Partial Least Squares (PLS) an alternative to the Quantification Theory Type I (QT1)

Héctor Rene Alvarez * y Humberto Alvarez **
* Universidad Politécnica de Cataluña (UPC). Barcelona, Spain
**Apsoluti de España S.L. Barcelona, Spain

Abstract

In this paper we propose a new method of estimation of Category Scores (CS) for a synthesis model in a kansei study, based in the regression PLS (Partial Least Squares). Partial Least Squares (PLS) regression is a multivariate data analysis technique which can be used to relate several response (Y) variables to several explanatory (X) variables. The method aims to identify the underlying factors, or linear combination of the X variables, which best model the Y dependent variables. PLS can deal efficiently with data sets where there are very many variables that are highly correlated and involving substantial random noise. We compare PLS method with different Quantification Theory Type I (QT1) algorithms referenced in the technical literature. (Hayashi 1952; Tanaka, 1979 and Murai 1983).

Introduction

Ever since Professor Mitsuo Nagamachi developed Kansei Engineering in the middle of the years 70s, many universities and companies have been using this methodology in Kansei Engineering studies.

Engineering Kansei is a methodology that allows translating the feelings and the emotions in concrete product parameters and provides a robust support for the product design future. Kansei Engineering is a methodology for develop new products oriented to the user. Provide procedures and tools to translate the perceptions, tastes and sensations that the consumer communicates about the product, in terms of design elements. With this methodology it is possible to investigate those aspects of surprise and product differentiation that proposes the Noriaki Kano Model (Apsoluti, 2006).

The goal of a Kansei Engineering project is to find the relationships between the product properties and the valuation of "kansei words" obtained through the denominated methodology "Semantic Differential". There are different statistical methods in order to find that relationships. The methods more used or informed by investigators are: General Linear Model (GLM), Hayashi´s Quantification Theory Type 1 (QT1), Neural Networks, Genetic Algorithm, Rough Set Analysis and Logistic Regression (Schütte, 2005)

Quantification Theory Type 1 (QT1)

The Quantification Theory Type 1 (QT1) was developed in the years 50´s by Professor Chikio Hayashi, like a method of quantification of qualitative data. QT1 method allows quantifying the relations that exist of a set of qualitative variables on a variable answer expressed in numerical values. The QT Method covers four methods (QT1, QT2, QT3, QT4) of quantification proposed by Hayashi (1952). These methods widely have been used in diverse applications and almost, all literature has been published in Japanese.

The form since it has defined method QT1 Hayashi (1952) it shows in table 1. The rows are the valuations done for n individuals on R qualitative variables and on the numerical variable Y.

Items	I_1				I_2				I_R		
	c_{11}	c_{12}	....	c_{1k_1}	c_{21}	....	c_{2k_2}	.....	c_{2K}	....	c_{RK_R}
Y_1		✔	....		✔	....		.....	✔	....	
Y_2	✔		....			....	✔	.....	✔	....	
.	.	.	.	.	.	.	.		.	.	.
.	.	.	.	.	.	.	.		.	.	.
.	.	.	.	.	.	.	.	.....	.	.	.
.	.	.	.	.	.	.	.		.	.	.
.	.	.	.	.	.	.	.		.	.	.
Y_i				✔	✔			.....			✔
.	.	.	.	.	.	.	.		.	.	.
.	.	.	.	.	.	.	.		.	.	.
.	.	.	.	.	.	.	.	.....	.	.	.
.	.	.	.	.	.	.	.		.	.	.
.	.	.	.	.	.	.	.		.	.	.
Y_n	✔						✔				✔

Table 1: Form of the Quantification Theory Type 1 according to Hayashi

For the quantification, we define variables $\delta_i(jk)$:

$$\delta_i(jk) = \begin{cases} 1 & : \text{when sample i correspond s to item j in category k} \\ 0 & \text{otherwise.} \end{cases}$$

From the variables $\delta_i(jk)$ it must estimate the variables X_{jk}, and the responses can be expressed that's follows:

$$Y_i = \sum_{j=1}^{R} \sum_{k=1}^{K_r} X_{jk} \delta_i(jk) + e_i \qquad (1)$$

Hayashi establishes that it must estimated the X_{jk}, and his defines as Category Scores (CS), such that it is maximized the correlation coefficient ρ between Y and $X_1+X_2 \cdots +X_R$.

There are several versions of algorithm of estimation of the CS. We compare those of Hayashi (1976), Tanaka (1979) and Murai (1988).

