Working Paper 06-35
Statistics and Econometrics Series 14
May 2006

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

IMPLEMENTING PLS FOR DISTANCE-BASED REGRESSION: COMPUTATIONAL ISSUES

Eva Boj[(1)], Aurea Grané[(2)], Josep Fortiana[(3)]
and M. Mercè Claramunt[(1)]

Abstract

Distance-based regression allows for a neat implementation of the Partial Least Squares recurrence. In this paper we address practical issues arising when dealing with moderately large datasets ($n \sim 10^4$) such as those typical of automobile insurance premium calculations.

Keywords: Distance-based prediction; PLS regression; Large datasets; Automobile insurance data.

[(1)] Dept. de Matemàtica Econòmica, Financera i Actuarial, Universitat de Barcelona, Avda. Diagonal, 690, 08034 Barcelona, Spain, e-mail evaboj@ub.edu .
[(2)] Dpto. de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain, e-mail: agrane@est-econ.uc3m.es .
[(3)] Dept. de Probabilitat, Lògica i Estadística, Universitat de Barcelona, C/ Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain, e-mail: fortiana@ub.edu .

Work partially supported by Spanish grant MTM2005-02370 (Ministerio de Educación, Ciencia y Tecnología and FEDER).

Implementing PLS for distance-based regression: computational issues

Eva Boj[1], Aurea Grané,[2] Josep Fortiana[3]
and M. Mercè Claramunt[1]

[1]Dept. de Matemàtica Econòmica, Financera i Actuarial, Fac. de Ciències Econòmiques i Empresarials, Univ. de Barcelona, Av. Diagonal, 690, 08034 Barcelona, Spain.
[2]Dept. de Estadística, Univ. Carlos III, Calle Madrid, 126, 28903 Getafe (Madrid), Spain.
[3]Dept. de Probabilitat, Lògica i Estadística, Fac. de Matemàtiques, Univ. de Barcelona, Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain.

Summary

Distance-based regression allows for a neat implementation of the Partial Least Squares recurrence. In this paper we address practical issues arising when dealing with moderately large datasets ($n \sim 10^4$) such as those typical of automobile insurance premium calculations.

Keywords: Distance-based prediction; PLS regression; Large datasets; Automobile insurance data.

1 Introduction

Distance-Based Regression (DBR) (see Cuadras 1989, Cuadras & Arenas 1990, Cuadras et al. 1996) is a method for predicting a numerical response y from a set $\boldsymbol{z}$ of both numerical and categorical predictors. The name of the procedure originates in the fact that it involves a metric in the space of predictors, $d(\cdot\,,\cdot)$, which must be Euclidean in the sense of Multidimensional Scaling (see Section 2).

In this paper we adapt PLS regression to the DBR context, with an emphasis on computational issues arising in treating moderately large ($n \leq 10^5$) datasets. Since DBR for such datasets entails a linear regression on a large number of predictors, it seems particularly adequate for PLS. It is so indeed, but huge matrices appearing at intermediate phases impose a careful out-of-core treatment.

A motivation for this study is the analysis of automobile insurance data, more precisely in rate-making, i.e., predicting total claim amounts from a set of *a priori* risk factors, whose results will be used to determine risk premia for new policy holders (Boj et al. 2004, 2005). Such data usually consist of a number of observations ranging from a moderately large to a very large size ($> 10^6$). In this paper we will concentrate on the former, which can be directly input to DBR, whereas the very large case requires a modified approach with additional processing, such as subsampling or stratification.

The paper is structured as follows: in Section 2 we outline the main characteristics of Distance Based Regression. In Section 3 we derive the DBR version of the PLS recurrence and in Section 4 we give some details on its implementation. As an illustration, in Section 5 we apply the method to an example with real data.

