PLS regression : a directional signal-noise approach

Pierre Druilhet [a,*] Alain Mom [b]

[a] *CREST-ENSAI, Campus de Ker Lann, 35 170 BRUZ, France*

[b] *Laboratoire de statistique, Université Rennes II, RENNES, France*

Abstract

We present a new approach of univariate partial least squares regression (PLSR) based on directional signal-noise ratios. We show how PLSR, unlike PCR, takes into account the actual value and not only the variance of the OLS estimator. We find an orthogonal sequence of directions associated with decreasing signal-noise ratios. Then, we state PLS estimators as least squares estimators constrained to be null onto the last directions. We also give another procedure that shows how PLSR rebuilds the OLS estimator iteratively by seeking at each step the direction with largest difference of signals over the noise. The latter approach does not involve any arbitrary scale or orthogonality constraints.

Key words: Biased regression, Constrained least squares, Latent variables, Partial least squares, Principal component, Shrinkage
1991 MSC: 62J05, 62J07

1 Introduction

Many different regression methods are now available to improve ordinary least squares (OLS) estimators in multiple linear regression when explanatory variables are strongly correlated. Like principal component regression (PCR), partial least squares regression (PLSR) is based on the construction of latent variables, i.e. new variables that are linear combinations of the original explanatory variables. In PCR, these latent variables have a statistical meaning: they explain the internal covariance structure of the explanatory variables.

* Corresponding author
Email addresses: druilhet@ensai.fr (Pierre Druilhet), alain.mom@uhb.fr (Alain Mom).

In PLSR, the statistical interpretation of the latent variables is not always so clear, but intuitively, they make a compromise between the internal structure of the explanatory variables and their relationship with the response. Continuum regression (Stone and Brooks, 1990) gives a more general latent variables method that includes PLSR and PCR. Latent variables approach is often related to an underlying common joint covariance distribution structure between the explanatory variables and the response for each individual: Helland (1990, 1992), Helland and Almøy (1994) give a population model for PLSR and establish asymptotic properties. However, in some situations, the sampling scheme used to collect data can be complex or unknown; on the other hand, the aim of the study can be to predict new individuals with the same conditional distribution but different joint distributions. In these cases, it is of interest to formulate a conditional model and then to consider the explanatory variables as fixed. See Breiman and Spector (1992) for further discussions and simulation studies upon the difference between fixed and random cases. In this paper, we assume a conditional model and focus on the improvement of the least squares estimator.

Like PCR and Ridge regression, PLSR provides a shrunk version of the OLS estimator (De Jong, 1995 and Goutis, 1996). For a comparison of these three techniques by simulations and case studies, see Frank and Friedman (1993); for a general discussion, see Brown (1993). When the explanatory variables are strongly correlated, the improvement of OLS estimator obtained by shrinking can be quite large w.r.t. the mean-squared error (MSE). However the MSE is a global measure of how the estimate is far from the true parameter and does not show how the estimate is improved on specific directions. Butler and Denham (2000) and Lingjærde and Christophersen (2000) study the shrinkage properties of PLS estimators onto directions given by the singular values decomposition of the design matrix.

In this paper, we aim to show that PLSR singles out some directions to shrink OLS estimators. More precisely, while PCR shrinks the OLS estimator onto directions with large variance (under a scale constraint), we show how PLSR do the same onto some directions corresponding to small signal-noise ratios. This signal-noise ratio is directly related with optimal directional shrinkage factors. We also show how the sequence of PLS estimators rebuilds successively the OLS estimator by seeking the direction of maximal difference of signals over the noise.

In section 2, we give a brief review of the classical constrained estimators. In section 3, we recall some known results on the link between shrinkage coefficients and signal-noise ratios for a one-dimensional parameter. Then, we apply these results to linear models. In section 4, we use a maximization procedure to obtain a sequence of orthogonal directions with decreasingly ordered signal-noise ratios. Then, we obtain PLS estimators as OLS estimators

under the constraint of nullity onto the last directions. In section 5, we propose a constraint free algorithm based on the difference of signals over the noise. This algorithm gives iteratively the PLS estimators.

2 Multiple linear model and biased estimation

We assume the following model for the n individuals:

$$\dot{y}_i = \mu + \dot{\mathrm{x}}_i'\beta + \varepsilon_i, \quad i = 1, ..., n, \tag{1}$$

where $\dot{y}_i$ is the real response observed on the i^{th} individual, $\dot{\mathrm{x}}_i$ are the p-vector of its explanatory variables considered as fixed, β is the unknown p-vector of parameters, and ε_i are i.i.d. mean zero variance σ^2 real random variables. Removing the dots on the original data means that their sample averages, denoted by $\bar{\dot{\mathrm{y}}}$ and $\bar{\dot{\mathrm{x}}}$, has been removed. In vector notation, we have for the centered model:

$$\mathrm{y} = X\beta + \varepsilon. \tag{2}$$

To avoid handling complicated generalized inverses, the design matrix X is assumed to be of full-rank. The non full-rank case will be discussed in section 7.

We write $\mathrm{s} = X'\mathrm{y}$ and $S = X'X$. We denote by I_n the (n, n) identity matrix. Let A be a (r,s) matrix, and M be a (r,r) symmetric positive matrix, we denote $P_A^M = A(A'MA)^- A'M$ the projector matrix onto Range(A) w.r.t. M. We also write $P_A = P_A^{I_r}$.