In order to use the QT1 in Kansei Engineering, it is necesary to relate the kansei average valuations done to n stimuli, with R properties, that have the n evaluated stimuli. Schütte (2005) raises that the model can be expressed like a model of linear regression expressed by

$$Y_i = \sum_{j=1}^{R} \sum_{k=1}^{K_r} \beta_{jk} \delta_i(jk) + e_i \qquad (2)$$

Where Y_i is the observed values of dependent variable, the β_{jk} are the parameters of model and the independent variables are $\delta_i(jk)$ (dummy). The estimation method used for Hayashi (1976) and Tanaka (1979) is the maximization the correlation coefficients between the kansei average valuations of stimuli and properties of stimuli. The estimation used for Murai is least square method that which minimizes the squared prediction error ε

$$L = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (3)$$

Regression PLS (Partial Least Squares)

Regression PLS (RPLS) was developed in the decade of 70s by Hermann Wold Wold, 1975, Wold y al. 1984). It is statistical tools specifically design to solve problems of multiple regression ill-conditional, when predictors are highly collinear, the ordinary least squares regression either fails or produces coefficients with high standard errors. These characteristics of the regression PLS had been prooven with real data a simulations (Garthwaite, 1994; Tenenhaus, 1998).

Partial Least Squares regression extends multiple linear regression without imposing the restrictions employed by discriminate analysis, principal components regression, and canonical correlation. In Partial Least Squares regression, prediction functions are represented by factors extracted from the $\boldsymbol{Y'XX'Y}$ matrix. The number of such prediction functions that can be extracted typically will exceed the maximum of the number of Y and X variables.

In short, partial least squares regression is probably the least restrictive of the various multivariate extensions of the multiple linear regression models. This flexibility allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than predictor variables. Furthermore, Partial Least Squares regression can be used as an exploratory analysis tool to select suitable predictor variables and to identify outliers before classical linear regression.

Partial Least Squares regression has been used in various disciplines such as chemistry, economics, medicine, psychology, and pharmaceutical science where predictive linear modeling, especially with a large number of predictors, is necessary. Especially in chemometrics, Partial Least Squares regression has become a standard tool for modeling linear relations between multivariate measurements (de Jong, 1993).

Originally, PLS was presented how a heuristic algorithm, based on algorithm NIPALS for the calculation of eigen vectors, but quickly it was interpreted with a statistical structure (Frank, & Friedman, 1993; Helland, 1990; Hoskuldsson, 1988; Tenenhaus, 1998). The methodology PLS generalizes and combines characteristics of Principal Component Analysis (PCA) and Regression Multiple Analysis (MRA).

The regression PLS applied to Kansei Engineering, it has the purpose to build a linear model $\boldsymbol{Y=XB+E}$,

where ***Y*** is an *n* rows by *m* variables response matrix, the rows corresponds to *n* stimulus and the columns to *m* kansei evaluates (average) in each stimuli, and ***X*** is a matrix where the *p* columns corresponds to *p* variables dummy of each properties of stimulus evaluates, ***B*** is a *p* by *m* regression coefficient matrix, and ***E*** is a noise term for the model which has the same dimensions as ***Y***. Usually, the variables in ***X*** and ***Y*** are centered by subtracting their means and scaled by dividing by their standard deviations (Geladi and Kowalski, 1986).

The matrix $\boldsymbol{X}=(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ it's possible to be written as a bilinear form (Kruskal 1978):

$$\mathbf{X}=\mathbf{TP'}+\mathbf{E}_k = t_1 p_1^{'} + t_2 p_2^{'} + \Lambda + t_k p_k^{'} + \mathbf{E}_k \quad (4)$$

Where $\boldsymbol{T}=[t_1, t_2, \ldots, t_k]$, $\boldsymbol{P}=[p_1, p_2, \ldots, p_k]$. The t_k are called **variables latent or scores** and the p_k are called the **loadings**. E_k is the residuals matrix and ***K*** is the number of components PLS.

We can assume the follow:

$$\begin{aligned} y &= X\beta \\ &= \mathbf{TQ}+f_k \\ &= t_1q_1+ t_2q_2+ \ldots + t_kq_k+f_k \end{aligned} \quad (5)$$

Where $\mathbf{Q}=[q_1, q_2, \ldots, q_k]$ and f_k is the residual. Therefore ***X*** and ***y*** are connected by the latent variables ***T***.

The criterion of construction of components PLS is the sequential maximization of covariance between the response ***y*** and ***Xg***, restricted to ***g'X'Xg=0***. The PLS components ***t=Xg*** * are orthogonal, where $g^*=\mathrm{Cov}(\boldsymbol{Xg},\boldsymbol{y})$ with ***g'g***=1. ***K*** is selected so that it is the rank of ***X*** (minimum between the rank of rows and columns) and if ***X*** is a full rank then the estimations PLS of ***b*** are identical to estimations of Ordinary Least Squares (OLS). Nevertheless, the regression PLS is usually applied in cases where *p* is more than *n*, a value of ***K*** least than the rank of ***X*** is frequently used. ***K*** can see how a hiper-parameter that it must are optimize. ***K*** is select by the cross-validation like the number of components PLS such that the sum of the predicted errors is minimized. Cross-validation calculates the predictive ability of potential models to help to determine the appropriate number of components to retain in your model.

