



Wiley Interdisciplinary Reviews:
Computational Statistics

Partial least squares regression

Journal:	<i>Wiley Interdisciplinary Reviews: Computational Statistics</i>
Manuscript ID:	EOCS-039.R1
Wiley - Manuscript type:	Focus article
Date Submitted by the Author:	08-Jan-2009
Complete List of Authors:	Abdi, Hervé
Keywords:	Partial least squares, Projection to latent structures, Principal component analysis, multicollinearity, Singular value decomposition



view

Partial Least Squares Regression and Projection on Latent Structure Regression (PLS-Regression)

Hervé Abdi

Abstract

Partial least squares (PLS) regression (*a.k.a* projection on latent structures) is a recent technique that combines features from and generalizes principal component analysis (PCA) and multiple linear regression. Its goal is to predict 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. These latent variables can be used to create displays akin to PCA displays. The quality of the prediction obtained from a PLS regression model is evaluated with cross-validation techniques such as the bootstrap and jackknife. There are two main variants of PLS regression: The most common one separates the rôles of independent and independent variables; the second one—used mostly to analyze brain imaging data—gives the same rôles to dependent and independent variables.

Keywords:

Partial least squares, Projection to latent structures, Principal component analysis, Principal component regression, Multiple regression, multicollinearity, NIPALS Eigenvalue decomposition, Singular value decomposition, Bootstrap, Jackknife, small N large P problem.

1 Introduction

PLS regression is an acronym which originally stood for *Partial Least Squares Regression*, but, recently, some authors have preferred to develop this acronym

1
2
3
4
5
6
7
8 as *Projection to Latent Structures*. In any case, PLS regression combines fea-
9 tures from and generalizes principal component analysis and multiple linear
10 regression. Its goal is to analyze or predict a set of dependent variables
11 from a set of independent variables or predictors. This prediction is achieved
12 by extracting from the predictors a set of orthogonal factors called *latent*
13 variables which have the best predictive power.

14
15 PLS regression is particularly useful when we need to predict a set of
16 dependent variables from a (very) large set of independent variables (*i.e.*,
17 predictors). It originated in the social sciences (specifically economy, Herman
18 Wold 1966) but became popular first in chemometrics (*i.e.*, computational
19 chemistry) due in part to Herman's son Svante, (Wold, 2001) and in sensory
20 evaluation (Martens & Naes, 1989). But PLS regression is also becoming
21 a tool of choice in the social sciences as a multivariate technique for non-
22 experimental (*e.g.*, Fornell, Lorange, & Roos, 1990; Hulland, 1999; Graham,
23 Evenko, Rajan, 1992) and experimental data alike (*e.g.*, neuroimaging, see
24 Worsley, 1997; McIntosh & Lobaugh, 2004; Giessing *et al.*, 2007; Kovacevic
25 & McIntosh, 2007; Wang *et al.*, 2008). It was first presented as an algorithm
26 akin to the power method (used for computing eigenvectors) but was rapidly
27 interpreted in a statistical framework. (see *e.g.*, Burnham, 1996; Garthwaite,
28 1994; Höskuldson, 2001; Phatak, & de Jong, 1997; Tenenhaus, 1998; Ter
29 Braak & de Jong, 1998).

30
31 Recent developments, including, extensions to multiple table analysis,
32 are explored in Höskuldson (in press), and in the volume edited by Esposito
33 Vinzi, Chin, Henseler, and Wang (2009).

34 35 36 37 38 39 40 41 **2 Prerequisite notions and notations**

42 The I observations described by K dependent variables are stored in an
43 $I \times K$ matrix denoted \mathbf{Y} , the values of J predictors collected on these I
44 observations are collected in an $I \times J$ matrix \mathbf{X} .

45 46 47 48 49 50 51 **3 Goal of PLS regression: 52 Predict \mathbf{Y} from \mathbf{X}**

53 The goal of PLS regression is to predict \mathbf{Y} from \mathbf{X} and to describe their
54 common structure. When \mathbf{Y} is a vector and \mathbf{X} is a full rank matrix, 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). This data configuration has been recently often called the “*small N large P problem*.” It is characteristic of recent data analysis domains such as, *e.g.*, bio-informatics, brain imaging, chemometrics, data mining, and genomics.

3.1 Principal Component Regression

Several approaches have been developed to cope with the multicollinearity problem. For example, one approach is to eliminate some predictors (*e.g.*, using stepwise methods, see Draper & Smith, 1998), another one is to use ridge regression (Hoerl & Kennard, 1970). One method, closely related to PLS regression is called principal component regression (PCR), it performs a principal component analysis (PCA) of the \mathbf{X} matrix and then use the principal components of \mathbf{X} as the independent variables of a multiple regression model predicting \mathbf{Y} . Technically, in PCA, \mathbf{X} is decomposed using its singular value decomposition (see Abdi 2007a,b for more details) as

$$\mathbf{X} = \mathbf{R}\mathbf{\Delta}\mathbf{V}^T \quad (1)$$

with:

$$\mathbf{R}^T\mathbf{R} = \mathbf{V}^T\mathbf{V} = \mathbf{I}, \quad (2)$$

(where \mathbf{R} and \mathbf{V} are the matrices 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 value which is the square root of the variance (*i.e.*, eigenvalue) of \mathbf{X} explained by the singular vectors. The columns of \mathbf{V} are called the *loadings*. The columns of $\mathbf{G} = \mathbf{R}\mathbf{\Delta}$ are called the *factor scores* or *principal components* of \mathbf{X} , or simply scores or components. The matrix \mathbf{R} of the left singular vectors of \mathbf{X} (or the matrix \mathbf{G} of the principal components) are then used to predict \mathbf{Y} using standard multiple linear regression. This approach works well because the orthogonality of the singular vectors eliminates the multicollinearity 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 were originally chosen to explain \mathbf{X} rather than \mathbf{Y} , and so, nothing guarantees that the principal components, which “explain” \mathbf{X} optimally, will be relevant for the prediction of \mathbf{Y} .