2 Distance Based Regression

The DBR procedure is as follows: Given n observed pairs $\{(y_i, \boldsymbol{z}_i), 1 \leq i \leq n\}$, we compute the matrix $\boldsymbol{D}$, with entries $\left(d^2(\boldsymbol{z}_i, \boldsymbol{z}_j)\right)$, and the doubly-centered *inner products matrix*,

$$\boldsymbol{G} = -\frac{1}{2}\,\boldsymbol{J} \cdot \boldsymbol{D} \cdot \boldsymbol{J},$$

where $\boldsymbol{J} = \boldsymbol{I} - \boldsymbol{1} \cdot \boldsymbol{1}'/n$ is the $n \times n$ centering matrix. The Euclidean requirement is equivalent to the positive semidefiniteness of $\boldsymbol{G}$, hence to the existence of an $\boldsymbol{X}$ such that $\boldsymbol{G} = \boldsymbol{X} \cdot \boldsymbol{X}'$, called in this context a *centered Euclidean configuration* of $\boldsymbol{D}$, meaning that $\boldsymbol{1}' \cdot \boldsymbol{X} = 0$ and that the squared Euclidean interdistances

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$$

between the rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\boldsymbol{X}$ coincide with the corresponding entries in $\boldsymbol{D}$. The verification of this equivalence involves some simple algebra for which we refer to any standard textbook on Multidimensional Scaling, such as Borg and Groenen (2005).

DBR of $\boldsymbol{y} = (y_i)$ on the predictor distance matrix $\boldsymbol{D}$ is defined as a Least Squares regression of $\boldsymbol{y}$ on a Euclidean configuration $\boldsymbol{X}$ of $\boldsymbol{D}$. The rationale supporting this definition is that DBR contains Ordinary Least Squares regression as a particular case. Specifically, with the Pythagorean metric (ℓ^2) on numerical $\boldsymbol{z}$ predictors DBR reduces to OLS.

For a given $\boldsymbol{G}$ there are many Euclidean configurations: Applying any nonsingular linear transformation or a translation to one of them generates another. DBR predictions, however, are independent of the choice. The reason is that both the *hat matrix* giving the fitted responses,

$$\widehat{\boldsymbol{y}} = \boldsymbol{H} \cdot \boldsymbol{y}$$

that is, the orthogonal projector

$$\boldsymbol{H} = \boldsymbol{X} \cdot (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}'$$

on the subspace $\mathcal{G} = \langle \boldsymbol{X} \rangle \subset \mathbb{R}^n$ generated by the columns of $\boldsymbol{X}$, and $\mathcal{G}$ itself can be expressed directly in terms of $\boldsymbol{G}$ or $\boldsymbol{D}$, hence they are independent of $\boldsymbol{X}$. It can be checked that

$$\boldsymbol{H} = \boldsymbol{G}^+ \cdot \boldsymbol{G} = \boldsymbol{G} \cdot \boldsymbol{G}^+,$$

where $\boldsymbol{G}^+$ is the Moore-Penrose generalized inverse of $\boldsymbol{G}$.

When implementing DBR a possible method is to actually compute an explicit Euclidean configuration $\boldsymbol{X}$. For instance, Cuadras et al. (1996) use the *Principal Coordinates Euclidean configuration*, $\boldsymbol{X} = \boldsymbol{U} \cdot \Lambda$, obtained from the eigendecomposition $\boldsymbol{G} = \boldsymbol{U} \cdot \Lambda^2 \cdot \boldsymbol{U}'$. This version of DBR amounts to performing a Principal Components Regression (PCR) (see, e.g., Jolliffe 2002). One of the pitfalls of PCR (Cuadras 1993, 1998, Hadi & Ling 1998) is the fact that the first few Principal Axes, with a greater variance, are not necesarily highly correlated with the response $\boldsymbol{y}$. Furthermore, diagonalization of large $n \times n$ matrices presents substantial computational problems.