Consider a new individual $\dot{\mathrm{x}}_o$ and assume that its response $\dot{\mathrm{y}}_o$ is given by

$$\dot{\mathrm{y}}_o = \mu + \dot{\mathrm{x}}_o'\,\beta + \varepsilon_o \tag{3}$$

where ε_o is independent of $(\varepsilon_1, ...\varepsilon_n)$ with mean 0 and variance σ_o^2. We want to predict $\dot{y}_o$ by $\bar{\dot{\mathrm{y}}} + \mathrm{x}_o'\hat{\beta}$, where $\mathrm{x}_o = \dot{\mathrm{x}}_o - \bar{\dot{\mathrm{x}}}_o$ and $\hat{\beta}$ is an estimate of β. It is well known that the best linear unbiased estimator of $\mathrm{x}_o'\,\beta$ is $\mathrm{x}_o'\,\hat{\beta}^{\text{OLS}}$, where $\hat{\beta}^{\text{OLS}} = (X'X)^{-1}X'\mathrm{y} = S^{-1}\,\mathrm{s}$. It is also well known that $\bar{\dot{\mathrm{y}}} + \mathrm{x}_o'\,\hat{\beta}^{\text{OLS}}$ is the best linear unbiased predictor of $\dot{\mathrm{y}}_o$. However, for some x_o, $\mathrm{x}_o'\,\hat{\beta}^{\text{OLS}}$ gives very poor estimate of $\mathrm{x}_o'\,\beta$ or predictor of $\dot{\mathrm{y}}_o$. There are many methods to improve the estimation of β by biased estimators.

For example, Ridge estimators are obtained by approximating S by $S + \alpha I$ in the normal equation $S\,\hat{\beta} = \mathrm{s}$, where the real α is a control parameter. They

can also be obtained as constrained least squares estimators:

$$\hat{\beta}_c^{\text{Ridge}} = \underset{\beta:\beta'\beta\leq c}{\text{ArgMin}} \, ||\mathrm{y} - X\beta||^2 \tag{4}$$

where c depends of α.

In a similar way, Lasso estimators (Tibshirani, 1996) are defined by

$$\hat{\beta}_c^{\text{Lasso}} = \underset{\beta:\sum|\beta_i|\leq c}{\text{ArgMin}} \, ||\mathrm{y} - X\beta||^2. \tag{5}$$

In PCR, it is postulated that bad estimation of $\mathrm{x}_o'\,\beta$ by OLS is due to the fact that $\text{var}(\mathrm{x}_o'\,\hat{\beta}^{\text{OLS}})$ can be very "large". To give a sense to "large", a scale constraint must be added: in PCR, the constraint is $x_o'\,x_o = 1$. Denote by $\vec{w}_1^{\text{PCR}}, ..., \mathrm{w}_p^{\text{PCR}}$ an orthogonal basis of unit eigenvectors of S corresponding to decreasingly ordered eigenvalues. Since $\text{var}(x_o'\,\hat{\beta}^{\text{OLS}}) = \sigma^2\,\mathrm{x}_o' S^{-1}\mathrm{x}_o$, PCR forces the OLS to be zero onto the directions given by the last eigenvectors. More precisely, define $\overline{W}_{(q)}^{\text{PCR}} = (\mathrm{w}_{q+1}^{\text{PCR}}|...|\mathrm{w}_p^{\text{PCR}})$, then

$$\beta_{(q)}^{\text{PCR}} = \underset{\beta:\overline{W}_{(q)}^{\text{PCR}\prime}\beta=0}{\text{ArgMin}} \, ||\mathrm{y} - X\beta||^2. \tag{6}$$

This expression shows how the OLS estimator is modified in a statistically meaningful way. On another way, $\beta_{(q)}^{\text{PCR}}$ can be obtained after a regression on the latent variables $\mathrm{t}_1^{\text{PCR}} = X\,\mathrm{w}_1^{\text{PCR}}, \mathrm{t}_2^{\text{PCR}} = X\,\mathrm{w}_2^{\text{PCR}}, ..., \mathrm{t}_q^{\text{PCR}} = X\,\mathrm{w}_q^{\text{PCR}}$. In that case, the vectors $\mathrm{w}_1^{\text{PCR}}, \mathrm{w}_2^{\text{PCR}}, ..., \mathrm{w}_q^{\text{PCR}}$ are weight vectors. The fact that the same vector w_i represent two different things, either a direction or a weight vector, comes from the fact that $\mathrm{w}_1^{\text{PCR}}, \mathrm{w}_2^{\text{PCR}}, ..., \mathrm{w}_p^{\text{PCR}}$ are orthogonal (see Lemma 5 and Formula 10). The main criticism of PCR is that the constraints $\overline{W}_{(q)}^{\text{PCR}\prime}\beta = 0$ are obtained by considering $\text{var}(x_o'\,\hat{\beta}^{\text{OLS}})$ (the noise) but not the actual value of $x_o'\,\hat{\beta}^{\text{OLS}}$ (the signal). In other words, PCR constructs the latent variables independently of their relationship with the response y.

In the usual approach, PLSR is a latent variables method that aims to overcome this undesirable feature of PCR. The first latent variables $\mathrm{t}_1^{\text{PLS}} = X\,\mathrm{w}_1^{\text{PLS}}$ maximizes the empirical covariance:

$$\text{cov}_e(\mathrm{t}, \mathrm{y}) = \text{cov}_e(X\,\mathrm{w}, \mathrm{y}) = \mathrm{w}'\,X\,\mathrm{y} \tag{7}$$

under the scale constraint $\mathrm{w}'\mathrm{w} = 1$. The i^{th} latent variables $\mathrm{t}_i = X\,\mathrm{w}_i^{\text{PLS}}$ maximizes (7) under the constraints $\mathrm{w}'\mathrm{w} = 1$ and $\mathrm{t}_i \perp (\mathrm{t}_1^{\text{PLS}}, ..., \mathrm{t}_{i-1}^{\text{PLS}})$. Define $T_{(q)}^{\text{PLS}} = (\mathrm{t}_1^{\text{PLS}}|\cdots|\mathrm{t}_q^{\text{PLS}})$ and $W_{(q)}^{\text{PLS}} = (\mathrm{w}_1^{\text{PLS}}|...|\mathrm{w}_q^{\text{PLS}})$. The PLS estimator $\hat{\beta}_{(q)}^{\text{PLS}}$ based on the q first latent variables is:

$$\hat{\beta}^{\text{PLS}}_{(q)} = W^{\text{PLS}}_{(q)} \left(T^{\text{PLS}\prime}_{(q)} \, T^{\text{PLS}}_{(q)} \right)^{-1} \, T^{\text{PLS}\prime}_{(q)} \, \mathrm{y}, \tag{8}$$

$$= W^{\text{PLS}}_{(q)} \left(W^{\text{PLS}\prime}_{(q)} \, S \, W^{\text{PLS}}_{(q)} \right)^{-1} W^{\text{PLS}\prime}_{(q)} S \, \hat{\beta}^{\text{OLS}}, \tag{9}$$

$$= \mathrm{P}^{S}_{W^{\text{PLS}}_{(q)}} \, \hat{\beta}^{\text{OLS}}. \tag{10}$$

Note that equations 8, 9 and 10 are general properties of a latent variables regressors.

