# Partial Least Square Regression PLS-Regression

Hervé Abdi[1]

## 1 Overview

PLS regression is a recent technique that generalizes and combines features from principal component analysis and multiple regression. Its goal is to predict or analyze a set of dependent variables from a set of independent variables or predictors. This prediction is achieved by extracting from the predictors a set of orthogonal factors called *latent* variables which have the best predictive power.

PLS regression is particularly useful when we need to predict a set of dependent variables from a (very) large set of independent variables (i.e., predictors). It originated in the social sciences (specifically economy, Herman Wold 1966) but became popular first in chemometrics (*i.e.,* computational chemistry) due in part to Herman's son Svante, (Wold, 2001) and in sensory evaluation (Martens & Naes, 1989). But PLS regression is also becoming a tool of choice in the social sciences as a multivariate technique for non-experimental and experimental data alike (e.g., neuroimaging, see Mcintosh & Lobaugh, 2004; Worsley, 1997). It was first presented

[1]In: Neil Salkind (Ed.) (2007). *Encyclopedia of Measurement and Statistics.* Thousand Oaks (CA): Sage.
Address correspondence to: Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
*E-mail:* herve@utdallas.edu http://www.utd.edu/~herve

as an algorithm akin to the power method (used for computing eigenvectors) but was rapidly interpreted in a statistical framework. (see *e.g.*, Phatak, & de Jong, 1997; Tenenhaus, 1998; Ter Braak & de Jong, 1998).

# 2 Prerequisite notions and notations

The $I$ observations described by $K$ dependent variables are stored in a $I \times K$ matrix denoted $\mathbf{Y}$, the values of $J$ predictors collected on these $I$ observations are collected in the $I \times J$ matrix $\mathbf{X}$.

# 3 Goal of PLS regression: Predict Y from X

The goal of PLS regression is to predict $\mathbf{Y}$ from $\mathbf{X}$ and to describe their common structure. When $\mathbf{Y}$ is a vector and $\mathbf{X}$ is full rank, this goal could be accomplished using ordinary multiple regression. When the number of predictors is large compared to the number of observations, $\mathbf{X}$ is likely to be singular and the regression approach is no longer feasible (i.e., because of multicollinearity). Several approaches have been developed to cope with this problem. One approach is to eliminate some predictors (e.g., using stepwise methods) another one, called principal component regression, is to perform a principal component analysis (PCA) of the $\mathbf{X}$ matrix and then use the principal components (i.e., eigenvectors) of $\mathbf{X}$ as regressors on $\mathbf{Y}$. Technically in PCA, $\mathbf{X}$ is decomposed using its singular value decomposition as

$$\mathbf{X} = \mathbf{S}\mathbf{\Delta}\mathbf{V}^{\mathsf{T}}$$

with:

$$\mathbf{S}^{\mathsf{T}}\mathbf{S} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I},$$

(these are the matriceds of the left and right singular vectors), and $\mathbf{\Delta}$ being a diagonal matrix with the singular values as diagonal elements. The singular vectors are ordered according to their corresponding singular values which correspond to the square root of

the variance of **X** explained by each singular vector. The left singular vectors (i.e., the columns of **S**) are then used to predict **Y** using standard regression because the orthogonality of the singular vectors eliminates the multicolinearity problem. But, the problem of choosing an *optimum* subset of predictors remains. A possible strategy is to keep only a few of the first components. But these components are chosen to explain **X** rather than **Y**, and so, nothing guarantees that the principal components, which "explain" **X**, are relevant for **Y**.

By contrast, PLS regression finds components from **X** that are also relevant for **Y**. Specifically, PLS regression searches for a set of components (called *latent vectors*) that performs a simultaneous decomposition of **X** and **Y** with the constraint that these components explain as much as possible of the *covariance* between **X** and **Y**. This step generalizes PCA. It is followed by a regression step where the decomposition of **X** is used to predict **Y**.

