***S**tatistica & **A**pplicazioni*
Vol. III, n. speciale 1, 2005
pp. 77-84

Discriminant Partial Least Squares: The Evaluation of Patient Satisfaction in Sanitary Services

Rosaria Lombardo[§]
Jean François Durand[‡]

Summary. *In this paper we propose an explorative study for the evaluation of Patient Satisfaction in hospital through linear and non linear Discriminant Partial Least Squares. This technique permits to study the dependence relationships between the ordinal variable "satisfaction" in function of variables of different nature, highly correlated, which reveal patient judgements on service quality. To decide on the optimal model dimension we compute the Generalized Cross Validation (GCV), as alternative to the PRESS criterion, well known in literature to check model validity and stability (Cross Validation).*

Parole Chiave: *PLS, PLSS, MAPLSS, Generalized Cross Validation, Patient Satisfaction.*

1. Introduction

From Customer Relationship Management studies, the problem to evaluate the Patient Satisfaction (PS) in sanitary services, in function of a set of predictors (tangible aspects, operator professionalism, organizations, information/communications, etc.) can be faced by linear or non-linear Discriminant Partial Least Squares (PLS, Wold, 1966; Tenenhaus, 1998; Durand, 2001; Durand and Lombardo, 2003; Lombardo *et al.*, 2006).
At the beginning one assumes that there exists an underlying linear relationship between response and predictor variables, but sometimes there is reason to doubt this assumption and a non-linear transformation of the

[§] Dipartimento di Strategie Aziendali e Metodologie Quantitative – Seconda Università degli Studi di Napoli – corso Gran Priorato di Malta, 81043 Capua (CE) Italia (e-mail: rosaria.lombardo@unina2.it).
[‡] University of MontpelierII, Montpellier, France (e-mail: jfd@helios.ensam.inra.fr). Download software at the address: http://jf.durand.club.fr/index.html.

variables might be useful to reveal the model underlying the data. So that non-linear relationships (PLS via Splines, i.e. PLSS; Durand, 2001) and interactions among predictors (Multivariate Additive PLSS, i.e. MAPLSS; Durand and Lombardo, 2003; Lombardo *et al.*, 2006) could drive to the choice of the model. To evaluate regression models of different complexities an heuristic strategy, consisting in increasing progressively the model parameters (degree and knot number) can be considered, as soon as it often tends to work quite well giving a visually pleasing fit to a set of data.
In the context of linear and non-linear PLS, in order to decide on the optimal space dimension the less expensive *GCV* criterion (Friedman, 1991; Lombardo *et al.*, 2006) is used as alternative to the *PRESS* criterion related to the well known cross-validation (*CV*) procedure.
The paper results structured so that, in the second section, the PLS, PLSS and MAPLSS models are briefly reviewed. Details of the *GCV* criterion are illustrated in section 3. In section 4 an application on a real data-set highlights the capabilities of the techniques in PS evaluation problems.

2. Discriminant Partial Least Squares regression models and some extensions towards multivariate additive models

In presence of a low-ratio of observations to variables and in case of multicollinearity in the predictors, a natural extension of the multiple linear regression is PLS regression model. It has been promoted in the chemometrics literature as an alternative to ordinary least squares (OLS) in the poorly or ill-conditioned problems. When the response variable is categorical, coded in a disjunctive form, we refer to the technique as Discriminant PLS (Tenenhaus, 1998).
Let $\mathbf{Y} = [\mathbf{Y}^1 | ... | \mathbf{Y}^q]$ be the categorical *n,q* response matrix, coded in disjunctive form, a column expressing a response category, and $\mathbf{X} = [\mathbf{X}^1 | ... | \mathbf{X}^p]$ the *n,p* matrix of the predictors observed on the same *n* statistical units.
The PLS regression can be viewed as a projection of response variables $\mathbf{Y} = [\mathbf{Y}^1 | ... | \mathbf{Y}^q]$ on latent structures (component scores from predictor matrix $\mathbf{X} = [\mathbf{X}^1 | ... | \mathbf{X}^p]$) that are of maximum covariance with $\mathbf{Y}$. It constructs a sequence of centered and uncorrelated exploratory variables, i.e. the PLS components $\mathbf{t}^1, ..., \mathbf{t}^A$ (Tenenhaus, 1998; Durand, 2001), in 3 steps:

- Set $\mathbf{E_0} = \mathbf{X}$ and $\mathbf{F_0} = \mathbf{Y}$ the design and response data matrices, respectively.