The choice of a metric is an important component in DBR model-building: it has points of contact with the choice of a link function for a Generalized Linear Model. In principle it is possible to tailor a metric to reflect specific information on predictors and on how their proximity relates to the particular prediction under study. Most of the times, however, it suffices to utilize an omnibus metric function which satisfies the Euclidean condition referred to in the Introduction. One very popular metric for mixtures of numerical continuous, nominal categorical and binary predictor variables is the one

based on Gower's general similarity coefficient, which for two p-dimensional vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ is equal to

$$s_{ij} = \frac{\sum_{h=1}^{p_1} (1 - |z_{ih} - z_{jh}| / R_h) + a + \alpha}{p_1 + (p_2 - d) + p_3}, \tag{1}$$

where $p = p_1 + p_2 + p_3$, p_1 is the number of continuous variables, a and d are the number of positive and negative matches, respectively, for the p_2 binary variables, and α is the number of matches for the p_3 multi-state categorical variables. R_h is the range of the h-th continuous variable. The squared distance is computed as:

$$d^2(\boldsymbol{z}_i, \boldsymbol{z}_j) = 1 - s_{ij}. \tag{2}$$

It can be proved (Gower 1971) that (2) satisfies the Euclidean condition.

3 PLS for DB regression

PLS regression provides us with a sensible alternative. The standard procedure (see, e.g., Helland 1988, Hoskuldsson 1988) can be described as follows: Given an $n \times p$ centered predictor matrix $\boldsymbol{X}$ and an $n \times 1$ centered response $\boldsymbol{y}$, for each k, $1 \leq k \leq p$, we obtain sets:

$$\begin{aligned}
&\{\boldsymbol{u}_j;\ 0 \leq j \leq k\}, \ \text{ of } n \times 1 \ \text{ vectors},\\
&\{\boldsymbol{E}_j; 0 \leq j \leq k\}, \ \text{ of } n \times p \ \text{ matrices},\\
&\{\boldsymbol{f}_j;\ 1 \leq j \leq k\}, \ \text{ of } n \times 1 \ \text{ orthogonal vectors},\\
&\{\boldsymbol{a}_j;\ 1 \leq j \leq k\}, \ \text{ of } p \times 1 \ \text{ vectors},\\
&\{b_j;\ \ 1 \leq j \leq k\}, \quad \text{of scalars},
\end{aligned}$$

for which the following decompositions hold:

$$\boldsymbol{X} = \boldsymbol{f}_1 \cdot \boldsymbol{a}_1' + \cdots + \boldsymbol{f}_k \cdot \boldsymbol{a}_k' + \boldsymbol{E}_k, \tag{3}$$

$$\boldsymbol{y} = \boldsymbol{f}_1 b_1 + \cdots + \boldsymbol{f}_k b_k + \boldsymbol{u}_k. \tag{4}$$

These sets are generated sequentially, through the following recursion: Let $\boldsymbol{E}_0 \equiv \boldsymbol{X}$ and $\boldsymbol{u}_0 \equiv \boldsymbol{y}$. Then, for each $k \geq 1$, $\boldsymbol{a}_k$ and b_k are the regression coefficients

$$\begin{aligned}
\boldsymbol{a}_k &= (\boldsymbol{E}_{k-1}' \cdot \boldsymbol{f}_k)/(\boldsymbol{f}_k' \cdot \boldsymbol{f}_k),\\
b_k &= (\boldsymbol{u}_{k-1}' \cdot \boldsymbol{f}_k)/(\boldsymbol{f}_k' \cdot \boldsymbol{f}_k),
\end{aligned}$$

and $\boldsymbol{f}_k$ is the linear combination

$$\boldsymbol{f}_k = \boldsymbol{E}_{k-1} \cdot \boldsymbol{w}_k, \tag{5}$$

with weights

$$\boldsymbol{w}_k = \boldsymbol{E}'_{k-1} \cdot \boldsymbol{u}_{k-1}. \tag{6}$$

The next recurrence step starts with the residuals

$$\boldsymbol{E}_k = \boldsymbol{E}_{k-1} - \boldsymbol{f}_k \cdot \boldsymbol{a}'_k, \tag{7}$$