Helland (1988) shows that $\text{Range}\left(W_{(q)}\right) = K_q$ where K_q is the Krylov subspace defined by:

$$K_q = \text{span}\left(\mathbf{s}, S\,\mathbf{s}, S^2\,\mathbf{s}, ..., S^{q-1}\,\mathbf{s}\right). \tag{11}$$

He also shows that Martens' algorithm (see Martens and Næs, 1989) is equivalent to a Gram-Schmidt orthogonalization procedure of the nested Krylov subspaces $(K_1, K_2, ...)$ and gives the same estimator, as easily seen with Formula 10. Note that Formulas 8, 9 and 10 are general expressions of a latent variables estimator.

The goal of the paper is to find an analogue to Formula 6 for PLSR. Of course, any estimator obtained by latent variables, whatever they are, can be expressed as a least squares estimator under some linear constraints. However, we seek for PLSR constraints corresponding to direction that have a clear statistical meaning. Since PLSR is a shrinkage method, shrinkage factors onto directions are of importance. This will be the key point of our approach.

3 Shrinkage factors and signal-noise ratio

The signal-noise ratio arises naturally in univariate estimation to deal with shrinkage. First, we recall some known facts on unidimensional shrinkage. Then, we apply these notions to directions in linear model.

3.1 Shrinking a one-dimensional estimator or predictor

Let $\hat{\theta}$ be an unbiased estimator of a real parameter θ. We denote by σ^2 the variance of $\hat{\theta}$. We want to improve $\hat{\theta}$ by shrinking, so we consider new estimators $a\,\hat{\theta}$ for $a \in \mathbb{R}$. Let a^* be the scalar that minimizes the MSE, i.e. :

$$a^* = \underset{a \in \mathbb{R}}{\text{ArgMin}}\, \mathbb{E}\left(a\,\hat{\theta} - \theta\right)^2. \tag{12}$$

We have,

$$a^* = \frac{\rho^2}{1+\rho^2} \quad \text{where} \quad \rho = \frac{|\theta|}{\sigma}. \tag{13}$$

When ρ is known, which arises in some scale models, $a^*\,\hat{\theta}$ is uniformly better under the quadratic lost among all the estimators $a\,\hat{\theta}$. When ρ is unknown but σ^2 is known, ρ can be estimated by

$$\hat{\rho} = \frac{|\hat{\theta}|}{\sigma} \quad \text{and} \quad \hat{a}^* = \frac{\hat{\rho}^2}{1+\hat{\rho}^2} = \frac{\hat{\theta}^2}{\sigma^2+\hat{\theta}^2}. \tag{14}$$

We see that the shrinkage factor $\hat{a}^*$ is an increasing function of the signal-noise ratio $\hat{\rho}$. Bibby and Toutenburg (1977) study the properties of the estimator $\hat{a}^*\,\hat{\theta}$. They also show that, in the normal case, the probability that $\hat{a}^*\,\hat{\theta}$ improve $\hat{\theta}$ is very high: 77% or more depending on ρ. When $\rho < 1$, the probability is equal to 1.

Consider now the following prediction problem: we want to predict $\mathrm{y}_o = \theta + \varepsilon_o$ where $\varepsilon_o \sim (0, \sigma_o^2)$ is independent of $\hat{\theta}$. The predictor $\tilde{\mathrm{y}}_o = \hat{\theta}$ is an unbiased predictor of y_o, i.e. $\mathbb{E}(\hat{\theta}) = \mathbb{E}(\mathrm{y}_o) = \theta$, and we want to improve the prediction by shrinking. The equivalent of the MSE in prediction is the mean-squared error of prediction (MSEP):

$$\mathrm{MSEP} = \mathbb{E}(\tilde{\mathrm{y}}_o - \mathrm{y}_o)^2 = \mathrm{MSE} + \sigma_o^2.$$

The predictor $a\,\tilde{\mathrm{y}}_o$ that minimizes the MSEP is $a^*\,\tilde{\mathrm{y}}_o$ with a^* given by equation 13. Thus in both estimation and prediction problem, the best shrinkage factor is the same.

3.2 Shrinkage factors in multiple linear regression

We apply the results of Section 3.1 to the linear model (2). For some $\mathrm{x}_o \in \mathbb{R}^p$, we are interested in the estimation of $\mathrm{x}_o'\beta$. The minimum-variance unbiased linear estimator of $\mathrm{x}_o'\beta$ is $\mathrm{x}_o'\hat{\beta}^{\text{OLS}}$. To improve this estimator, we apply the shrinkage factor presented in Section 3.1. So, we have:

$$a^*_{\mathrm{x}_o} = \frac{\rho^2_{\mathrm{x}_o}}{1+\rho^2_{\mathrm{x}_o}} \quad \text{where} \quad \rho_{\mathrm{x}_o} = \frac{|\mathrm{x}_o'\,\beta|}{\sigma\,\sqrt{\mathrm{x}_o'\,S^{-1}\mathrm{x}_o}}, \tag{15}$$

that can be estimated by

$$\hat{a}^*_{\mathrm{x}_o} = \frac{\hat{\rho}^2_{\mathrm{x}_o}}{1+\hat{\rho}^2_{\mathrm{x}_o}} \quad \text{where} \quad \hat{\rho}_{\mathrm{x}_o} = \frac{|\mathrm{x}'_o\ \hat{\beta}^{\text{OLS}}|}{\sigma\ \sqrt{\mathrm{x}'_o\, S^{-1}\mathrm{x}_o}}. \tag{16}$$