3.2 Simultaneous decomposition of predictors and dependent variables

Principal component regression decomposes \mathbf{X} in order to obtain components which best explains \mathbf{X} . By contrast, PLS regression finds components *from* \mathbf{X} that best predict \mathbf{Y} . Specifically, PLS regression searches for a set of components (called *latent vectors*) that performs a simultaneous decomposition of \mathbf{X} and \mathbf{Y} with the constraint that these components explain as much as possible of the *covariance* between \mathbf{X} and \mathbf{Y} . This step generalizes PCA. It is followed by a regression step where the latent vectors obtained from \mathbf{X} are used to predict \mathbf{Y} .

PLS regression decomposes both \mathbf{X} and \mathbf{Y} as a product of a common set of orthogonal factors and a set of specific loadings. So, the independent variables are *decomposed* as

$$\mathbf{X} = \mathbf{TP}^T \quad \text{with} \quad \mathbf{T}^T \mathbf{T} = \mathbf{I}, \quad (3)$$

with \mathbf{I} being the identity matrix (some variations of the technique do not require \mathbf{T} to have unit norms, these variations differ mostly by the choice of the normalization, they do not differ in their final prediction, but the differences in normalization may make delicate the comparisons between different implementations of the technique). By analogy with PCA, \mathbf{T} is called the *score* matrix, and \mathbf{P} the *loading* matrix (in PLS regression the loadings are not orthogonal). Likewise, \mathbf{Y} is *estimated* as

$$\hat{\mathbf{Y}} = \mathbf{TBC}^T, \quad (4)$$

where \mathbf{B} is a diagonal matrix with the “regression weights” as diagonal elements and \mathbf{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 \mathbf{T} are the *latent vectors*. When their number is equal to the rank of \mathbf{X} , they perform an exact decomposition of \mathbf{X} . Note, however, that the latent vectors provide only an *estimate* of \mathbf{Y} (*i.e.*, in general $\hat{\mathbf{Y}}$ is not equal to \mathbf{Y}).

4 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 \mathbf{X} could be used to play the rôle of \mathbf{T} . In order to specify \mathbf{T} , additional conditions are required. For PLS regression this amounts to finding two sets of weights denoted \mathbf{w} and \mathbf{c} in order to create (respectively) a linear combination of the columns of \mathbf{X} and \mathbf{Y} such that these two linear combinations have maximum covariance. Specifically, the goal is to obtain a first pair of vectors

$$\mathbf{t} = \mathbf{X}\mathbf{w} \text{ and } \mathbf{u} = \mathbf{Y}\mathbf{c} \quad (5)$$

with the constraints that $\mathbf{w}^T\mathbf{w} = 1$, $\mathbf{t}^T\mathbf{t} = 1$ and $\mathbf{t}^T\mathbf{u}$ is maximal. When the first latent vector is found, it is *subtracted* from both \mathbf{X} and \mathbf{Y} and the procedure is re-iterated until \mathbf{X} becomes a null matrix (see the algorithm section for more).

5 NIPALS: A PLS algorithm

The properties of PLS regression can be analyzed from a sketch of the original algorithm (called NIPALS). 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 \mathbf{u} is initialized with random values. The NIPALS algorithm then performs the following steps (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 \mathbf{X} weights).

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

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

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

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

6 PLS regression and the singular value decomposition

The NIPALS algorithm is obviously similar to the power method (for a description, see, *e.g.*, Abdi, Valentin, & Edelman, 1999) which finds eigenvectors. So PLS regression is likely to be closely related to the eigen- and singular value decompositions (see, Abdi, 2007a,b for an introduction to these notions) and this is indeed the case. For example, if we start from Step 1 of the algorithm, which computes: $\mathbf{w} \propto \mathbf{E}^\top \mathbf{u}$, and substitute the rightmost term iteratively, we find the following series of equations:

$$\mathbf{w} \propto \mathbf{E}^\top \mathbf{u} \propto \mathbf{E}^\top \mathbf{F} \mathbf{c} \propto \mathbf{E}^\top \mathbf{F} \mathbf{F}^\top \mathbf{t} \propto \mathbf{E}^\top \mathbf{F} \mathbf{F}^\top \mathbf{E} \mathbf{w} . \quad (6)$$

This shows that the weight vector \mathbf{w} is the first right singular vector of the matrix

$$\mathbf{S} = \mathbf{E}^\top \mathbf{F} . \quad (7)$$

Similarly, the first weight vector \mathbf{c} is the left singular vector of \mathbf{S} . The same argument shows that the first vectors \mathbf{t} and \mathbf{u} are the first eigenvectors of $\mathbf{E} \mathbf{E}^\top \mathbf{F} \mathbf{F}^\top$ and $\mathbf{F} \mathbf{F}^\top \mathbf{E} \mathbf{E}^\top$. This last observation is important from a computational point of view because it shows that the weight vectors can also be obtained from matrices of size I by I (Rännar, Lindgren, Geladi, & Wold, 1994). This is useful when the number of variables is much larger than the number of observations (*e.g.*, as in the “small N , large P problem”).

