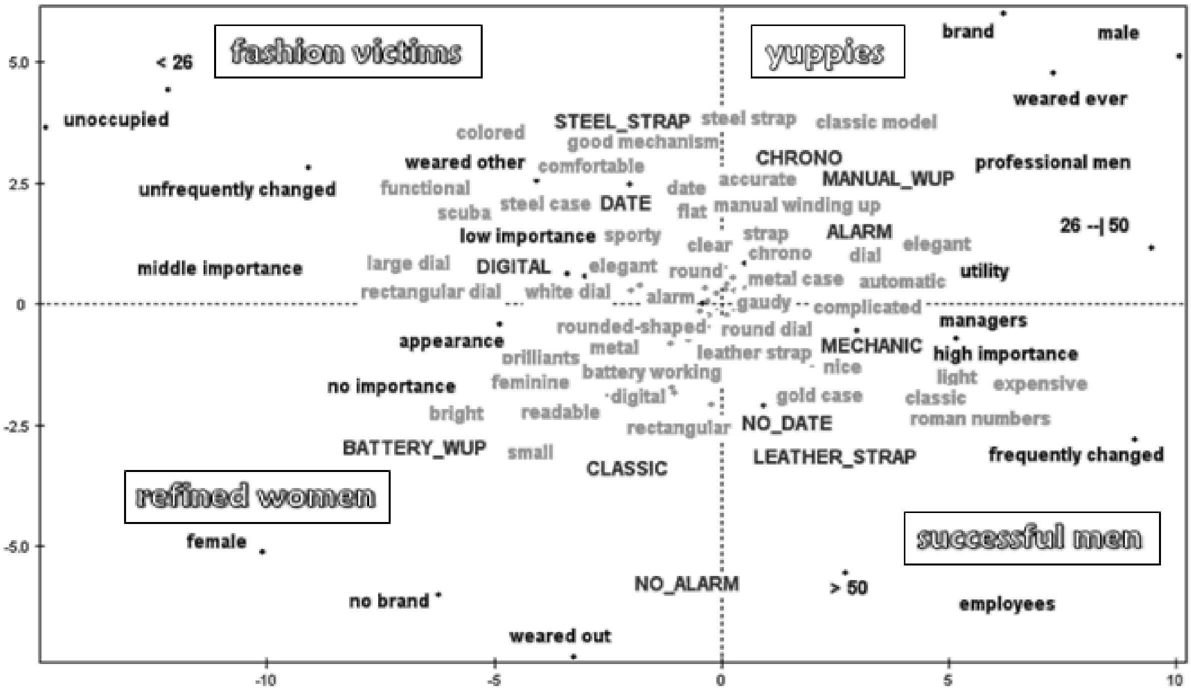

Figure 4 – Factorial representations of the market segmentation

In Figure 4 words of the verbal descriptions are represented on the first factorial plane. If we project the supplementary points (socio-demographic variables) we obtain a powerful market segmentation tool. In particular we visualize in small black letters the categories of the socio-demographic variables, in capital black letters the C.A. stimuli and in small grey letters the textual forms describing the ideal watch.

By performing on the factorial axes a clustering procedure we can divide watch market into four different segments:

- *Young informal people*: young people (less than 26) that want a comfortable and sporty digital watch with steel strap and date.
- *Refined women*: women that want a classic watch with battery. This must be small and feminine, almost a jewel.
- *Successful men* adults that frequently change the watch. It must be an elegant mechanic watch with leather strap and Roman numbers and without date. It does not matter if it is expensive!
- *Yuppies*: people wanting a chrono with manual winding up and alarm. They are mainly professionals.

5. Final remarks

The strategy proposed in the present paper allows to verify and improve the segmentation process by means of statistical tools. All survey data, structured (i.e. numerical data) and not structured (i.e. textual data), are jointly analysed. Therefore, the interpretation of classical conjoint analysis results is enriched by the verbal description of the ideal product. Further developments are envisaged, by the introduction of graphical tools used in textual data analysis frame. One research direction is related to the bootstrap convex hulls on term coordinates (Balbi, 1995), in order to deal with disambiguation problems for evaluating the stability of verbal descriptions, e.g. in terms of quasi-synonyms.

An application on real data has shown the potentiality of our proposal. Textual data, considered as external information on the judges, in a factorial analysis strategy, allows to better understand the individual preferences thanks to the use of both designed and observational information.

References

Balbi S. (1995). Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms. In Bolasco S. et al. editors, *Actes des 3es Journées Internationales d'Analyse statistique des Données Textuelles*. CISU, 2, pages 5-12.

Balbi S. and Giordano G. (2001). A Factorial Technique for Analyzing Textual Data with External Information. In Borra S., Rocci R., Vichi M. and Schader M. editors, *Advances in Classification and Data Analysis*. Springer-Verlag, pages 169-176.

Giordano G. and Scepi G. (1999). Different Informative Structures for Quality Design. *Journal of Italian Statistical Society*, 8(2-3), pages 139-149.

Giordano G. (1997) *L'analisi multidimensionale dei dati di preferenza: una strategia esplorativa per la conjoint analysis*. Doctoral Thesis, Università di Napoli Federico II.

Green P. E. and Srinivasan V. (1990). Conjoint analysis in Marketing: new developments with implications for research and practise. *Journal of Marketing*, 54(4), pages 3-19.

Lauro N. C., Giordano G. and Verde R. (1998). A multidimensional approach to conjoint analysis. *Applied Stochastic Models and Data Analysis*, pages 265-274.

Lebart L. (2004). Validité des visualisations de données textuelles. In Purnelle G., Fairon C. and Dister A. editors, *Le poid des mots. JADT2004*. UCL Presses Universitaires de Louvain, 2, pages 708-715.

Lebart L., Morineau A. and Warwick K. M. (1984). *Multivariate Descriptive Statistical Analysis*. Wiley & Sons.

Takane Y. and Shibayama T. (1991). Principal component analysis with external information on both subjects and variables. *Psychometrika*, 56, pages 97-120.