A two-step method for linear prediction with connections to PLS

Ying Li and Dietrich von Rosen

Department of Energy and Techology
Swedish University of Agricultural Sciences

LINSTAT, 2014

Outline

- Partial least squares regression
- A two-step method
- MLEs in the two-step method
- Extensions
- Summary

PLS

- Partial least squares originates from a system analysis approach of Herman Wold. As a regression method in the chemometrics field, it has been mainly developed by Svante Wold and Harald Martens.
- Partial least squares regression (PLS) was proposed for situations with collinear explanatory variables where the number of variables often is relatively large.
- PLS is (mostly) an algorithmic approach.
- In order to further develop and improve data analysis, it is of interest to make theoretical contributions.

The overall aim is to connect PLS with linear models.

PLS-algorithm

At every step, according to Helland(1990), the PLS algorithm forms the two linear representations:

$$\mathbf{x} = \boldsymbol{\mu}_x + \mathbf{p}_1 t_1 + \mathbf{p}_2 t_2 + \cdots + \mathbf{p}_a t_a + \mathbf{e}_a,$$

$$y = \mu_y + q_1 t_1 + q_2 t_2 + \cdots + q_a t_a + f_a,$$

where the component t_i is defined as:

$$t_i = \boldsymbol{\omega}_i' \mathbf{e}_{i-1}, \text{ with weights } \boldsymbol{\omega}_i = C[\mathbf{e}_{i-1}, y].$$

Note that $\mathbf{e}_0 = \mathbf{x} - \boldsymbol{\mu}_x$, $\boldsymbol{\omega}_1 = \boldsymbol{\omega}$. The loadings $\mathbf{p}_i$ and q_i are determined by regressing y, $\mathbf{x}$ on t using OLS methodology.

PLS

According to Helland (1988), the population PLS predictor at step a equals

$$\hat{y}_{a,PLS} = \boldsymbol{\omega}'\mathbf{G}_a(\mathbf{G}_a'\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}_a'(\mathbf{x} - \boldsymbol{\mu}_x) + \mu_y, \qquad (1)$$

where $\mathbf{G}_a$ **is the most important quantity**, i.e.

$$\zeta(\mathbf{G}_a) = \zeta(\boldsymbol{\omega}_1 : \boldsymbol{\omega}_2 : \cdots : \boldsymbol{\omega}_a) = \zeta(\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a-1}\boldsymbol{\omega}).$$

The space $\zeta(\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a-1}\boldsymbol{\omega})$ is called a Krylov space.

The above PLS predictor can be viewed to have been obtained from a two-step approach. We assume $\boldsymbol{\mu}_x$, μ_y, ω and $\boldsymbol{\Sigma}$ to be known. First, it is supposed that $x - \boldsymbol{\mu}_x$ follows a linear model:

$$\mathbf{x} - \mu_x \quad = \quad \boldsymbol{\Sigma}\mathbf{G}_a\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $E(\boldsymbol{\varepsilon}) = 0$, $D(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$, $\boldsymbol{\beta}$ is an unknown vector and

$$\mathbf{G}_a = (\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a-1}\boldsymbol{\omega})$$

is a Krylov structured matrix. By the algorithm of PLS, $\zeta(\mathbf{G}_a) \subseteq \zeta(\boldsymbol{\Sigma})$, so the model satisfies the assumption of being a weakly singular Gauss-Markov model.

Then, a least square predictor is given by,

$$\begin{aligned}\widehat{\mathbf{x}-\boldsymbol{\mu}_x} &= \boldsymbol{\Sigma}\mathbf{G}_a(\mathbf{G}_a'\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}_a'\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_x)\\ &= \boldsymbol{\Sigma}\mathbf{G}_a(\mathbf{G}_a'\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}_a'(\mathbf{x}-\boldsymbol{\mu}_x).\end{aligned}$$

It may be of help to view μ_x as a baseline parameter rather than a population mean of the explanatory variables. Hence, PLS is not modeling the residuals of explanatory variables but the mean with adjustment for the baseline. In the second step

$$\hat{y} = \boldsymbol{\omega}'\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{x}-\boldsymbol{\mu}_x}) + \mu_y = \boldsymbol{\omega}'\mathbf{G}_a(\mathbf{G}_a'\boldsymbol{\Sigma}\mathbf{G}_a)^{-}\mathbf{G}_a'(\mathbf{x}-\boldsymbol{\mu}_x) + \mu_y$$

is used and is identical to $\hat{y}_{a,PLS}$ in (1) and is completely free of $\boldsymbol{\Sigma}^{-1}$.

- The idea of assuming a model for x, which also can be regarded as explanatory variables, is not new and fairly reasonable, Especially in situations with almost collinear explanatory variables.
- Often there are a huge number of explanatory variables available. Some of them would jointly mirror the same latent effect and then also influence the response variable, i.e. the explanatory variables x are governed by a latent effect and some random effect.
- How to handle x is an open question. Among others, Stone and Brooks (1990) use potential additional regressors in continuum regression. Helland (1992) proposed an approach with relevant components which is similar to principle component regression.