7 Prediction of the dependent variables

The dependent variables are predicted using the multivariate regression formula as

$$\hat{\mathbf{Y}} = \mathbf{T} \mathbf{B} \mathbf{C}^\top = \mathbf{X} \mathbf{B}_{\text{PLS}} \text{ with } \mathbf{B}_{\text{PLS}} = (\mathbf{P}^{\top+}) \mathbf{B} \mathbf{C}^\top \quad (8)$$

(where \mathbf{P}^{\dagger} is the Moore-Penrose pseudo-inverse of \mathbf{P}^{\top} , see Abdi, 2001). This last equation assumes that both \mathbf{X} and \mathbf{Y} have been standardized prior to the prediction. In order to predict a non-standardized matrix \mathbf{Y} from a non-standardized matrix \mathbf{X} , we use $\mathbf{B}_{\text{PLS}}^*$ which is obtained by re-introducing the original units into \mathbf{B}_{PLS} and adding a first column corresponding to the intercept (when using the original units, \mathbf{X} needs to be augmented with a first columns of 1, as in multiple regression).

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.

The interpretation of the latent variables is often facilitated 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).

8 Statistical inference: Evaluating the quality of the prediction

8.1 Fixed effect model

The quality of the prediction obtained from PLS regression described so far corresponds to a fixed effect model (*i.e.*, the set of observations is considered as the population of interest, and the conclusions of the analysis are restricted to this set). In this case, the analysis is *descriptive* and the amount of variance (of \mathbf{X} and \mathbf{Y}) explained by a latent vector indicates its importance for the set of data under scrutiny. In this context, latent variables are worth considering if their interpretation is meaningful within the research context.

For a fixed effect model, the overall quality of a PLS regression model using L latent variables is evaluated by first computing the predicted matrix of dependent variables denoted $\hat{\mathbf{Y}}^{[L]}$ and then measuring the similarity between $\hat{\mathbf{Y}}^{[L]}$ and \mathbf{Y} . Several coefficients are available for the task. The squared coefficient of correlation is sometimes used as well as its matrix specific cousin the R_V coefficient (Abdi, 2007c). The most popular coefficient, however, is the residual sum of squares, abbreviated as RESS. It is computed as:

$$\text{RESS} = \|\mathbf{Y} - \hat{\mathbf{Y}}^{[L]}\|^2, \quad (9)$$

(where $\|\cdot\|$ is the norm of \mathbf{Y} , *i.e.*, the square root of the sum of squares of the elements of \mathbf{Y}). The smaller the value of RESS, the better the prediction,

1
2
3
4
5
6
7
8 with a value of 0 indicating perfect prediction. For a fixed effect model, the
9 larger L (*i.e.*, the number of latent variables used), the better the prediction.
10

11 8.2 Random effect model

12
13
14 In most applications however, the set of observations is a *sample* from some
15 population of interest. In this context, the goal is to *predict* the value of the
16 dependent variables for *new* observations originating from the same popula-
17 tion as the sample. This corresponds to a *random* model. In this case, the
18 amount of variance explained by a latent variable indicates its importance
19 in the prediction of \mathbf{Y} . In this context, a latent variable is relevant only
20 if it improves the prediction of \mathbf{Y} for new observations. And this, in turn,
21 opens the problem of which and how many latent variables should be kept
22 in the PLS regression model in order to achieve optimal generalization (*i.e.*,
23 optimal prediction for new observations). In order to estimate the general-
24 ization capacity of PLS regression, standard parametric approaches cannot
25 be used, and therefore the performance of a PLS regression model is evalu-
26 ated with computer-based resampling techniques such as the bootstrap and
27 cross-validation techniques where the data are separated into *learning* set (to
28 build the model) and *testing* test (to test the model). A popular example of
29 this last approach is the jackknife (sometimes called the “leave-one-out” ap-
30 proach). In the jackknife (Quenouille, 1956; Efron, 1982), each observation
31 is, in turn, dropped from the data set, the remaining observations consti-
32 tute the learning set and are used to build a PLS regression model that is
33 applied to predict the left-out observation which then constitutes the testing
34 set. With this procedure, each observation is predicted according to a ran-
35 dom effect model. These predicted observations are then stored in a matrix
36 denoted $\tilde{\mathbf{Y}}$.
37
38
39
40
41
42

43 For a random effect model, the overall quality of a PLS regression model
44 using L latent variables is evaluated by using L variables to compute—
45 according to the random model—the matrix denoted $\tilde{\mathbf{Y}}^{[L]}$ which stores the
46 predicted values of the observations for the dependent variables. The qual-
47 ity of the prediction is then evaluated as the similarity between $\tilde{\mathbf{Y}}^{[L]}$ and
48 \mathbf{Y} . As for the fixed effect model, this can be done with the squared coeffi-
49 cient of correlation (sometimes called, in this context, the “cross-validated
50 r ,” Wakeling & Morris, 1993) as well as the R_V coefficient. By analogy with
51 the RESS coefficient, one can also use the predicted residual sum of *squares*,
52
53
54
55
56
57
58
59
60

abbreviated PRESS. It is computed as:

$$\text{PRESS} = \|\mathbf{Y} - \tilde{\mathbf{Y}}^{[L]}\|^2. \quad (10)$$

The smaller the value of PRESS, the better the prediction for a random effect model, with a value of 0 indicating perfect prediction.