If σ^2 is not known, it can be replaced by an estimation. However, σ^2 is a positive constant and will not play any role in the following. Note that $\mathrm{x} \mapsto \hat{\rho}_{\mathrm{x}}$ is positively homogeneous, that is:

$$\forall \alpha \neq 0, \quad \hat{\rho}_{\alpha \mathrm{x}} = \hat{\rho}_{\mathrm{x}}. \tag{17}$$

Consequently, the signal-noise ratio $\hat{\rho}_{\mathrm{x}_o}$ depends only on the direction given by x_o, not on the exact x_o. More precisely:

$$\hat{\rho}_{\mathrm{x}_o} = \frac{|\mathrm{x}'_o\ \hat{\beta}|}{\sigma\ \sqrt{\mathrm{x}'_o\, S^{-1}\mathrm{x}_o}} = \frac{|\mathrm{x}'_o\ S^{-1}\mathrm{s}|}{\sigma\ \sqrt{\mathrm{x}'_o\, S^{-1}\mathrm{x}_o}} = \frac{\sqrt{\mathrm{s}\, S^{-1}\mathrm{s}}}{\sigma} \left|\cos_{S^{-1}}(\widehat{\mathrm{x}_o, \mathrm{s}})\right|, \tag{18}$$

where $\cos_{S^{-1}}(\widehat{\mathrm{x}_o, \mathrm{s}})$ is the cosine of the angle $(\widehat{\mathrm{x}_o, \mathrm{s}})$ w.r.t. the quadratic form S^{-1}. Thus, $\hat{\rho}_{\mathrm{x}_o}$ depends only on the angle between x_o and $\mathrm{s} = X'\mathbf{y}$ w.r.t. the quadratic form S^{-1}. The greater the angle w.r.t S^{-1} between $\mathbf{x}$ and s, the less its signal ratio. Thus, the space of null signal-noise, which coincides with the space of null signal, is the orthogonal complement *w.r.t* the quadratic form S^{-1} of $\mathbf{s}$. Moreover, the relationship between $a^*_{\mathrm{x}_o}$ and x_o is intrinsically non linear and depends only on ρ_{x_o}.

4 PLSR and directional signal-noise algorithm

In the classical algorithms used to construct PLS estimators, two kinds of arbitrariness are involved: a scale constraint and an orthogonality constraint. In this section, we present an algorithm based on the signal-noise ratio. Since the signal-noise ratio is scale invariant, no scale constraint is necessary. The algorithm seeks a sequence of orthogonal directions corresponding to decreasing signal-noise ratios. Then, we show that PLS estimators as least squares estimators constrained to be null onto the last directions of this sequence.

4.1 The signal-noise maximization procedure

Let start to seek the direction w_1 that maximizes the signal-noise ratio, i.e.

$$\mathrm{w}_1 = \underset{\mathrm{w}\in\mathbb{R}^p}{\mathrm{ArgMax}}\,(\hat{\rho}_{\mathrm{w}}) = \underset{\mathrm{w}\in\mathbb{R}^p}{\mathrm{ArgMax}}\,\frac{|\mathrm{w}'\hat{\beta}|}{\sigma\,\sqrt{\mathrm{w}'S^{-1}\mathrm{w}}}. \tag{19}$$

By a direct application of Cauchy-Schwarz's inequality and because $\mathrm{w}'\hat{\beta} = \mathrm{w}'S^{-1}\mathrm{s}$, we have:

Proposition 1 *The first direction is spanned by* s*, that is* w_1 *is any vector proportional to* s.

It is interesting to see that the direction of maximal noise play a central role in PLSR.

Iteratively, at step i, the direction w_i maximizes the signal-noise ratio $\widehat{\rho}_{\mathrm{w}}$ under the orthogonality constraints $\mathrm{w} \perp (\mathrm{w}_1, ..., \mathrm{w}_{i-1})$.

Define q^* the lowest integer q such that $K_{q+1} = K_q$, where K_q is the Krylov subspace given by (11). We have:

$$S\,K_{q^*} = K_{q^*}. \tag{20}$$

This invariance property implies that $\widehat{\beta}^{\text{OLS}}$ belongs to K_{q^*}.

Lemma 2 *The orthogonal subspace of* K_{q^*} *w.r.t. the quadratic form* S *is the same as the orthogonal subspace w.r.t.* I*. Moreover,* $K_{q^*}^{\perp}$ *is included in the null-signal space, that is*

$$\forall \mathrm{w} \in K_{q^*}^{\perp}, \mathrm{w}'\hat{\beta}^{\text{OLS}} = 0.$$

PROOF. Let w be in the orthogonal to K_{q^*} w.r.t. I and w^* be any vector in K_{q^*}, then $\mathrm{w}'S\mathrm{w}^* = \mathrm{w}'\,(S\mathrm{w}^*) = 0$ since $S\mathrm{w}^* \in K_{q^*+1} = K_{q^*}$. Hence w lies in the orthogonal to K_{q^*} w.r.t. S. Consider now $\mathrm{w} \in K_{q^*}^{\perp}$. Then, $\forall \mathrm{w}^* \in K_{q^*}$, $\mathrm{w}'\mathrm{w}^* = 0$. Equation 20 gives $s = S\mathrm{w}^{**}$ for some $\mathrm{w}^{**} \in K_{q^*}$ and then $\mathrm{w}'\hat{\beta}^{\text{OLS}} = \mathrm{w}'S^{-1}s = \mathrm{w}'S^{-1}S\mathrm{w}^{**} = \mathrm{w}'\mathrm{w}^{**} = 0$. □

We can now formulate the theorem that characterizes the sequence $\mathrm{w}_1, ..., \mathrm{w}_p$ and shows its relationship with the Krylov subspaces $K_1, ..., K_{q^*}$.