# 4 Simultaneous decomposition of predictors and dependent variables

PLS regression decomposes both **X** and **Y** as a product of a common set of orthogonal factors and a set of specific loadings. So, the independent variables are *decomposed* as $\mathbf{X} = \mathbf{TP}^{\mathsf{T}}$ with $\mathbf{T}^{\mathsf{T}}\mathbf{T} = \mathbf{I}$ with **I** being the identity matrix (some variations of the technique do not require **T** to have unit norms). By analogy with PCA, **T** is called the *score* matrix, and **P** the *loading* matrix (in PLS regression the loadings are not orthogonal). Likewise, **Y** is *estimated* as $\widehat{\mathbf{Y}} = \mathbf{TBC}^{\mathsf{T}}$ where **B** is a diagonal matrix with the "regression weights" as diagonal elements and **C** is the "weight matrix" of the dependent variables (see below for more details on the regression weights and the weight matrix). The columns of **T** are the *latent vectors*. When their number is equal to the rank of **X**, they perform an exact decomposition of **X**. Note, however, that they only *estimate* **Y**. (i.e., in general $\widehat{\mathbf{Y}}$ is not equal to **Y**).

# 5 PLS regression and covariance

The latent vectors could be chosen in a lot of different ways. In fact in the previous formulation, any set of orthogonal vectors spanning the column space of **X** could be used to play the rôle of **T**. In order to specify **T**, additional conditions are required. For PLS regression this amounts to finding two sets of weights **w** and **c** in order to create (respectively) a linear combination of the columns of **X** and **Y** such that their covariance is maximum. Specifically, the goal is to obtain a first pair of vectors $\mathbf{t} = \mathbf{X}\mathbf{w}$ and $\mathbf{u} = \mathbf{Y}\mathbf{c}$ with the constraints that $\mathbf{w}^T\mathbf{w} = 1$, $\mathbf{t}^T\mathbf{t} = 1$ and $\mathbf{t}^T\mathbf{u}$ be maximal. When the first latent vector is found, it is *subtracted* from both **X** and **Y** and the procedure is re-iterated until **X** becomes a null matrix (see the algorithm section for more).

# 6 A PLS regression algorithm

The properties of PLS regression can be analyzed from a sketch of the original algorithm. The first step is to create two matrices: $\mathbf{E} = \mathbf{X}$ and $\mathbf{F} = \mathbf{Y}$. These matrices are then column centered and normalized (i.e., transformed into $Z$-scores). The sum of squares of these matrices are denoted $SS_X$ and $SS_Y$. Before starting the iteration process, the vector **u** is initialized with random values. (in what follows the symbol $\propto$ means "to normalize the result of the operation").

Step 1. $\mathbf{w} \propto \mathbf{E}^T\mathbf{u}$ (estimate **X** weights).

Step 2. $\mathbf{t} \propto \mathbf{E}\mathbf{w}$ (estimate **X** factor scores).

Step 3. $\mathbf{c} \propto \mathbf{F}^T\mathbf{t}$ (estimate **Y** weights).

Step 4. $\mathbf{u} = \mathbf{F}\mathbf{c}$ (estimate **Y** scores).

If **t** has not converged, then go to Step 1, if **t** has converged, then compute the value of $b$ which is used to predict **Y** from **t** as $b = \mathbf{t}^T\mathbf{u}$, and compute the factor loadings for **X** as $\mathbf{p} = \mathbf{E}^T\mathbf{t}$. Now subtract (i.e., partial out) the effect of **t** from both **E** and **F** as follows $\mathbf{E} =$

$\mathbf{E} - \mathbf{t}\mathbf{p}^{\mathsf{T}}$ and $\mathbf{F} = \mathbf{F} - b\mathbf{t}\mathbf{c}^{\mathsf{T}}$. The vectors $\mathbf{t}$, $\mathbf{u}$, $\mathbf{w}$, $\mathbf{c}$, and $\mathbf{p}$ are then stored in the corresponding matrices, and the scalar $b$ is stored as a diagonal element of $\mathbf{B}$. The sum of squares of $\mathbf{X}$ (respectively $\mathbf{Y}$) explained by the latent vector is computed as $\mathbf{p}^{\mathsf{T}}\mathbf{p}$ (respectively $b^2$), and the proportion of variance explained is obtained by dividing the explained sum of squares by the corresponding total sum of squares (i.e., $SS_X$ and $SS_Y$).