8.3 How many latent variables?

By contrast with the fixed effect model, the quality of prediction for a random model does not always increase with the number of latent variables used in the model. Typically, the quality first increases and then decreases. If the quality of the prediction decreases when the number of latent variables increases this indicates that the model is *overfitting* the data (*i.e.*, the information useful to fit the observations from the learning set is not useful to fit *new* observations). Therefore, for a random model, it is critical to determine the optimal number of latent variables to keep for building the model. A straightforward approach is to stop adding latent variables as soon as the PRESS decreases. A more elaborated approach (see, *e.g.*, Tenenhaus, 1998) starts by computing for the ℓ th latent variable the ratio Q_ℓ^2 defined as:

$$Q_\ell^2 = 1 - \frac{\text{PRESS}_\ell}{\text{RESS}_{\ell-1}}, \quad (11)$$

with PRESS_ℓ (resp. $\text{RESS}_{\ell-1}$) being the value of PRESS (resp. RESS) for the ℓ th (resp. $\ell - 1$) latent variable [where $\text{RESS}_0 = K \times (I - 1)$]. A latent variable is kept if its value of Q_ℓ^2 is larger than some arbitrary value generally set equal to $(1 - 95^2) = .0975$ (an alternative set of values sets the threshold to .05 when $I \leq 100$ and to 0 when $I > 100$, see Tenenhaus, 1998, Wold, 1995). Obviously, the choice of the threshold is important from a theoretical point of view, but, from a practical point of view, the values indicated above seem satisfactory.

8.4 Bootstrap confidence intervals for the dependent variables

When the number of latent variables of the model has been decided, confidence intervals for the predicted values can be derived using the bootstrap (Efron & Tibshirani, 1993). When using the bootstrap, a large number of

1
2
3
4
5
6
7
8 samples is obtained by drawing, for each sample, observations with replace-
9 ment from the learning set. Each sample provides a value of \mathbf{B}_{PLS} which
10 is used to estimate the values of the observations in the testing set. The
11 distribution of the values of these observations is then used to estimate the
12 sampling distribution and to derive confidence intervals.
13
14

15 16 9 A small example

17
18 We want to predict the subjective evaluation of a set of 5 wines. The depen-
19 dent variables that we want to predict for each wine are its likeability, and
20 how well it goes with meat, or dessert (as rated by a panel of experts, see
21 Table 1). The predictors are the price, sugar, alcohol, and acidity content of
22 each wine (see Table 2).
23
24

25 The different matrices created by PLS regression are given in Tables 3
26 to 13. From Table 9, one can find that two latent vectors explain 98% of
27 the variance of \mathbf{X} and 85% of \mathbf{Y} . This suggests that these two dimensions
28 should be kept for the final solution as a fixed effect model. The examination
29 of the two-dimensional regression coefficients (*i.e.*, \mathbf{B}_{PLS} , see Table 10) shows
30 that sugar is mainly responsible for choosing a dessert wine, and that price is
31 negatively correlated with the perceived quality of the wine (at least in this
32 example ...), whereas alcohol is positively correlated with it. Looking at
33 the latent vectors shows that \mathbf{t}_1 expresses price and \mathbf{t}_2 reflects sugar content.
34 This interpretation is confirmed and illustrated in Figures 1a and b which
35 display in (a) the projections on the latent vectors of the wines (matrix
36 \mathbf{T}) and the predictors (matrix \mathbf{W}), and in (b) the correlation between the
37 original dependent variables and the projection of the wines on the latent
38 vectors.
39
40
41

42 From Table 9, we find that PRESS reaches its minimum value for a model
43 including only the first latent variable and that Q^2 is larger than .0975 only
44 for the first latent variable. So, both PRESS and Q^2 suggest that a model
45 including only the first latent variable is optimal for generalization to new
46 observations. Consequently, we decided to keep one latent variable for the
47 random PLS regression model. Tables 12 and 13 display the predicted value
48 of $\hat{\mathbf{Y}}$ and $\tilde{\mathbf{Y}}$ when the prediction uses one latent vector.
49
50
51
52
53
54
55
56
57
58
59
60

Table 1: The matrix \mathbf{Y} of the 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 \mathbf{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

Table 3: The matrix \mathbf{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 \mathbf{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 \mathbf{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{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