$$\boldsymbol{u}_k = \boldsymbol{u}_{k-1} - \boldsymbol{f}_k \, b_k. \tag{8}$$

In the DBR context, $\boldsymbol{X}$ is assumed to be a centered Euclidean configuration of the square distance matrix $\boldsymbol{D}$, but its explicit computation is not required for the PLS recursion, since all the steps are invariant under a change of $\boldsymbol{X}$ and may be performed by operating directly with $\boldsymbol{G}$. Indeed,

$$\boldsymbol{f}_1 = \boldsymbol{X} \cdot \boldsymbol{X}' \cdot \boldsymbol{y} = \boldsymbol{G} \cdot \boldsymbol{y},$$

$$b_1 = \frac{\boldsymbol{y}' \cdot \boldsymbol{f}_1}{\boldsymbol{f}'_1 \cdot \boldsymbol{f}_1} = \frac{\boldsymbol{y}' \cdot \boldsymbol{G} \cdot \boldsymbol{y}}{\boldsymbol{y}' \cdot \boldsymbol{G}^2 \cdot \boldsymbol{y}},$$

$$\widehat{\boldsymbol{y}}_1 = \boldsymbol{f}_1 \, b_1 = \frac{\boldsymbol{G} \cdot \boldsymbol{y} \, (\boldsymbol{y}' \cdot \boldsymbol{G} \cdot \boldsymbol{y})}{\boldsymbol{y}' \cdot \boldsymbol{G}^2 \cdot \boldsymbol{y}},$$

which can be written more compactly in terms of $\boldsymbol{f}_1 = \boldsymbol{G} \cdot \boldsymbol{y}$:

$$\widehat{\boldsymbol{y}}_1 = \left(\frac{\boldsymbol{f}_1 \cdot \boldsymbol{f}'_1}{\|\boldsymbol{f}_1\|^2} \right) \cdot \boldsymbol{y} = \boldsymbol{P}_1 \cdot \boldsymbol{y},$$

where $\boldsymbol{P}_1 = \boldsymbol{f}_1 \cdot \boldsymbol{f}'_1 / \|\boldsymbol{f}_1\|^2$ is the orthogonal projector on the one-dimensional linear space spanned by $\boldsymbol{f}_1$. The residual

$$\widetilde{\boldsymbol{y}}_1 = \boldsymbol{y} - \widehat{\boldsymbol{y}}_1 = \boldsymbol{Q}_1 \cdot \boldsymbol{y},$$

where $\boldsymbol{Q}_1 = \boldsymbol{I} - \boldsymbol{P}_1$ is the complementary orthogonal projector, also depends on $\boldsymbol{X}$ through $\boldsymbol{G}$. Similarly,

$$\boldsymbol{a}_1 = \frac{\boldsymbol{X}' \cdot \boldsymbol{f}_1}{\boldsymbol{f}'_1 \cdot \boldsymbol{f}_1} = \frac{\boldsymbol{X}' \cdot \boldsymbol{G} \cdot \boldsymbol{y}}{\boldsymbol{y}' \cdot \boldsymbol{G}^2 \cdot \boldsymbol{y}} = \frac{\boldsymbol{X}' \cdot \boldsymbol{f}_1}{\|\boldsymbol{f}_1\|^2},$$

$$\widehat{\boldsymbol{X}}_1 = \boldsymbol{f}_1 \cdot \boldsymbol{a}'_1 = \boldsymbol{P}_1 \cdot \boldsymbol{X},$$

$$\widetilde{\boldsymbol{X}}_1 = \boldsymbol{X} - \widehat{\boldsymbol{X}}_1 = \boldsymbol{Q}_1 \cdot \boldsymbol{X}.$$

Thus we see that it is possible to define the following recursive procedure: Given an $n \times n$ doubly centered positive semidefinite matrix $\boldsymbol{G}$ and an $n \times 1$ centered response $\boldsymbol{y}$, for each k, $1 \leq k \leq \operatorname{rank}(\boldsymbol{G})$, we obtain sets:

$\{\boldsymbol{u}_j; 0 \le j \le k\}$,	of $n \times 1$ vectors,
$\{\boldsymbol{G}_j; 0 \le j \le k\}$,	of $n \times n$ doubly centered positive semidefinite matrices,
$\{\boldsymbol{f}_j; 1 \le j \le k\}$,	of $n \times 1$ orthogonal vectors,
$\{(\boldsymbol{P}_j, \boldsymbol{Q}_j); 1 \le j \le k\}$,	of pairs of complementary $n \times n$ orthogonal projectors,
$\{b_j; 1 \le j \le k\}$,	of scalars,

starting with $\boldsymbol{u}_0 \equiv \boldsymbol{y}$ and $\boldsymbol{G}_0 \equiv \boldsymbol{G}$. The recursion, for $k \ge 1$, follows by:

$$\begin{aligned}
\boldsymbol{f}_k &= \boldsymbol{G}_{k-1} \cdot \boldsymbol{u}_{k-1}, & b_k &= \frac{1}{\|\boldsymbol{f}_k\|^2} \boldsymbol{f}_k' \cdot \boldsymbol{u}_{k-1}, \\
\boldsymbol{P}_k &= \frac{1}{\|\boldsymbol{f}_k\|^2} \boldsymbol{f}_k \cdot \boldsymbol{f}_k', & \boldsymbol{Q}_k &= \boldsymbol{I} - \boldsymbol{P}_k, \\
\boldsymbol{u}_k &= \boldsymbol{u}_{k-1} - \boldsymbol{f}_k\, b_k = \boldsymbol{Q}_k \cdot \boldsymbol{u}_{k-1}, \\
\boldsymbol{G}_k &= \boldsymbol{Q}_k \cdot \boldsymbol{G}_{k-1} \cdot \boldsymbol{Q}_k.
\end{aligned}$$

showing that each recurrence step, hence the whole sequence of computations, involves only the distances $\boldsymbol{D}$. The fitted response vector at step k is given by the sum

$$\widehat{\boldsymbol{y}}_k = \boldsymbol{f}_1\, b_1 + \cdots + \boldsymbol{f}_k\, b_k.$$

It is worth noting that, for $1 \le j \le k$, $\boldsymbol{f}_j\, b_j = \boldsymbol{P}_j \cdot \boldsymbol{u}_{j-1}$ and, by recurrence,

$$\boldsymbol{f}_j\, b_j = \boldsymbol{P}_j \cdot \boldsymbol{y},$$

hence the k-th hat matrix, i.e., the orthogonal projector $\boldsymbol{H}_k$ such that $\widehat{\boldsymbol{y}}_k = \boldsymbol{H}_k \cdot \boldsymbol{y}$, is given by the sum of mutually orthogonal orthogonal projectors:

$$\boldsymbol{H}_k = \boldsymbol{P}_1 + \cdots + \boldsymbol{P}_k.$$

4 Implementation

Very large datasets, with a number of observations greater than 10^5, require specific treatments with the goal of reducing the computational effort to a feasible range. Such an endeavor would involve a displacement of the theoretical emphasis to designing and analyzing suitable subsampling strategies, a problem which will not be treated here. Instead, our current targets are those datasets referred to in the Introduction as *moderately large*, consisting of $n \sim 10^4$ observations. Even for this size, both the distance matrix $\boldsymbol{D}$ and the inner products matrix $\boldsymbol{G}$ are too large to fit comfortably into the main

memory of a standard computer. For instance, if $n = 10^4$, double precision storage requires about 0.8 GB. The PLS-DB scheme outlined above can be easily adapted to this situation. To this end, we implement the following pieces of software:

1. `calcd:` Computes $\boldsymbol{D}$ from the observed predictors $\boldsymbol{z}_i$ and writes it sequentially (one row at a time) to the disk. Several matrices along the computations will be stored out-of-core in this way and henceforth we name them *big* matrices. We are not taking advantage of the symmetry of $\boldsymbol{D}$ for storage since the resulting code complexity and the increased access time override the intended saving.