The regression PLS like the regression of Principal Components extracts factorial loadings ***T*** that they are calculated by ***T=XW*** for a appropriate matrix of weights ***W***, next it consider the model of linear regression ***Y=TQ+E***, where ***Q*** is a matrix of regression coefficient (loadings) for ***T***, and ***E*** is an error term (noise). Once the loads ***Q*** are calculate, the previous model of regression is equivalent to ***Y=XB+E***, where ***B=WQ***, which can be used like a predictive model of regression.

Regression PLS has an advantage because the estimation process is made by steps in algorithm form. There are two forms of estimation: Regression PLS1 (Regression PLS univariate), this form estimates the parameters of model when ***y*** is a vector *nx1*, and regression PLS2 (Regression PLS multivariate) estimates the parameters when ***Y*** is a matrix *nxm*.

Comparison between QT1 and PLS

For the comparison between the both methods of estimation we are used two examples. For the comparisons we used algorithms written in MATLAB, and for the estimations PLS we used the number components obtains with the cross-validation analysis.

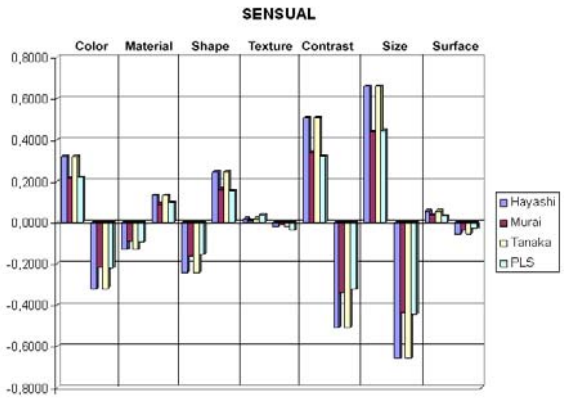


Figure 1: Comparisons of CS with the algorithms of Hayashi, Murai, Tanaka and PLS. Properties Matrix Orthogonal

In the first example, we consider a case when the properties matrix is orthogonal. The case, the properties space has seven items, each one with two characteristics. We observed that in all models the estimations results similar, as it shows the figure 1. Particularly, with the Hayashi's and Tanaka's algorithms, the values are almost equals and the same case with the algorithms of Murai and PLS.

The second example, we consider a case when the properties matrix it is not orthogonal, in this case we take nine stimuli with fourth items and all with equal number of characteristics (table 2).

Estimuli	Color	Shape	Contrast	Material
Bottle 1	Monocolor	Linear	Kill	Glass
Bottle 2	Multicolor	Linear	Brillant	Glass
Bottle 3	Multicolor	Irregular	Brillant	Glass
Bottle 4	Multicolor	Curve	Brillant	Glass
Bottle 5	Monocolor	Linear	Kill	Glass
Bottle 6	Monocolor	Curve	Brillant	Plastic
Bottle 7	Monocolor	Curve	Kill	Plastic
Bottle 8	Monocolor	Irregular	Kill	Glass
Bottle 9	Monocolor	Curve	Kill	Glass

Table 2: Properties matrix non orthogonal

In this case, we observed a pattern non stable. Some models, CS has similar values. An example shows in figure 2.

Other models, Some CS has different values and opposed pattern with the Hayashi's algorithm, no always the CS are equal. And we found one more analysis with same pattern. The figure 3 shows two cases where some of estimations has different signs.

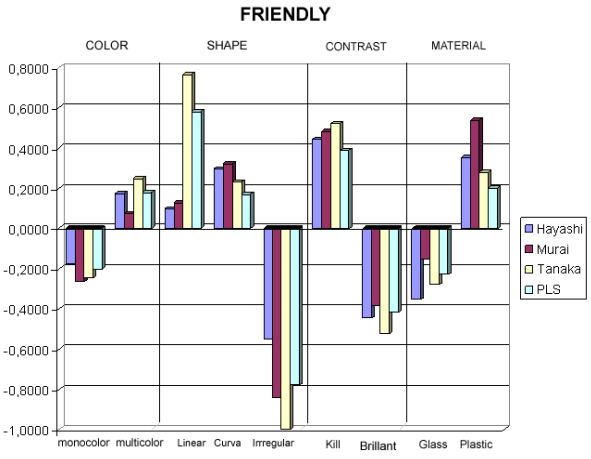
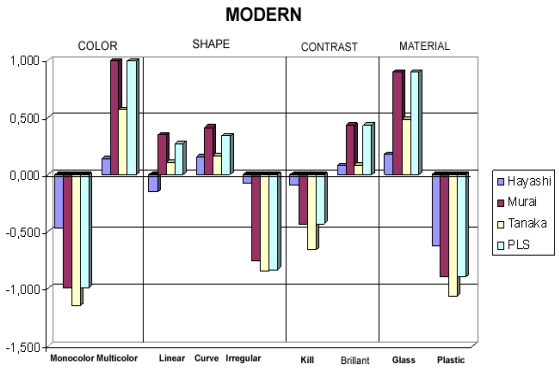