Theorem 3 *For* $i = 1, ..., p$*, the directions* w_i *can be obtained as follow:*

(1) If $i \leq q^*$*, then* w_i *belongs to* K_i *and is orthogonal to* K_{i-1}*. Moreover,* $\mathrm{w}_i'\hat{\beta}^{\text{OLS}} \neq 0$.

(2) If $i > q^*$*, then* w_i *is any vector orthogonal to* $(\mathrm{w}_1, ..., \mathrm{w}_{i-1})$*. In that case,* $\mathrm{w}_i'\hat{\beta}^{\text{OLS}} = 0$*, i.e.* w_i *belongs to the null-signal space.*

PROOF.

We use a recurrence procedure. It has already been shown that $\mathrm{w}_1 = \mathrm{s}$. We assume that the theorem is true for $k = 1, ..., i-1$. At step i, we seek $\mathbf{w}_i$ that maximizes the signal-noise ratio and that is orthogonal to $\mathrm{w}_1, ..., \mathrm{w}_{i-1}$. By Formula 17, we actually seek a direction. Thus, without loss of generality, we may impose a scale constraint. The choice of the constraint will change the actual w_i but not the direction given by w_i. So, we choose the simplest constraint which is $\mathrm{w}'\, S^{-1}\, \mathrm{w} = 1$. Under this constraint, the signal-noise ratio reduce to $|\mathrm{w}'\hat{\beta}^{\text{OLS}}|$. If w maximizes $|\mathrm{w}'\hat{\beta}^{\text{OLS}}|$ then either w either $-\mathrm{w}$ maximize $\mathrm{w}'\hat{\beta}^{\text{OLS}}a$. So, we want to maximize $\mathrm{w}'\,\hat{\beta}^{\text{OLS}}$ under the constraint $\mathrm{w}'\, S^{-1}\, \mathrm{w} = 1$ and $\mathrm{w}'\, \mathrm{w}_k = 0$ for $k = 1, ..., i-1$.

Case 1: $i \leq q^*$

The maximizing w_i will be a solution to the Lagrange multiplier equation

$$S^{-1}\, \mathrm{s} + \lambda\, S^{-1}\, \mathrm{w}_i + \sum_{k=1}^{i-1} \mu_k\, \mathrm{w}_k = 0. \tag{21}$$

By recurrency, $\mathrm{w}_k \in K_k$ for $k = 1, ..., i-1$, thus $S\, \mathrm{w}_k \in K_{k+1} \subseteq K_i$. Multiplying Equation 21 by S gives:

$$\lambda\, \mathrm{w}_i = -\mathrm{s} - \sum_{k=1}^{i-1} \mu_k\, S\, \mathrm{w}_k. \tag{22}$$

Since $k < q^*$, $K_{k+1} \neq K_k$ and $\mathrm{w}_1, S\, \mathrm{w}_1, ..., S\, \mathrm{w}_{i-1}$ are linearly independent. Hence, λ cannot be equal to 0 and w_i belongs necessarily to K_i. Point 1 is established.

Multiplying Equation 21 by w_i' gives

$$\mathrm{w}_i' S^{-1}\, \mathrm{s} + \lambda \mathrm{w}_i\, S^{-1}\, \mathrm{w}_i = 0. \tag{23}$$

Thus, $\mathrm{w}_i'\hat{\beta}^{\text{OLS}} \neq 0$.

Case 2: $i > q^*$

Since w_i is orthogonal to $K_{q^*} = \text{span}\,\{\mathrm{w}_1, ..., \mathrm{w}_{q^*}\}$, the signal $\mathrm{w}_i'\hat{\beta}^{\text{OLS}} = 0$ according to Lemma 2. Hence, the function to be maximized is null and every w_i in the orthogonal to $\mathrm{w}_1, ..., \mathrm{w}_{i-1}$ is a solution of the maximization problem. Point 2 is established. □

Theorem 3 shows that the sequence $(\mathrm{w}_1, ..., \mathrm{w}_{q^*})$ can be found by a Gram-Schmidt orthogonalization procedure from the nested Krylov subspaces $K_1, ..., K_{q^*}$.

Therefore, it corresponds to Martens' weight vectors sequence. However, in our context, w_i corresponds to a direction for the parameter β or its estimate rather than weight vectors used to define latent variables. Denote

$$W_{(q)} = (\mathrm{w}_1|...|\mathrm{w}_q) \text{ and } \overline{W}_{(q)} = (\mathrm{w}_{q+1}|...|\mathrm{w}_p)\,. \tag{24}$$

We obviously have:

$$\text{Range}\left(\overline{W}_{(q)}\right) = \text{Range}\left(W_{(q)}\right)^{\perp}. \tag{25}$$

Therefore, signal-noise ratios and shrinkage factors of directions in Range $\left(\overline{W}_{(q)}\right)$ are bounded above, i.e.

Proposition 4

$$\forall \mathrm{x} \in \textit{Range}\left(\overline{W}_{(q)}\right), \quad \hat{\rho}_{\mathrm{x}} \leq \hat{\rho}_{\mathrm{w}_{q+1}} \text{ and } \hat{a}^*_{\mathrm{x}} \leq \hat{a}^*_{\mathrm{w}_{q+1}}. \tag{26}$$

4.2 PLS estimators as constrained least squares estimators

We show that PLS estimators are least squares estimators constrained to be null onto the last directions of the sequence exhibited in Section 4.1. By Formula 26, we know that these directions are associated to the lowest shrinkage factors or the lowest signal-noise ratios of the sequence.

Lemma 5 *Let G be a $p \times g$ matrix of rank g $(g \leq p)$ and F a $p \times f$ matrix $(f = p - g)$ of rank f such that $G'F = 0$, then:*

$$\underset{\beta: G'\beta = 0}{\text{ArgMin}} \|\mathrm{y} - X\beta\|^2 = S^{-1}\left(I - P_G^{S^{-1}}\right)(\mathrm{s}) = P_F^S \widehat{\beta}.$$

PROOF. The proof of this Lemma can be found in Seber (1984).