- Define $\mathbf{t}_k = \mathbf{E}_{k-1}\mathbf{w}_k$ and $\mathbf{u}_k = \mathbf{F}_{k-1}\mathbf{c}_k$, where the vectors $\mathbf{w}_k$ and $\mathbf{c}_k$ of unit length are computed by maximizing the covariance between explanatory and response variables $Max[\text{cov}(\mathbf{t}_k, \mathbf{u}_k)]$.
- Update $\mathbf{E}_k$ and $\mathbf{F}_k$ using the orthogonal projection operator $\mathbf{P}_{\mathbf{t}_k} = \mathbf{t}_k\mathbf{t'}_k/(\mathbf{t'}_k\mathbf{t}_k)$ on the component $\mathbf{t}_k$, as follows:

$$\mathbf{E}_k = \mathbf{E}_{k-1} - \mathbf{P}_{\mathbf{t}_k}\mathbf{E}_{k-1}$$
$$\mathbf{F}_k = \mathbf{F}_{k-1} - \mathbf{P}_{\mathbf{t}_k}\mathbf{F}_{k-1}.$$

Because the *A* components are linear compromises of the original predictors **X**, we can write the linear PLS model for the response *j* as the following expression

$$\hat{y}_A^j = f(\mathbf{X}) = \sum_{i=1}^{p} \hat{\beta}_i^j(A)x_i$$

where $\hat{\beta}_i^j(A)$ is the regression coefficient of the predictor x_i on the fitted response y^j, which depends on *A*. In the non-linear context of PLS via Spline (PLSS; Durand, 2001), the **X** design has been replaced by the centered supercoding matrix **B** obtained by transforming the predictors through a basis of B-spline functions. PLSS is defined as the usual linear PLS regression of **Y** onto the space spanned by the centered coding matrix **B**. Given the centered main effects coding matrix we get PLSS(**X,Y**) = PLS(**B,Y**) where the latent variables from **B** are of maximum covariance with **Y**. The *tuning parameters* are given by the number *A* of components and the nature of the spline space (the degree of the polynomials and the number and locations of the knots). The PLSS regression, like PLS, is:

- efficient in spite of low ratio of observations on column dimension of **B**
- efficient in the multi-collinear context for predictors (concurvity)
- robust against extreme values of predictors (local polynomials).

Furthermore, using spline functions, the PLSS permits:

- to treat with predictor variables of different nature
- to evaluate non-linear relationships.

The non-linear PLSS model for the generic response can be so expressed

$$\hat{y}_A^j = f(\mathbf{X}) = \sum_{i=1}^{p} s_i[x_i, \hat{\beta}_i^j(A)]$$

where $s_i[x_i, \hat{\beta}_i^j(A)]$ is the spline function expressing the main effect of x_i on the fitted response, also called the coordinate function of the predictor. The PLSS generalization towards multivariate predictors, called Multivariate Additive PLSS, i.e. MAPLSS (Lombardo *et al*., 2006) enhances the variable

predictive power of PLSS including interaction terms. The MAPLSS model allows:

- to include interesting interactions as predictors.

This kind of decomposition is particularly suitable to investigate in predictive models and identify the particular variables that enter into the model, whether they enter purely additively or are involved in interactions with other variables.
The use of interaction terms means looking for functions of two or more variables. The interaction degree depends on the number of variables involved in the analysis. We define the design matrix **B** from univariate and multivariate B-splines (tensor product of two or more functions). The MAPLSS model for the response *j* has been presented by using the ANOVA decomposition (Lombardo *et al.*, 2006).

$$\hat{y}_A^j = f(\mathbf{X}) = \sum_{i \in K_1} s[x_i; \hat{\beta}_i^j(A)] + \sum_{(k,l) \in K_2} s[x_k, x_l; \hat{\beta}_{k,l}^j(A)] + ...$$

where K_1 and K_2 are index sets pointing out the main effects and the bivariate interactions, respectively. The higher interaction order in MAPLSS implies dimension expansion of design matrix. The risk of overfitting related to an increasing column dimension for the new design matrix **B** is well supported by MAPLSS like the standard PLS method.