Figure 2: Comparisons of CS with the algorithms of Hayashi, Murai, Tanaka and PLS. Properties Matrix non orthogonal. Similar values.

Of all analyzed cases with properties of non orthogonal matrix, the Murai's and PLS algorithms are stable in all the analysis. For example, in the case of analysis of kansei "modern" we observed that the SC estimated with Hayashi's algorithm are small for the items shape and contrast **y** for the linear shape has opposite sign (figure 3a), while in the analysis of kansei "soft" we observed that SC estimated are small and with opposite sign for curve shape, in addition SC estimate for item "color" has opposite sign (figure 3b).

(a)



(b)

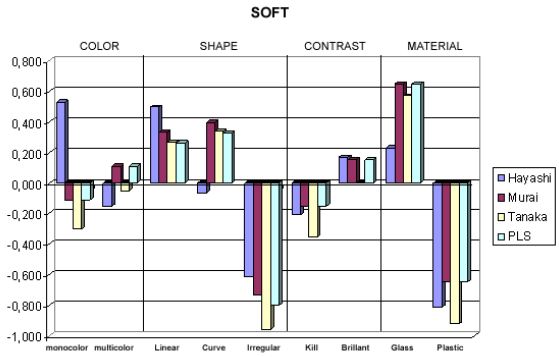


Figure 3: Comparisons of CS with the algorithms of Hayashi, Murai, Tanaka and PLS. Properties Matrix non orthogonal. Non similar values.

Conclusions

We observed that the algorithms of Murai and PLS are those that can be more robust and with similar values. Notice that they are developed using least square method and inverse generalized for that final estimation.

Regression PLS is a good alternative method for estimate the SC in a synthesis model of Kansei Engineering Systems. It's based in a sequential method, that it proof in great deal applications.

The next step in this research, we will compare the PLS results with several algorithms QT1 used in commercial software. The purpose of this second part is, to establish statistics that allows to compare technically these algorithms, and to select most robust.

This work is part of a developed doctoral thesis in the UPC.

References

Apsoluti. (2006) Ingeniería Kansei: Metodología para el desarrollo de productos con alto contenido emocional. Internal report of research.

Frank, I.E., & Friedman, J.H. (1993). A statistical view of chemometrics regression tools. *Technometrics*, 35 pp: 109-148.

Geladi, P., & Kowlaski B. (1986). Partial least square regression: A tutorial. *Analytica Chemica Acta*, 35, pp: 1-17.

Helland I.S. (1990). PLS regression and statistical Models. *Scandivian Journal of Statistics,* 17, pp: 97-114.

Hayashi C. (1950) On the Quantification of Qualitative Data from the Mathematic-Statistical Point of View. *Annals of the Institute of Statistical Mathematics*. Vol. II Nº 1.

Hayashi C. (1952) On the Prediction of phenomena from Qualitative Data and the Quantification of Qualitative Data the Mathematic-Statistical Point of View. *Annals of the Institute of Statistical Mathematics*. Vol. III Nº 1.

Höskuldson, A. (1988). PlS regression methods. *Journal of Chemometrics*, 2, pp: 211-228.

De Jong, S. (1993) PLS fits closer than PCR. *Jourbal of Chemometrics.* Vol. 7, pp. 551-557.

Komazawa, T. and Hayashi, C. (1976). A Statistical Method for Quantification of Categorical Data and its Applications to Medical Science. de Dombal, F. T. and Gremy, F. (ed.), North-Holland Publishing Company, pp.

Murai R. and Kobayashi K. (1989) Method for Generating rules for an Expert Systems for use in controlling a plant. *United States Patent.* Patent Nº 4931951

Schütte. S (2005) Engineering Emotional Values in Product Design: Kansei Engineering in Development. *Dissertation Thesis.* Linköping Universitet, Institute of Technology. Studies in Science and Technology,.

Tanaka Y. (1979) Review of the Methods of Quantification *Environmental Health Perspectives*. Vol 32. pp: 113-123.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P.R. Krishnaiaah (Ed.). Multivariate Analysis. (pp.391-420) New York: Academic Press.

Tenenhaus, M. (1998). La regression PLS. Paris: Technip.