2. `bigproduct:` An auxiliary function for multiplying an $n \times n$ big matrix, $\boldsymbol{M}$, times an $n \times 1$ in-core vector, resulting another $n \times 1$ in-core vector.

3. `bigproject:` An auxiliary function to compute

$$(\boldsymbol{I} - \boldsymbol{v} \cdot \boldsymbol{v}') \cdot \boldsymbol{M} \cdot (\boldsymbol{I} - \boldsymbol{v} \cdot \boldsymbol{v}')$$

   for an $n \times 1$ in-core unit vector $\boldsymbol{v}$ and an $n \times n$ big matrix $\boldsymbol{M}$, resulting another big matrix, the orthogonal projection of $\boldsymbol{M}$ onto the hyperplane $\langle \boldsymbol{v} \rangle^{\perp}$, the orthogonal complement of $\boldsymbol{v}$. We need `bigproject` both to obtain $\boldsymbol{G}$ from $\boldsymbol{D}$ (with $\boldsymbol{v} = \boldsymbol{1}/\sqrt{n}$) and in each PLS step, to obtain the residual.

4. `PLSstep:` From an $n \times 1$ in-core $\widetilde{\boldsymbol{y}}$ and a big $\widetilde{\boldsymbol{G}}$, computes the new $(\widehat{\boldsymbol{y}}, \widetilde{\boldsymbol{y}})$ and $\widetilde{\boldsymbol{G}}$.

A main function controls the sequence and measures convergence. The PLS-DBR procedure can integrate formal stopping rules for the iterative algorithm, by adapting usual devices such as the Akaike or the Bayes Information Criteria or a Crossvalidation Statistic, which can be implemented in a straightforward manner. As a final observation, it is not necessary to allocate disk space for several Gigabyte-sized matrices. Three of them will suffice, one for $\boldsymbol{D}$ and two more to flip-flop between the two $\widetilde{\boldsymbol{G}}$'s of consecutive steps.

A package with the set of MATLAB functions implementing DBR-PLS is available from the authors upon request.

5 An illustrative case in insurance

We consider the problem of rate-making in automobile insurance, that is, prediction of *total claim amounts* from a set of *a priori* risk factors (for terminology and context see Boj et al. 2004). The results are used to determine

Table 1: Set of predictors for the insurance dataset

- *Continuous predictors:*
  - Power of the vehicle (in HP)
  - Vehicle age
  - Price (Original list price)
  - Age of the main driver
  - Driving license age
- *Categorical predictors:*
  - Sex of the main driver
  - Geographical zone (10 levels)

risk premia for new policy holders. As an illustration we use a real dataset with $n = 11028$ cases, obtained from a portfolio from a Spanish insurer in the period 1996–1997 corresponding to compulsory civil liability insurance.

We set up a DB regression model, where the response is the total claim amount per policyholder and the set of predictors is of a mixed type, comprising both continuous and categorical predictors (Table 1).

The metric we use is Gower's distance (2), based on the general similarity coefficient (1). In addition to being one of the most popular measures of proximity for mixed data types, as mentioned above, it appears to be ade-

Table 2: R^2 as a function of the number of PLS steps

k	R^2
1	0.7698
2	0.8034
3	0.8214
10	0.8633
20	0.8806
30	0.8864
49	0.8934

Table 3: Comparing PLS and PCR: R^2 as a function of the number of steps

	R^2 (PLS-DBR)	R^2 (PCR-DBR)
1	0.1601	0.0825
2	0.2072	0.1072
3	0.2274	0.1075
4	0.2471	0.1080
9	0.3091	0.1516
24	0.4610	0.2276
49	0.7132	0.2771
74	0.8531	0.2908
99	0.9254	0.3294
149	0.9736	0.3640
199	0.9904	0.4003
642	1.0000	0.7202

quate for the type of actuarial prediction we are currently dealing with (Boj et al. 2002).