3. The dimensionality problem: PRESS and GCV criteria

In literature, the usual approach for choosing the best dimension *A* is based on cross-validation, which implies high computational costs. It works by leaving points (x_i, y_i) out one at a time and estimating the model on the remaining $(n - 1)$ points. The study of model stability through the *PRESS* criterion (residual sum of squares of reduced model) computed for $(n - 1)$ times, has been loosely justified by the fact that the expected value of this statistics is approximately equal to the prediction sum of squares (*PSE*): $E\{PRESS(A)\} \approx PSE(A)$.
A more direct way of constructing an estimate of *PSE* is to correct the average squared residuals (*ASR*), which leads to the C_p statistic strictly related to an other index the *Generalized Cross Validation* criterion (*GCV*; Hastie and Tibshirani, 1990).
In literature (Friedman, 1991) a modified form of the *GCV* criterion has been presented as a suitable alternative to the *PRESS*. Promoting dimension reduction by *GCV* in PLS, but principally in PLSS and MAPLSS, could represent a parsimonious and convenient alternative.

***S**tatistica & **A**pplicazioni*
Vol. III, n. speciale 1, 2005
pp. 77-84

The *GCV* statistic computed in the linear regression case is the same in PLSS and MAPLSS as well as in the usual PLS, because in all these cases regressions are made on components. The *GCV* statistic takes the form

$$GCV(\alpha, A) = \frac{ASR(A)}{\left(1-\alpha\frac{A}{n}\right)^2}$$

where ASR(*A*) is the average squared residuals and α represents a penalty constant to be fixed (Lombardo *et al.*, 2006). The *GCV* criterion depends on the selected initial space-dimension *A*, on α and does not depend on samples. It is interesting to observe that setting $\alpha = 1$, through the approximation $(1-x)^{-2} \cong 1+2x$ where $x = (A/n)\alpha$, leads to the C_p statistic (Hastie and Tibshirani, 1990), originally proposed as a covariate-selection criterion for linear regression models. Without inferential theory, a campaign of few experiments is usually processed to calibrate α, that is, to find α making the *GCV* value as close as possible to the *PRESS*. The building-model stage consists of finding a balance between "thriftiness" of model (evaluated by *GCV*) and "goodness" (of fit and prediction). In order to evaluate the goodness-of-fit, we look at the criterion $R^2(A)$, that is the proportion of the total **Y** variance accounted for by the components $\mathbf{t}^1,...,\mathbf{t}^A$:

$$R^2(A) = \frac{1}{q}\sum_{j=1}^{q} R^2(\mathrm{y}^j, \hat{y}_A^j) = \frac{1}{q}\sum_{j=1}^{q} R^2[\mathrm{y}^j, span(\mathbf{t}^1,...,\mathbf{t}^A)]$$

which is an increasing function of *A*.

4. An application: evaluation of Patient satisfaction in sanitary service

To illustrate the usefulness of linear and non-linear discriminant PLS we consider the data set resulting from a survey on sanitary service of the Second University of Naples (October 2004). The questionnaire instruments considered to investigate on the patient satisfaction (PS) in an hospital of south Italy (Aversa, CE, Italy) is the SERVPERF (Parasuraman *et al.*, 1985) adapted at the health care services (Babakus and Mangold, 1992). Let be the response matrix coded in disjunctive form. It collects the degree of satisfaction on the perceived experience in the hospital, in an ordered scale: from 1 (optimum level of satisfaction) to 5 (very bad, dissatisfaction), in function of 15 ordinal predictors, whose natural scale is also coded by integers from 1 to 5. The predictors observed on 235 individuals are in relation with the five service quality dimensions indicated by Babakus and Mangold (1992): 1) Tangibility (**T1**, **T2**, **T3**); 2) Reliability (**R4**, **R5**, **R6**); 3)