If $\mathbf{E}$ is a null matrix, then the whole set of latent vectors has been found, otherwise the procedure can be re-iterated from Step 1 on.

# 7 PLS regression and the singular value decomposition

The iterative algorithm presented above is similar to the power method (for a description, see Abdi, Valentin,& Edelman, 1999) which finds eigenvectors. So PLS regression is likely to be closely related to the eigen- and singular value decompositions, and this is indeed the case. For example, if we start from Step 1 which computes: $\mathbf{w} \propto \mathbf{E}^{\mathsf{T}}\mathbf{u}$, and substitute the rightmost term iteratively, we find the following series of equations: $\mathbf{w} \propto \mathbf{E}^{\mathsf{T}}\mathbf{u} \propto \mathbf{E}^{\mathsf{T}}\mathbf{F}\mathbf{c} \propto \mathbf{E}^{\mathsf{T}}\mathbf{F}\mathbf{F}^{\mathsf{T}}\mathbf{t} \propto \mathbf{E}^{\mathsf{T}}\mathbf{F}\mathbf{F}^{\mathsf{T}}\mathbf{E}\mathbf{w}$. This shows that the first weight vector $\mathbf{w}$ is the first right singular vector of the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{Y}$. Similarly, the first weight vector $\mathbf{c}$ is the left singular vector of $\mathbf{X}^{\mathsf{T}}\mathbf{Y}$. The same argument shows that the first vectors $\mathbf{t}$ and $\mathbf{u}$ are the first eigenvectors of $\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\mathbf{Y}^{\mathsf{T}}$ and $\mathbf{Y}\mathbf{Y}^{\mathsf{T}}\mathbf{X}\mathbf{X}^{\mathsf{T}}$.

# 8 Prediction of the dependent variables

The dependent variables are predicted using the multivariate regression formula as $\widehat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^{\mathsf{T}} = \mathbf{X}\mathbf{B}_{\mathrm{PLS}}$ with $\mathbf{B}_{\mathrm{PLS}} = (\mathbf{P}^{\mathsf{T}+})\mathbf{B}\mathbf{C}^{\mathsf{T}}$ (where $\mathbf{P}^{\mathsf{T}+}$ is the Moore-Penrose pseudo-inverse of $\mathbf{P}^{\mathsf{T}}$). If all the latent variables of $\mathbf{X}$ are used, this regression is equivalent to principal component regression. When only a subset of the latent variables is used, the prediction of $\mathbf{Y}$ is optimal for this number of predictors.

Table 1: The **Y** matrix of dependent variables.

| Wine | Hedonic | Goes with meat | Goes with dessert |
|---|---|---|---|
| 1 | 14 | 7 | 8 |
| 2 | 10 | 7 | 6 |
| 3 | 8 | 5 | 5 |
| 4 | 2 | 4 | 7 |
| 5 | 6 | 2 | 4 |

Table 2: The **X** matrix of predictors.

| Wine | Price | Sugar | Alcohol | Acidity |
|---|---|---|---|---|
| 1 | 7 | 7 | 13 | 7 |
| 2 | 4 | 3 | 14 | 7 |
| 3 | 10 | 5 | 12 | 5 |
| 4 | 16 | 7 | 11 | 3 |
| 5 | 13 | 3 | 10 | 3 |

An obvious question is to find the number of latent variables needed to obtain the best generalization for the prediction of *new* observations. This is, in general, achieved by cross-validation techniques such as bootstrapping.

The interpretation of the latent variables is often helped by examining graphs akin to PCA graphs (e.g., by plotting observations in a $\mathbf{t}_1 \times \mathbf{t}_2$ space, see Figure 1).