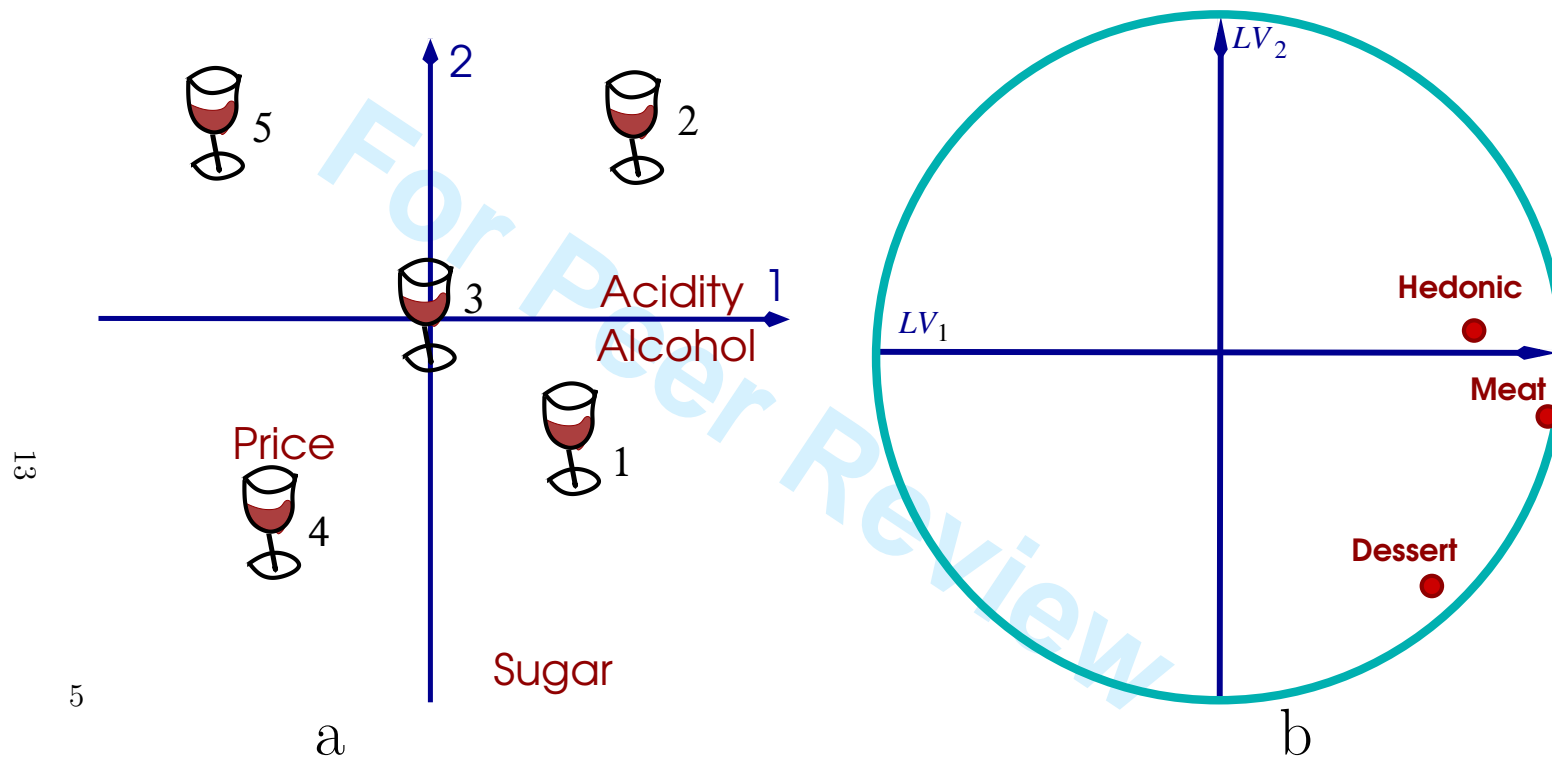


Figure 1: PLS regression-regression. (a) Projection of the wines and the predictors on the first 2 latent vectors (respectively matrices \mathbf{T} and \mathbf{W}). (b) Circle of correlation showing the correlation between the original dependent variables (matrix \mathbf{Y}) and the latent vectors (matrix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

Table 8: The \mathbf{b} vector.

b_1	b_2	b_3
2.7568	1.6272	1.1191

10 Symmetric PLS regression: BPLS regression

Interestingly, two different, but closely related, techniques exist under the name of PLS regression. The technique described so far originated from the work of Wold and Martens. In this version of PLS regression, the latent variables are computed from a succession of singular value decompositions followed by deflation of both \mathbf{X} and \mathbf{Y} . The goal of the analysis is to *predict* \mathbf{Y} from \mathbf{X} and therefore the rôles of \mathbf{X} and \mathbf{Y} are asymmetric. As a consequence, the latent variables computed to predict \mathbf{Y} from \mathbf{X} are different from the latent variables computed to predict \mathbf{X} from \mathbf{Y} .

A related technique, also called PLS regression, originated from the work of Bookstein (1994, see also Tucker, 1958 for early related ideas; McIntosh, Bookstein, Haxby, & Grady, C.L., 1996; and Bookstein, Steigsguth, Sampson, Conner, & Barr, 2002; for later applications). To distinguish this version of PLS regression from the previous one, we will call it BPLS regression.

This technique is particularly popular for the analysis of brain imaging data (probably because it requires much less computational time, which is critical taking into account the very large size of brain imaging data sets). Just like standard PLS regression (*cf.* Equations 6 and 7), BPLS regression starts with the matrix

$$\mathbf{S} = \mathbf{X}^T \mathbf{Y} . \quad (12)$$

The matrix \mathbf{S} is then decomposed using its singular value decomposition as:

$$\mathbf{S} = \mathbf{W} \mathbf{\Theta} \mathbf{C}^T \text{ with } \mathbf{W}^T \mathbf{W} = \mathbf{C}^T \mathbf{C} = \mathbf{I} , \quad (13)$$

(where \mathbf{W} and \mathbf{C} are the matrices of the left and right singular vectors of \mathbf{S} and $\mathbf{\Theta}$ is the diagonal matrix of the singular values, *cf.* Equation 1). In BPLS regression, the latent variables for \mathbf{X} and \mathbf{Y} are obtained as (*cf.* Equation 5):

$$\mathbf{T} = \mathbf{XW} \text{ and } \mathbf{U} = \mathbf{YC} . \quad (14)$$

Table 9: Variance of \mathbf{X} and \mathbf{Y} explained by the latent vectors, RESS, PRESS and Q^2 .