Theorem 6 *For $q \leq q^*$, the PLS estimator $\widehat{\beta}^{\text{PLS}}_{(q)}$ given by (8) is the least squares estimator constrained to have null signal onto the space $\overline{W}_{(q)}$:*

$$\widehat{\beta}^{\text{PLS}}_{(q)} = \underset{\beta: \overline{W}'_{(q)}\beta = 0}{\text{ArgMin}} \|\mathrm{y} - X\beta\|^2. \tag{27}$$

PROOF. By Theorem 3, $\overline{W}'_{(q)} K_q = 0$, $\text{Rank}(K_q) = q$ and $\text{Rank}(\overline{W}_{(q)}) = p - q$, Lemma 5 with $G = \overline{W}_{(q)}$ and $F = K_q$ gives the result. □

Actually, the constraints $\overline{W}'_{(q)}\beta = 0$ do not need to take into account the vectors $\mathrm{w}_{q^*+1}, ..., \mathrm{w}_p$ since they belong to the null-signal space. More precisely, let $\widetilde{W}_{(q)} = (\mathrm{w}_{q+1}, ..., \mathrm{w}_{q^*})$ for $q < q^*$ and $\widetilde{W}_{(q)} = \varnothing$ for $q \geq q^*$, then, forcing the least squares estimator to be zero either on $\overline{W}_{(q)}$ or on $\widetilde{W}_{(q)}$ both gives $\widehat{\beta}^{\text{PLS}}_{(q)}$. First, we give two technical results:

Lemma 7 *Let G_1 be a $p \times g_1$ matrix of rank g_1 and G_2 be a $p \times g_2$ matrix of rank g_2 such that $G_2' S^{-1} G_1 = 0$. Let $G = (G_1|G_2)$. Then,*

$$G_2'\widehat{\beta} = 0 \Longrightarrow \underset{\beta: G'\beta = 0}{\text{ArgMin}} \left\|\mathrm{y} - X\beta\right\|^2 = \underset{\beta: G_1'\beta = 0}{\text{ArgMin}} \left\|\mathrm{y} - X\beta\right\|.^2$$

PROOF. $G_2' S^{-1} G_1 = 0$ and $G_2'\widehat{\beta} = 0$ imply respectively that $P_G^{S^{-1}} = P_{G_1}^{S^{-1}} + P_{G_2}^{S^{-1}}$ and $P_{G_2}^{S^{-1}}(\mathrm{s}) = 0$. Whence, from Lemma 5,

$$\begin{aligned}
\underset{\beta: G'\beta = 0}{\text{ArgMin}} \left\|\mathrm{y} - X\beta\right\|^2 &= S^{-1}\left(I - P_G^{S^{-1}}\right)(\mathrm{s}) \\
&= S^{-1}\left(I - P_{G_1}^{S^{-1}}\right)(\mathrm{s}) \\
&= \underset{\beta: G_1'\beta = 0}{\text{ArgMin}} \left\|\mathrm{y} - X\beta\right\|^2.
\end{aligned}$$

□

Corollary 8 *If $G'\widehat{\beta} = 0$, then $P_G^{S^{-1}}(\mathrm{s}) = 0$ and*

$$\underset{\beta: G'\beta = 0}{\text{ArgMin}} \left\|\mathrm{y} - X\beta\right\|^2 = \widehat{\beta}^{\text{OLS}}.$$

Theorem 9 *We have the following characterizations for $\widehat{\beta}^{\text{PLS}}_{(q)}$:*

1) $\forall q < q^*$,

$$\widehat{\beta}^{\text{PLS}}_{(q)} = \underset{\beta: \widetilde{W}'_{(q)}\beta = 0}{\text{ArgMin}} \left\|\mathrm{y} - X\beta\right\|^2.$$

2) $\forall q \geq q^*$,

$$\widehat{\beta}^{\text{PLS}}_{(q)} = \underset{\beta: \widetilde{W}'_{(q)}\beta = 0}{\text{ArgMin}} \left\|\mathrm{y} - X\beta\right\|^2 = \widehat{\beta}^{ols}.$$

PROOF. From Theorem 3, we have $\overline{W}'_{(q^*)}\widehat{\beta} = 0$. Moreover, if $q < q^*$,

$$\widetilde{W}_{(q)} \subset K_{q^*} \perp \overline{W}_{(q^*)} \Longrightarrow K_{q^*} = S K_{q^*} \perp_{S^{-1}} \overline{W}_{(q^*)}.$$

The conditions of Lemma 7 are fulfilled with $G_1 = \widetilde{W}_{(q)}$, $G_2 = \overline{W}_{(q^*)}$ and $G = \overline{W}_{(q)}$. Then, Theorem 6 and Lemma 7 give the first part of the theorem.

If $q \geq q^*, \overline{W}_{(q)} \subset \overline{W}_{(q^*)}$. Thus, $\overline{W}'_{(q)}\widehat{\beta} = 0$. Theorem 6 and Corollary 8 give $\widehat{\beta}^{\text{PLS}}_{(q)} = \widehat{\beta}_{ols}$. $\square$

Thus, when the underlying space of an estimator on latent variables contains K_{q^*}, this latter estimator coincides with $\widehat{\beta}^{\text{OLS}}$. This follows from the fact that any estimator on latent variables can be obtained by projecting $\widehat{\beta}^{\text{OLS}}$ on the space spanned by its weight vectors. Hence, K_{q^*} indicates the end of the procedure of shrinkage in the sequence of nested Krylov subspaces.

5 A constraint free algorithm for PLSR

In the classical approach of PLSR and the one presented in Section 4, the maximization algorithm involves some arbitrary constraints. In this section, we give another procedure for PLSR with no such arbitrariness. The main idea is to rebuild iteratively $\widehat{\beta}^{\text{OLS}}$ according to a new criterion which is the difference between signals over the noise. At each step, we attempt to bring closer the current estimator to $\widehat{\beta}^{\text{OLS}}$ by adding the direction that maximizes this criterion. This direction is the one where the two estimators differ the most.