***S**tatistica & **A**pplicazioni*
Vol. III, n. speciale 1, 2005
pp. 77-84

Response capacity (**C7**, **C8**, **C9**); 4) Assure capacity (**A10**, **A11**, **A12**, **A13**); 5) Empathy (**E14**, **E15**). The aim is to predict the patient dissatisfaction (PS degree equal to 5) given the judgement of patients on the previous service quality dimensions (15 items). Performing usual linear PLS, we read in Table 1 the *PRESS* and *GCV* values, the dimension retained is 1. *PRESS*(0.1,1) denotes the *PRESS* with 10 percent of the observations out and $A=1$. Clearly the *GCV* criterion, is a surrogate of *PRESS*, their similar values can make reliable analysis results, difference in values of *GCV* can be due to α that should be properly determined. The goodness of fit (R^2 values), retaining 1 component as indicated by *GCV* and *PRESS* criteria, is low 10.98%. The *reconstituted variance* according to component is only 63.7%.

Table 1: First experience with Linear PLS

GCV(3,**1**) = 4.57 *PRESS*(0.1,**1**) = 4.51		
dimension	Y var.	cum. R^2 %
1	0.5490	10.98
2	0.0882	12.74

Passing to PLSS models (Table 2 and 3) we increase the goodness-of-fit and prediction of the models. In the present context of an opinion poll, the splines used to transform the predictors are local polynomials of degree 0 (piecewise constant functions) with 4 knots at (1.5, 2.5, 3.5, 4.5). Because no variable interaction was accepted in MAPLSS, we present the results of two different multi-response PLSS models. First, with all the 5 satisfaction response levels at hand, Table 2 shows the *PRESS* and *GCV* values, the dimension retained is 2. The goodness of fit (R^2 values) performing PLSS is increased to 17.7%. The *reconstituted variance* according to components is 88.6%.

Table 2: second experience with PLSS, degree = 0, knots = 4

GCV(5,**2**) = 4.49 *PRESS*(0.1,**2**) = 4.34		
dimension	Y var.	cum. R^2 %
1	0.5538	11.08
2	0.3324	17.73

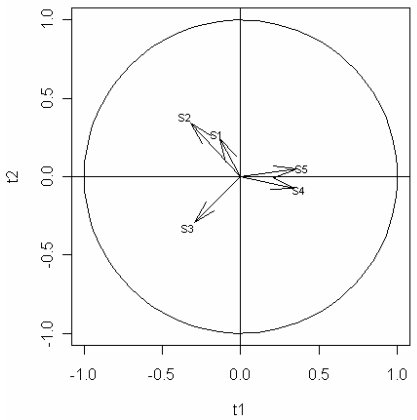

Figure 1: second experience, correlation circle of the responses

***S**tatistica & **A**pplicazioni*
Vol. III, n. speciale 1, 2005
pp. 77-84

Looking at the correlation circle (Figure 1) of the response categories we note that **S1** and **S2**, **S4** and **S5**, are highly correlated so that we synthesize, in the second PLSS experience, the **Y** responses in three ordered categories (**S1** = very satisfied; **S2** = medium satisfied; **S3** = not satisfied) and repeat the analysis improving the goodness of model (Table 3, Figure 2). Reading the *PRESS* and *GCV* values, the dimension retained is 2. The goodness of fit (R^2 values) performing PLSS is increased to 33.3%. For the first two components has been indicated the explained variability. The *reconstituted variance* according to components is now equal to 99.8%.