Latent Vector	Percentage of Explained Variance for \mathbf{X}	Cumulative Percentage of Explained Variance for \mathbf{X}	Percentage of Explained Variance for \mathbf{Y}	Cumulative Percentage of Explained Variance for \mathbf{Y}	RESS	PRESS	Q^2
1	70	70	63	63	32.11	95.11	7.93
2	28	98	22	85	25.00	254.86	-280
3	2	100	10	95	1.25	101.56	-202.89

Table 10: The matrix \mathbf{B}_{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

Table 11: The matrix \mathbf{B}_{PLS}^* when 2 latent vectors are used.

	Hedonic	Goes with meat	Goes with dessert
Intercept	-3.2809	-3.3770	-1.3909
Price	-0.2559	-0.1129	0.0063
Sugar	0.1418	0.3401	0.6227
Alcohol	0.8080	0.4856	0.2713
Acidity	0.6870	0.3957	0.1919

Table 12: The matrix $\hat{\mathbf{Y}}$ when one latent vector is used.

Wine	Hedonic	Goes with meat	Goes with dessert
1	11.4088	6.8641	6.7278
2	12.0556	7.2178	6.8659
3	8.0000	5.0000	6.0000
4	4.7670	3.2320	5.3097
5	3.7686	2.6860	5.0965

Table 13: The matrix $\tilde{\mathbf{Y}}$ when one latent vector is used.

	Wine	Hedonic	Goes with meat	Goes with dessert
1	1	8.5877	5.7044	5.5293
2	2	12.7531	7.0394	7.6005
3	3	8.0000	5.0000	6.2500
4	4	6.8500	3.1670	4.4250
5	5	3.9871	4.1910	6.5748

Because BPLS regression uses a single singular value decomposition to compute the latent variables, they will be identical if the rôles of \mathbf{X} and \mathbf{Y} are reversed: BPLS regression treats \mathbf{X} and \mathbf{Y} symmetrically. So, while standard PLS regression is akin to multiple regression, BPLS regression is akin to correlation or canonical correlation (Gittins, 1985). BPLS regression, however, differs from canonical correlation because BPLS regression extracts the *variance* common to \mathbf{X} and \mathbf{Y} whereas canonical correlation seeks linear combinations of \mathbf{X} and \mathbf{Y} having the *largest correlation*. In fact, the name of partial least squares *covariance* analysis or canonical *covariance* analysis would probably be more appropriate for BPLS regression.

10.1 Varieties of BPLS regression

BPLS regression exists in three main varieties, one of which being specific to brain imaging. The first variety of BPLS regression is used to analyze experimental results, it is called *Behavior* BPLS regression if the \mathbf{Y} matrix consists of measures or *Task* BPLS regression if the \mathbf{Y} matrix consists of contrasts or describes the experimental conditions with dummy coding.

The second variety is called *mean centered task* BPLS regression and is closely related to barycentric discriminant analysis (*e.g.*, discriminant correspondence analysis, see, Abdi, 2007d). Like discriminant analysis, this approach is suited for data in which the observations originate from groups defined *a priori*, but, unlike discriminant analysis, it can be used for small N , large P problems. The \mathbf{X} matrix contains the deviations of the observations to the average vector of all the observations, and the \mathbf{Y} matrix uses a dummy code to identify the group to which each observation belongs (*i.e.*, \mathbf{Y} has as

1
2
3
4
5
6
7
8 many columns as there are groups, with a value of 1 at the intersection of
9 the i th row and the k th column indicating that the i th row belongs to the
10 k th group, whereas a value of 0 indicates that it does not). With this coding
11 scheme, the \mathbf{S} matrix contains the group barycenters and the BPLS regression
12 analysis of this matrix is equivalent to a PCA of the matrix of the barycenters
13 (which is the first step of barycentric discriminant analysis).
14

15 The third variety, which is specific to brain imaging, is called *Seed* PLS
16 regression. It is used to study patterns of connectivity between brain regions.
17 Here the columns of a matrix of brain measurements (where rows are scans
18 and columns are voxels) are partitioned into two sets: A small one called
19 the *seed* and a larger one representing the rest of the brain. In this context,
20 the \mathbf{S} matrix contains the correlation between the columns of the seed and
21 the rest of the brain. The analysis of the \mathbf{S} matrix reveals the pattern of
22 connectivity between the seed and the rest of the brain.
23
24
25
26

27 11 Relationship with other techniques

28
29
30 PLS regression is obviously related to canonical correlation (see Gittins,
31 1985), STATIS, and multiple factor analysis (see Abdi & Valentin, 2007a,b
32 for an introduction to these techniques). These relationships are explored
33 in detail by Tenenhaus (1998), Pagès and Tenenhaus (2001), Abdi (2003b),
34 Rosipal and Krämer (2006), and in the volume edited by Esposito Vinzi,
35 Chin, Henseler, and Wang (2009). The main originality of PLS regression is
36 to preserve the asymmetry of the relationship between predictors and depen-
37 dent variables, whereas these other techniques treat them symmetrically.
38
39