The algorithm starts with the null estimator. At step one, we put $\widehat{\beta}^{\text{PLS}}_o = 0$ and we define

$$\Delta\text{SN}_o(\text{x}) = \frac{|\text{x}'\widehat{\beta}^{\text{OLS}} - \text{x}'\widehat{\beta}^{\text{PLS}}_o|}{\sigma\sqrt{\text{x}'S^{-1}\text{x}}}.$$

Then, we seek the direction that maximizes $\Delta\text{SN}_o(\text{x})$. Since in that case, $\Delta\text{SN}_o(\text{x}) = \hat{\rho}_{\text{x}}$, this direction is spanned by $\text{w}_1 = \text{s}$. Now we want to rebuild from 0 the least squares estimator on that direction. So, we have:

$$\underset{\beta \in \text{span}\{\text{w}_1\}}{\text{ArgMin}} \|y - X\beta\|^2 = \widehat{\beta}^{\text{PLS}}_{(1)}. \tag{28}$$

Iteratively, at step $q+1 \leq q^*$, the current estimator is $\hat{\beta}^{\text{PLS}}_q = P^S_{W(q)}\widehat{\beta}^{\text{OLS}}$ where $W_{(q)}$ is defined in Section 4.1. To construct the subspace of dimension $q+1$ which gives $\widehat{\beta}^{\text{PLS}}_{q+1}$, we add to $W_{(q)}$ the direction that maximizes $\Delta\text{SN}_q(\text{x})$, where

$$\Delta\text{SN}_q(\text{x}) = \frac{|\text{x}'\widehat{\beta}^{\text{OLS}} - \text{x}'P^S_{W_{(q)}}\widehat{\beta}^{\text{OLS}}|}{\sigma\sqrt{\text{x}'\,S^{-1}\,\text{x}}} = \frac{|\text{x}'\widehat{\beta}^{\text{OLS}} - \text{x}'\widehat{\beta}^{\text{PLS}}_q|}{\sigma\sqrt{\text{x}'\,S^{-1}\,\text{x}}}. \tag{29}$$

Theorem 10 $\forall q \leq q^* - 1$,

$$\underset{\mathrm{x}\in\mathbb{R}^p}{\mathrm{ArgMax}}\, \Delta\mathrm{SN}_q(\mathrm{x}) = \mathrm{w}_{q+1} \tag{30}$$

where w_{q+1} *is the* $q{+}1^{th}$ *element of the sequence given by Theorem 3. Moreover,*

$$\max_{\mathrm{x}\in\mathbb{R}^p}\ \Delta\mathrm{SN}_q(\mathrm{x}) = \widehat{\rho}_{\mathrm{w}_{q+1}}. \tag{31}$$

PROOF. We have: $P^S_{W_q}\, S^{-1} = S^{-1}\, P^{S^{-1}}_{SW_q}$. By Cauchy-Schwarz's inequality:

$$\Delta\mathrm{SN}_q(\mathrm{x}) = \frac{\left|\mathrm{x}'\, S^{-1}\, \left(\mathrm{s} - \mathrm{P}^{S^{-1}}_{S\,W_q}\,\mathrm{s}\right)\right|}{\sqrt{\mathrm{x}'S^{-1}\mathrm{x}}} \leq \left\|\mathrm{s} - \mathrm{P}^{S^{-1}}_{S\,W_q}\,\mathrm{s}\right\|_{S^{-1}}.$$

Equality holds iff $\mathrm{x} \propto (I - \mathrm{P}^{S^{-1}}_{S\,W_q})\,\mathrm{s}$. Since $\mathrm{Range}(\overline{W}_q)$ is the orthogonal to $\mathrm{Range}(W_q)$ w.r.t. I, $\overline{W}_q$ is the orthogonal to $\mathrm{Range}(S\,W_q)$ w.r.t. S^{-1}. So, $(I - \mathrm{P}^{S^{-1}}_{S\,W_q})\,\mathrm{s} = P^{S^{-1}}_{\overline{W}_q}\,\mathrm{s} \in \mathrm{Range}(\overline{W}_q)$. Thus, $\Delta\mathrm{SN}_q(x)$ attains its maximum on $\mathrm{Range}(\overline{W}_q)$. Since $\Delta\mathrm{SN}_q(x)$ and $\hat{\rho}_x$ coincide on $\mathrm{Range}(\overline{W}_q)$,

$$\max_{\mathrm{x}\in\mathbb{R}^p}\Delta\mathrm{SN}_q(\mathrm{x}) = \max_{\mathrm{x}\in\mathrm{Range}(\overline{W}_q)}\Delta\mathrm{SN}_q(\mathrm{x}) = \max_{\mathrm{x}\in\mathrm{Range}(\overline{W}_q)}\hat{\rho}_{\mathrm{x}} = \hat{\rho}_{\mathrm{w}_{q+1}}.$$

The last equality follows from Theorem 3. □

Thus, $\mathrm{Range}(W_{(q+1)})$ is the subspace that we are looking for, where $W_{q+1} = (w_1|...|w_{q+1})$. The corresponding rebuilt least squares estimator is then $\widehat{\beta}^{\text{PLS}}_{(q+1)}$ since

$$\widehat{\beta}^{\text{PLS}}_{(q+1)} = \underset{\beta\in\mathrm{Range}(W_{(q+1)})}{\mathrm{ArgMin}} \left\|\mathrm{y} - X\beta\right\|^2. \tag{32}$$

Formulas 28 and 32 follows from Formula 27 since $\beta \in \mathrm{Range}(W_{(q+1)})$ is equivalent to $\overline{W}'_{(q+1)}\beta = 0$.

At step q^*, $\widehat{\beta}^{\text{PLS}}_{q^*} = \widehat{\beta}^{\text{OLS}}$ and $\widehat{\rho}_{\mathrm{w}_{q^*+1}} = 0$.