Table 2 shows the successive determination coefficients obtained as a function of the number of PLS steps. We observe a quite steep ascent, as it is to be expected for the PLS algorithm. The purpose of this table is to demonstrate the comparatively effortless computation required by PLS-DBR to attain a given predictive quality, neatly defeating other methods that require an explicit Euclidean configuration.

For instance, computation of DBR through the Principal Coordinates Euclidean configuration (PCR-DBR) for datasets in the range of sizes we are dealing with here, is at least impractical, perchance utterly impossible, since it would involve obtaining eigenpairs of a huge out-of-core matrix. We have extracted a relatively small subset ($n = 1000$) from the dataset used above in order to compare both methods. Table 3 shows the results as pairs of R^2 for equal number of steps (number of latent variables taken as predictors) in both methods.

6 Conclusion

We have presented an implementation of the PLS recurrence for DB regression. The fact that it does not need explicit matrix decompositions or eigen-

value extraction is a crucial property that enables the method to handle moderately large datasets, such as those found in automobile insurance prediction.

Acknowledgments

Work supported in part by the Spanish Ministerio de Ciencia y Tecnología and FEDER grant MTM2005-02370.

References

Boj, E., Claramunt, M. M., Fortiana, J. & Vidiella, A. (2002), 'The use of distance-based regression and generalized linear models in the rate-making process'. *Mathematics Preprint Series, Institut de Matemàtica de la Universitat de Barcelona*, Num. 305.

Boj, E., Claramunt, M. M. & Fortiana, J. (2004), *Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación*. Madrid: Fundación MAPFRE Estudios. Cuadernos de la Fundación MAPFRE Estudios, Num. 88

Boj, E., Claramunt, M. M., Fortiana, J. & Vegas, A. (2005), 'Bases de datos y estadísticas del seguro de automóviles en España: influencia en el cálculo de primas', *Estadística Española* **47**, 539–566.

Borg, I. & Groenen, P. J. F. (2005), *Modern Multidimensional Scaling: Theory and Applications. 2nd Edition*, Springer-Verlag, New York.

Cuadras, C. M. (1989), 'Distance Analysis in discrimination and classification using both continuous and categorical variables', *in* Yadolah Dodge (ed.), 'Statistical Data Analysis and Inference', North-Holland Publishing Co., Amsterdam, pp. 459–473.

Cuadras, C. M. (1993), 'Intepreting and inequality in multiple regression', *The American Statistician* **47**, 256–258.

Cuadras, C. M. (1998), 'Some cautionary notes on the use of principal components regression (Revisited)', *The American Statistician* **52**, 371 (Letters to the Editor).

Cuadras, C. M. & Arenas, C. (1990), 'A Distance Based Regression Model for Prediction with Mixed Data', *Communications in Statistics A. Theory and Methods* **19**, 2261–2279.

Cuadras, C. M., Arenas, C. & Fortiana, J. (1996), 'Some computational aspects of a Distance-Based model for Prediction', *Communications in Statistics B. Simulation and Computation* **25**, 593–609.

Gower, J. C. (1971), 'A general coefficient of similarity and some of its properties', *Biometrics* **27**, 857–874.

Hadi, A. S. & Ling, R. F. (1998), 'Some cautionary notes on the use of principal components regression', *The American Statistician* **52**, 15–19.

Helland, I. S. (1988), 'On the structure of partial least squares regression', Communications in Statistics - Simulation and Computation **17**, 581–607.

Hoskuldsson, A. (1988), 'PLS Regression Methods' *Journal of Chemometrics* **2**, 211–228.

Jolliffe, I. T. (2002), *Principal Component Analysis. 2nd Edition*, Springer-Verlag, New York.