# 9 A small example

Table 3: The matrix **T**.

| Wine | $\mathbf{t}_1$ | $\mathbf{t}_2$ | $\mathbf{t}_3$ |
|---|---|---|---|
| 1 | 0.4538 | −0.4662 | 0.5716 |
| 2 | 0.5399 | 0.4940 | −0.4631 |
| 3 | 0 | 0 | 0 |
| 4 | −0.4304 | −0.5327 | −0.5301 |
| 5 | −0.5633 | 0.5049 | 0.4217 |

Table 4: The matrix **U**.

| Wine | $\mathbf{u}_1$ | $\mathbf{u}_2$ | $\mathbf{u}_3$ |
|---|---|---|---|
| 1 | 1.9451 | −0.7611 | 0.6191 |
| 2 | 0.9347 | 0.5305 | −0.5388 |
| 3 | −0.2327 | 0.6084 | 0.0823 |
| 4 | −0.9158 | −1.1575 | −0.6139 |
| 5 | −1.7313 | 0.7797 | 0.4513 |

Table 5: The matrix **P**.

| | $\mathbf{p}_1$ | $\mathbf{p}_2$ | $\mathbf{p}_3$ |
|---|---|---|---|
| Price | −1.8706 | −0.6845 | −0.1796 |
| Sugar | 0.0468 | −1.9977 | 0.0829 |
| Alcohol | 1.9547 | 0.0283 | −0.4224 |
| Acidity | 1.9874 | 0.0556 | 0.2170 |

Table 6: The matrix $\mathbf{W}$.

| | $\mathbf{w}_1$ | $\mathbf{w}_2$ | $\mathbf{w}_3$ |
|---|---|---|---|
| Price | −0.5137 | −0.3379 | −0.3492 |
| Sugar | 0.2010 | −0.9400 | 0.1612 |
| Alcohol | 0.5705 | −0.0188 | −0.8211 |
| Acidity | 0.6085 | 0.0429 | 0.4218 |

Table 7: The matrix $\mathbf{B}_{\text{PLS}}$ when 3 latent vectors are used.

| | Hedonic | Goes with meat | Goes with dessert |
|---|---|---|---|
| Price | −1.0607 | −0.0745 | 0.1250 |
| Sugar | 0.3354 | 0.2593 | 0.7510 |
| Alcohol | −1.4142 | 0.7454 | 0.5000 |
| Acidity | 1.2298 | 0.1650 | 0.1186 |

Table 8: The matrix $\mathbf{B}_{\text{PLS}}$ when 2 latent vectors are used.

| | Hedonic | Goes with meat | Goes with dessert |
|---|---|---|---|
| Price | −0.2662 | −0.2498 | 0.0121 |
| Sugar | 0.0616 | 0.3197 | 0.7900 |
| Alcohol | 0.2969 | 0.3679 | 0.2568 |
| Acidity | 0.3011 | 0.3699 | 0.2506 |

We want to predict the subjective evaluation of a set of 5 wines. The dependent variables that we want to predict for each wine are its likeability, and how well it goes with meat, or dessert (as rated by a panel of experts) (see Table 1). The predictors are the price, the sugar, alcohol, and acidity content of each wine (see Table 2).

The different matrices created by PLS regression are given in Tables 3 to 11. From Table 11 one can find that two latent vec-

Table 9: The matrix **C**.

| | $\mathbf{c}_1$ | $\mathbf{c}_2$ | $\mathbf{c}_3$ |
|---|---|---|---|
| Hedonic | 0.6093 | 0.0518 | 0.9672 |
| Goes with meat | 0.7024 | −0.2684 | −0.2181 |
| Goes with dessert | 0.3680 | −0.9619 | −0.1301 |

Table 10: The **b** vector.

| $b_1$ | $b_2$ | $b_3$ |
|---|---|---|
| 2.7568 | 1.6272 | 1.1191 |

tors explain 98% of the variance of **X** and 85% of **Y**. This suggests to keep these two dimensions for the final solution. The examination of the two-dimensional regression coefficients (i.e., $\mathbf{B}_{\text{PLS}}$, see Table 8) shows that sugar is mainly responsible for choosing a dessert wine, and that price is negatively correlated with the perceived quality of the wine, whereas alcohol is positively correlated with it (at least in this example ...). Looking at the latent vectors shows that $\mathbf{t}_1$ expresses price and $\mathbf{t}_2$ reflects sugar content. This interpretation is confirmed and illustrated in Figures 1*a* and *b* which display in (a) the projections on the latent vectors of the wines (matrix **T**) and the predictors (matrix **W**), and in (b) the correlation between the original dependent variables and the projection of the wines on the latent vectors.