Table 3: Third experience with PLSS, degree = 0 knots = 4

GCV(8,**2**) = 2.30 *PRESS*(0.1,**2**) = 2.18		
dimension	Y var.	cum. R^2 %
1	0.6605	22.02
2	0.3384	33.30

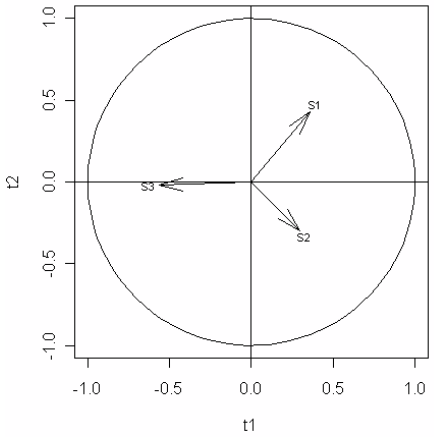

Figure 2: Third experience, correlation circle of the responses

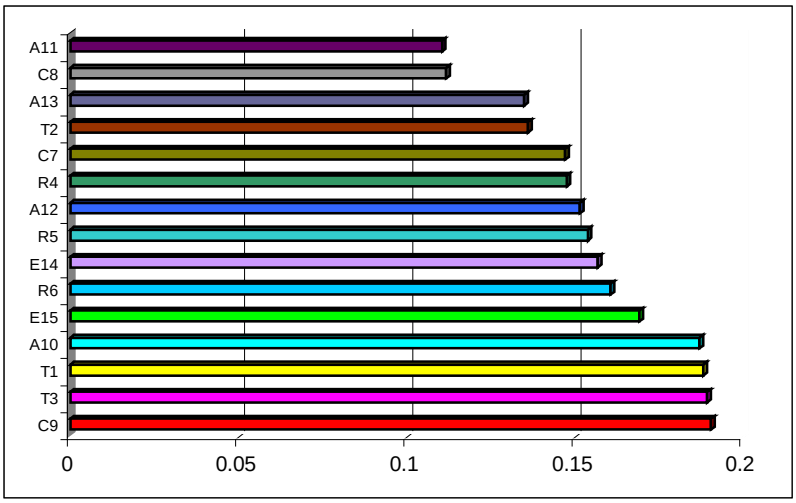

Figure 3: Decreasing influence of the predictors on the dissatisfaction S3

Figure 3 presents the importance of the predictors on the dissatisfaction degree (**S3**) measured by the range of the coordinate functions (Durand, 2001). The first important variable is related with the response capacity dimension **C9**, the second and third ones concerns the tangibility aspects (**T3** and **T1**), etc.. The management should take into account of the importance of variables in predicting the worst degree of satisfaction. In particular increasing attention should be paid at structure aspects (**T3**,**T1**), as well as at improving response and assurance capacities of professional operators (**C9**, **A10**).

5. Conclusion

Linear and non-linear discriminant PLS models have been proposed to study and predict patient satisfaction, particularly suitable in presence of a large number of correlated variables (linear PLS), or when predictors are of mixed nature (PLSS and MAPLSS). An application on a real data set collecting the satisfaction levels of patients in sanitary services in function of a lot variables characterizing the service quality dimensions permits to appreciate the usefulness of the technique as a decision model.

Bibliography

Babakus E., Mangold G. (1992). Adapting the Servqual scale to hospital services: an empirical investigation. *Health Services Research J.*, 767-786.
Durand J.F. (2001). Local Polynomial additive regression through PLS and Splines: PLSS. *In Chemometrics & Intelligent Laboratory Systems*, 58, 235.
Durand J.F., Lombardo R. (2003). Interaction Terms in Non-linear PLS via Additive Spline Transformation. *In Between Data Science and Applied Data Analysis.* (Eds.) Schader P., Gaul J., Vichi M., Springer, 22-30.
Friedman J.H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19, 1: 1-141.
Hastie T.J., Tibshirani R.J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC.
Lombardo R., Durand J.F., De Veaux R. (2006). Multivariate Additive Partial Least Squares via Splines. *Submitted.*
Parasuraman A., Zeithaml V.A., Berry L.L. (1985). A conceptual model of service quality and its implications for future research. *J. Marketing*, v. 49.
Tenenhaus M. (1998). *La Règression PLS, Thèorie et Pratique*. Editions Technip, Paris.
Wold H. (1966). Estimation of principal components and related models by iterative least squares. *In Multivariate Analysis*. (Eds.) P.R. Krishnaiah, New York: Academic Press, 391-420.