6 Some comments:

The two algorithms presented in Sections 4 and 5 focus on the way PLSR modify the OLS estimator from a directional point of view. The first one is based on the signal-noise ratio $\hat{\rho}_x$. This ratio appears naturally to improve the estimate of $x'\beta$ by applying a shrinkage factor. It plays the same role for PLSR as the variance (or noise) for PCR. The second one is based on $\Delta\mathrm{SN}_q(x)$, which is a measurement of the difference between the current estimator and the OLS

estimator according to x. The interesting feature of these two criteria is that there are scale invariant. Therefore, unlike the latent variables approach, no arbitrary scale constraints are needed .

Both algorithms differ in their structure. In the first one, the maximization procedure of $\hat{\rho}_x$ is carried out to get the whole sequence of directions. Then, the latest directions are used as constraints in the least squares minimization. In the second approach, the alternation of $\Delta \mathrm{SN}_q(x)$ maximization and least squares minimization at each step gives the PLS estimators. This alternation allows to release the orthogonality constraints used in the first algorithm.

Both algorithms differ also in their interpretation. The first one shows PLS estimators as least squares estimators forcing to be null onto the smallest signal-noise ratio directions of the sequence. The second algorithm shed the light on how PLSR reduce the difference in terms of $\Delta \mathrm{SN}_q(x)$ between the current estimator and the OLS estimator.

7 The non full-rank case

We discuss briefly the non full-rank case. If $\mathrm{Rank}(S) = r < p$, the same procedure describe in Section 4 can be applied but need to be adjusted. The OLS estimator $\hat{\beta}^{\text{OLS}}$ is non unique, but if $\mathrm{x}_o'\beta$ is estimable, i.e. if $\mathrm{x}_o \in \mathrm{Range}\, S$, $\mathrm{x}'\hat{\beta}^{\text{OLS}}$ and $\mathrm{var}(\mathrm{x}'\hat{\beta}^{\text{OLS}})$ do not depend on the choice of $\hat{\beta}^{\text{OLS}}$. So, we seek directions $\mathrm{w}_1, ..., \mathrm{w}_r$ in $\mathrm{Range}(S)$ rather than in the whole space. With this restriction, part (a) of Theorem 3 and Theorem 10 are still valid and $\mathrm{w}_1, ..., \mathrm{w}_{q^*}$ are identical to Martens' weight vectors. Denote by $\mathrm{w}_{r+1}, ..., \mathrm{w}_p$ an orthogonal sequence in $\mathrm{Range}(S)^\perp$. For $q \leq r$, we define $\overline{W}_{(q)} = (w_{q+1}|...|w_p)$. It can be shown that for $q < q^*$,

$$\widehat{\beta}^{\text{PLS}}_{(q)} = \underset{\beta : \overline{W}'_{(q)}\beta = 0}{\mathrm{ArgMin}} \, \|\mathrm{y} - X\beta\|^2 \tag{33}$$

and for $q^* \leq q \leq r$:

$$\underset{\beta : \overline{W}'_{(q)}\beta = 0}{\mathrm{ArgMin}} \, \|\mathrm{y} - X\beta\|^2 = \hat{\beta}^{\text{OLS}*}. \tag{34}$$

where $\hat{\beta}^{\text{OLS}*} = S^+ s$ with S^+ the Moore-Penrose inverse of S. The constraints $\mathrm{w}'_{r+1}\beta = 0, ..., \mathrm{w}'_p\beta = 0$ ensures that the estimator belongs to $\mathrm{Range}(S)$. The interpretation of estimators in term of projectors can be easily adapted by using generalized projectors (see Rao and Mitra, 1974) and choosing $\hat{\beta}^{\text{OLS}*}$ as initial OLS estimator.

References

[1] Bibby, J., Toutenburg, H., 1977. Prediction and improved estimation in linear models. John Wiley & Sons.

[2] Breiman, L., Spector, P., 1992. Submodel selection and evaluation in regression. The X-random case. International Statistical Review 60, 291–319.

[3] Brown, P. J., 1993. Measurement, regression, and calibration. Oxford University Press.

[4] Butler, N. A., Denham, M. C., 2000. The peculiar shrinkage properties of partial least squares regression. Journal of the Royal Statistical Society, Series B, Methodologica 62, 585–593.

[5] De Jong, S., 1995. PLS shrinks. Journal of Chemometrics 9, 323–326.

[6] Frank, I. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools (Disc: p136-148). Technometrics 35, 109–135.

[7] Goutis, C., 1996. Partial least squares algorithm yields shrinkage estimators. The Annals of Statistics 24, 816–824.

[8] Helland, I. S., 1988. On the structure of partial least squares regression. Communications in Statistics, Part B – Simulation and Computation [Split from: @J(CommStat)] 17, 581–607.

[9] Helland, I. S., 1990. Partial least squares regression and statistical models. Scandinavian Journal of Statistics 17, 97–114.

[10] Helland, I. S., 1992. Maximum likelihood regression on relevant components. Journal of the Royal Statistical Society, Series B, Methodological 54, 637–647.

[11] Helland, I. S., Almøy, T., 1994. Comparison of prediction methods when only a few components are relevant. Journal of the American Statistical Association 89, 583–591.

[12] Lingjærde, O. C., Christophersen, N., 2000. Shrinkage structure of partial least squares. Scandinavian Journal of Statistics 27 (3), 459–473.

[13] Martens, H., Næs, T., 1989. Multivariate calibration. John Wiley & Sons.

[14] Rao, C., Mitra, S., 1974. Projections under semi-norms and generalized inverse of matrices. Linear Algebra and Applications 9, 155–167.

[15] Seber, G. A. F., 1984. Multivariate observations. John Wiley & Sons.

[16] Stone, M., Brooks, R. J., 1990. Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (Corr: V54 p906-07). Journal of the Royal Statistical Society, Series B, Methodological 52.

[17] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, Methodological 58, 267–288.