40 By contrast, BPLS regression is a symmetric technique and therefore is
41 closely related to canonical correlation, but BPLS regression seeks to extract
42 the variance common to \mathbf{X} and \mathbf{Y} whereas canonical correlation seeks linear
43 combinations of \mathbf{X} and \mathbf{Y} having the largest correlation (some connections
44 between BPLS regression and other multivariate techniques relevant for brain
45 imaging are explored in Kherif *et al.*, 2003; Friston & Büchel, 2004; Lazar,
46 2008). The relationships between BPLS regression, and STATIS or multiple
47 factor analysis have not been analyzed formally, but these techniques are
48 likely to provide similar conclusions.
49
50
51
52
53
54
55
56
57
58
59
60

12 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. 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, several standard commercial statistical packages (*e.g.*, SAS, SPSS, STATISTICA), include PLS regression. The public domain R language also includes PLS regression. A dedicated public domain called SmartPLS is also available.

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. Staying with MATLAB, the statistical toolbox includes a PLS regression routine.

For brain imaging (a domain where the Bookstein approach is, by far, the most popular PLS regression approach), 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. (2001). Linear algebra for neural networks. In N.J. Smelser, P.B. Baltes (Eds.): *International encyclopedia of the social and behavioral sciences*. Oxford (UK): Elsevier.
- [2] Abdi, H. (2003a). PLS regression. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks: Sage. (pp.792–795).
- [3] Abdi, H. (2003b). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks: Sage. (pp.699–702).
- [4] Abdi, H. (2007a). Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). In N.J. Salkind (Ed): *Encyclo-*

- 1
2
3
4
5
6
7
8 *pedia of measurement and statistics*. Thousand Oaks: Sage. (pp.907–
9 912).
- 10 [5] Abdi, H. (2007b). Eigen-decomposition: eigenvalues and eigenvecteurs, In
11 N.J. Salkind (Ed): *Encyclopedia of measurement and statistics*. Thou-
12 sand Oaks: Sage. (pp.304–308).
- 13 [6] Abdi, H. (2007c). RV coefficient and Congruence coefficient, In N.J.
14 Salkind (Ed): *Encyclopedia of measurement and statistics*. Thousand
15 Oaks: Sage. (pp.849–853).
- 16 [7] Abdi, H. (2007d). Discriminant correspondence analysis, In N.J. Salkind
17 (Ed): *Encyclopedia of measurement and statistics*. Thousand Oaks:
18 Sage. (pp.270–275).
- 19 [8] Abdi, H., & Valentin (2007a). STATIS, In N.J. Salkind (Ed): *Encyclopedia*
20 *of measurement and statistics*. Thousand Oaks: Sage. (pp.955–962).
- 21 [9] Abdi, H., & Valentin (2007b). Multiple factor analysis, In N.J. Salkind
22 (Ed): *Encyclopedia of measurement and statistics*. Thousand Oaks:
23 Sage. (pp.651–657).
- 24 [10] Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Thousand
25 Oaks (CA): Sage.
- 26 [11] Burnham, A.J., Viveros, R., MacGregor, J.F. (1996). Frameworks for
27 latent variable multivariate regression. *Journal of Chemometrics*, **10**,
28 31–45.
- 29 [12] Draper, N.R., & Smith H. (1998). *Applied regression analysis (3rd Edi-*
30 *tion)*. New York: Wiley.
- 31 [13] Efron, B. (1982). *The Jackknife, the bootstrap and other resampling plans*.
32 Philadelphia: SIAM.
- 33 [14] Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*.
34 Chapman and Hall, New York.
- 35 [15] Escofier, B., & Pagès, J. (1988). *Analyses factorielles multiples*. Paris:
36 Dunod.
- 37 [16] Esposito Vinzi, V., Chin, W.W., Henseler, J., & Wang, H. (Eds.) *Hand-*
38 *book of partial least squares concepts, methods and applications in mar-*
39 *keting and related fields*. New York: Springer Verlag.
- 40 [17] Fornell C., Lorange, P., Roos, J. (1990). The cooperative venture forma-
41 tion process: A latent variable structural modeling approach. *Manage-*
42 *ment Science*, **36**, 1246–1255.
- 43 [18] Frank, I.E., & Friedman, J.H. (1993). A statistical view of chemometrics
44 regression tools. *Technometrics*, **35** 109–148.
- 45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3
4
5
6
7
8 [19] Friston, K., Büchel, C. (2004) Functional integration. In Frackowiak,
9 R.S.J., Friston, K.J., Frith, C.D., Dolan, R.J., Price, C.J., Zeki, S.,
10 Ashburner, J.T., Penny, W.D. (Eds.), *Human brain function*. New York:
11 Elsevier. (pp.999–1019).
12
13 [20] Kovacevic, N., & McIntosh, R. (2007). Groupwise independent component
14 decomposition of EEG data and partial least square analysis *NeuroIm-*
15 *age*, **35**, 1103–1112.
16
17 [21] Garthwaite, P. (1994). An interpretation of partial least squares. *Journal*
18 *of the American Statistical Association*, **89**, 122–127.
19
20 [22] Graham, J.L., Evenko, L.I., Rajan, M.N. (1992). An empirical comparison
21 of soviet and american business negotiations. *Journal of International*
22 *Business Studies*, **5**, 387–418.
23
24 [23] Giessing, C., Fink, G.R., Rošler, F., & Thiel, C.M. (2007). fMRI data
25 predict individual differences of behavioral effects of nicotine: A partial
26 least square analysis. *Journal of Cognitive Neuroscience* **19**, 658–670.
27
28 [24] Gittins, R. (1985). *Canonical analysis: A review with applications in ecol-*
29 *ogy*. New York: Springer.
30
31 [25] Helland, I.S. (1990). PLS regression and statistical models. *Scandinavian*
32 *Journal of Statistics*, **17**, 97–114.
33
34 [26] Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: Biased estima-
35 tion for nonorthogonal problem. *Technometrics*, **12**, 55–67.
36
37 [27] Höskuldson, A. (1988). PLS regression methods. *Journal of Chemomet-*
38 *rics*, **2**, 211–228.
39
40 [28] Höskuldson, A. (in Press). Modelling Procedures for Directed Network of
41 Data Blocks. *Chemometrics and Intelligent Laboratory Systems*.
42
43 [29] Höskuldson, A. (2001). Weighting schemes in multivariate data analysis.
44 *Journal of Chemometrics*, **15**, 371–396.
45
46 [30] Hulland J. (1999). Use of partial least square in strategic management
47 research: A review of four recent studies. *Strategic Management Journal*,
48 **20**, 195–204.
49
50 [31] Geladi, P., & Kowalski B. (1986). Partial least square regression: A tuto-
51 rial. *Analytica Chimica Acta*, **35**, 1–17.
52
53 [32] Kherif, F., Poline, J.B., Flandin, G., Benali, H., Simon, O., Dehaene, S.,
54 & Worsley, K.J. (2002). Multivariate model specification for fMRI data.
55 *NeuroImage*, **16**, 1068–1083.
56
57 [33] Lazar, N.A. (2008). *The statistical analysis of functional MRI data*. New
58 York: Springer.
59
60