# 10 Relationship with other techniques

PLS regression is obviously related to canonical correlation, STATIS, and to multiple factor analysis. These relationships are explored in details by Tenenhaus (1998), Pagès and Tenenhaus (2001), and Abdi (2003b). The main originality of PLS regression is to preserve the asymmetry of the relationship between predictors and depen-

Table 11: Variance of **X** and **Y** explained by the latent vectors.

| Latent Vector | Percentage of Explained Variance for **X** | Cumulative Percentage of Explained Variance for **X** | Percentage of Explained Variance for **Y** | Cumulative Percentage of Explained Variance for **Y** |
|---|---|---|---|---|
| 1 | 70 | 70 | 63 | 63 |
| 2 | 28 | 98 | 22 | 85 |
| 3 | 2 | 100 | 10 | 95 |

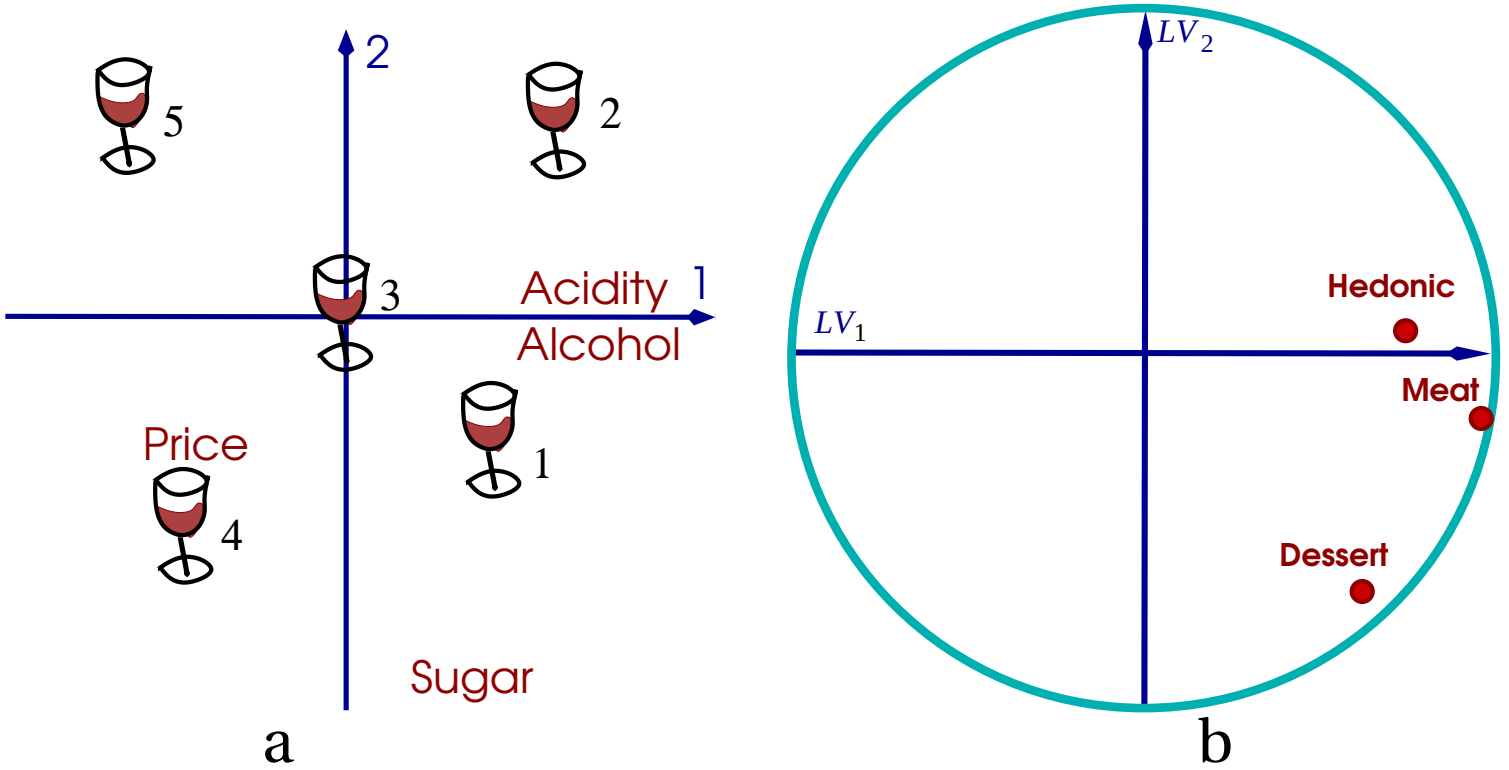


Figure 1: PLS-regression. (a) Projection of the wines and the predictors on the first 2 latent vectors (respectively matrices $\mathbf{T}$ and $\mathbf{W}$). (b) Circle of correlation showing the correlation between the original dependent variables (matrix $\mathbf{Y}$) and the latent vectors (matrix $\mathbf{T}$).

dent variables, whereas these other techniques treat them symmetrically.

# 11 Software

PLS regression necessitates sophisticated computations and therefore its application depends on the availability of software. For chemistry, two main programs are used: the first one called SIMCA-P was developed originally by Wold, the second one called the UNSCRAMBLER was first developed by Martens who was another pioneer in the field. For brain imaging, SPM, which is one of the most widely used programs in this field, has recently (2002) integrated a PLS regression module. Outside these domains, SAS PROC PLS is probably the most easily available program. In addition, interested readers can download a set of MATLAB programs from the author's home page (`www.utdallas.edu/~herve`). Also, a public domain set of MATLAB programs is available from the home page of the *N*-Way project (`www.models.kvl.dk/source/nwaytoolbox/`) along with tutorials and examples. From brain imaging, a special toolbox written in MATLAB (by McIntosh, Chau, Lobaugh, & Chen) is freely available from `www.rotman-baycrest.on.ca:8080`. And finally, a commercial MATLAB toolbox has also been developed by EIGENRESEARCH.

# References

[1] Abdi, H. (2003a&b). PLS-Regression; Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks: Sage.

[2] Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Thousand Oaks (CA): Sage.

[3] Escofier, B., & Pagès, J. (1988). *Analyses factorielles multiples*. Paris: Dunod.

[4] Frank, I.E., & Friedman, J.H. (1993). A statistical view of chemometrics regression tools. *Technometrics*, **35** 109–148.

[5] Helland I.S. (1990). PLS regression and statistical models. *Scandivian Journal of Statistics*, **17**, 97–114.

[6] Höskuldson, A. (1988). PLS regression methods. *Journal of Chemometrics*, **2**, 211-228.

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

[8] McIntosh, A.R., & Lobaugh N.J. (2004). Partial least squares analysis of neuroimaging data: applications and advances. *Neuroimage*, **23**, 250–263.

[9] Martens, H, & Naes, T. (1989). *Multivariate Calibration*. London: Wiley.

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

[11] Phatak, A., & de Jong, S. (1997). The geometry of partial least squares. *Journal of Chemometrics*, **11**, 311–338.

[12] Tenenhaus, M. (1998). *La régression PLS*. Paris: Technip.

[13] Ter Braak, C.J.F., & de Jong, S. (1998). The objective function of partial least squares regression. *Journal of Chemometrics*, **12**, 41–54.

[14] 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.

[15] Wold, S. (2001). Personal memories of the early PLS development. *Chemometrics and Intelligent Laboratory Systems*, **58**, 83–84.

[16] Worsley, K.J. (1997). An overview and some new developments in the statistical analysis of PET and fMRI data. *Human Brain Mapping*, **5**, 254–258.