- 1
2
3
4
5
6
7
8 [34] McIntosh, A.R., & Bookstein, F.L., Haxby, J.V., & Grady, C.L. (1996).
9 Spatial pattern analysis of functional brain images using partial least
10 squares. *Neuroimage*, **3**, 143–157.
11 [35] McIntosh, A.R., & Lobaugh N.J. (2004). Partial least squares analysis of
12 neuroimaging data: applications and advances. *Neuroimage*, **23**, 250–
13 263.
14 [36] Martens, H., & Naes, T. (1989). *Multivariate calibration*. London: Wiley.
15 [37] Martens, H., & Martens, M. (2001). *Multivariate analysis of quality : An*
16 *introduction*. London Wiley.
17 [38] Pagès, J., & Tenenhaus, M. (2001). Multiple factor analysis combined
18 with PLS regression path modeling. Application to the analysis of rela-
19 tionships between physicochemical variables, sensory profiles and hedonic
20 judgments. *Chemometrics and Intelligent Laboratory Systems*, **58**,
21 261–273.
22 [39] Phatak, A., & de Jong, S. (1997). The geometry of partial least squares.
23 *Journal of Chemometrics*, **11**, 311–338.
24 [40] Quenouille, M. (1956). Notes on Bias and Estimation. *Biometrika*, **43**,
25 353–360.
26 [41] Rännar, Lindgren, Geladi, & Wold, (1994) A PLS regression kernel al-
27 gorithm for data sets with many variables and fewer objects. Part 1:
28 Theory and algorithms. *Journal of Chemometrics*, **8**, 111–125.
29 [42] Rosipal, R., & Krämer, N. (2006). Overview and recent advances in partial
30 least squares. in C. Saunders et al. (Eds). *Subspace, Latent Structure and*
31 *Feature Selection: Statistical and Optimization Perspectives Workshop*
32 *(SLSFS 2005)*. New York: Springer-Verlag. pp. 34–51.
33 [43] Tenenhaus, M. (1998). *La régression PLS regression*. Paris: Technip.
34 [44] Ter Braak, C.J.F., & de Jong, S. (1998). The objective function of partial
35 least squares regression. *Journal of Chemometrics*, **12**, 41–54.
36 [45] Tucker, L.R. (1958). Determination of parameters of a functional relation
37 by factor analysis. *Psychometrika*, **23**, 19–23.
38 [46] Wang, J.Y., Bakhadirov, K., Devous, M.D. Sr., Abdi, H., McColl, R.,
39 Moore, C. Marquez de la Plata, C.D., Ding, K., Whittemore, A. Bab-
40 cock, E., Rickbeil, E.T., Dobervich, J., Kroll, D., Dao, B., Mohindra, N.,
41 & Diaz-Arrastia, R. (2008). Diffusion tensor tractography of traumatic
42 diffuse axonal injury. *Archives of Neurology*, **65**, 619–626.
43 [47] Wakeling, I.N., & Morris, J. (1993). A test for significance for partial least
44 squares regression. *Journal of Chemometrics*, **7**, 291–304.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3
4
5
6
7
8 [48] Wold, H. (1966). Estimation of principal components and related mod-
9 els by iterative least squares. In P.R. Krishnaiah (Ed.). *Multivariate*
10 *analysis*. (pp.391–420) New York: Academic Press.
11
12 [49] Wold, S. (2001). Personal memories of the early PLS development. *Chemo-*
13 *metrics and Intelligent Laboratory Systems*, **58**, 83–84.
14
15 [50] Worsley, K.J. (1997). An overview and some new developments in the
16 statistical analysis of PET and fMRI data. *Human Brain Mapping*, **5**,
17 254–258.
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review