A two-step method

A two-step method is formulated as follows: let $\boldsymbol{\omega}$ be known,

(i). suppose $\mathbf{x} = \boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N_p(0, \boldsymbol{\Sigma})$ and γ is an unknown parameter,

(ii). predict via the conditional expectation: $\hat{y} = \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}_x}) + \mu_y$.

The major problem is to estimate $\boldsymbol{\Sigma}^{-1}$ in Step (ii)

Based on the Cayley-Hamilton theorem, $\boldsymbol{\Sigma}^{-1}$ can be presented in a polynomial form, i.e.

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{p} c_i \boldsymbol{\Sigma}^{i-1} \approx \sum_{i=1}^{a} c_i \boldsymbol{\Sigma}^{i-1}$$

for some constants c_i and $a \leq p$.

So the model in the first step can be formulated as:

$$\begin{aligned}\mathbf{x} &= \boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon} = \boldsymbol{\Sigma}\sum_{i=1}^{p} c_i\boldsymbol{\Sigma}^{i-1}\boldsymbol{\omega}\gamma + \boldsymbol{\varepsilon} \\ &\approx \sum_{i=1}^{a}\boldsymbol{\Sigma}^{i}\boldsymbol{\omega}(c_i\gamma) + \boldsymbol{\varepsilon} = \boldsymbol{\Sigma}\mathbf{G}_a\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\end{aligned}$$

where $\boldsymbol{\beta} = (\beta_i)$, with $\beta_i = c_i\gamma$, is an unknown vector and

$$\mathbf{G}_a = (\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a-1}\boldsymbol{\omega})$$

is a Krylov structured matrix.

Three types of approximation of $\boldsymbol{\Sigma}^{-1}$

1. $\boldsymbol{\Sigma}^{-1}$ is approximated by removing eigenvectors with small eigenvalues (shrinkage).
2. $\boldsymbol{\Sigma}^{-1}$ is approximated by regularization, i.e. $(\boldsymbol{\Sigma} + k\mathbf{I})^{-1}$ (ridge method).
3. $\boldsymbol{\Sigma}^{-1}$ is approximated by a Krylov sequence (PLS).

- In the sample version of PLS, before applying the algorithm, the parameters are usually replaced by unbiased estimators.
- In our approach, a $semi - population$ PLS version is introduced, i.e. we will assume the covariance between the response variable and explanatory variables to be known as well as the mean of the response variable, whereas, the other parameters, i.e. the mean and variance for the explanatory variable are unknown and hence should be estimated.
- The object is to estimate the unknown parameters in a linear model for the explanatory variables.

The two-step method and Data

In the first step, a design matrix $\mathbf{A}$(function of $\boldsymbol{\Sigma}$) is determined, such that

$$\mathbf{X} = \mathbf{A}\boldsymbol{\beta}\mathbf{1}_n' + \mathbf{E}. \quad \mathbf{E} \sim N_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_n), \tag{2}$$

where $\mathbf{1}_n'$: $1 \times n$ is a vector of n 1s, $\mathbf{A} = \boldsymbol{\Sigma}\mathbf{G}_a$, $\mathbf{G}_a$ is the Krylov matrix used previously, $\boldsymbol{\Sigma}$: $p \times p$ is *p.d.*, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are unknown. In the second step, y is determined via the conditional mean

$$\hat{\mathbf{y}} = \boldsymbol{\omega}'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \hat{\boldsymbol{\mu}}_x\mathbf{1}_0') + \boldsymbol{\mu}_y$$

MLEs in the two-step method

Theorem

Let the model be given by (2) and suppose that ω *in* $\mathbf{A}$ *is known, where* $\mathbf{A} = \mathbf{\Sigma G_a} = (\mathbf{\Sigma}\omega, \mathbf{\Sigma^2}\omega, \ldots, \mathbf{\Sigma^a}\omega)$, *and* $\mathbf{S} = \mathbf{X}(\mathbf{I} - \mathbf{1}_n\mathbf{1}_n' n^{-1})\mathbf{X}'$. *Then, if* $n > p$, *the maximum likelihood estimators of* $\mathbf{\Sigma}$ *and* $\mathbf{A}\beta$ *are given by*

$$\widehat{\mathbf{A}\beta} = \hat{\mathbf{A}}(\hat{\mathbf{A}}'\mathbf{S}^{-\mathbf{1}}\hat{\mathbf{A}})^{-\mathbf{1}}\hat{\mathbf{A}}'\mathbf{S}^{-\mathbf{1}}\mathbf{X}\mathbf{1}_n n^{-1},$$

$$\hat{\mathbf{A}} = (\frac{1}{n}\mathbf{S}\omega, \frac{1}{n^2}\mathbf{S}^2\omega, \cdots, \frac{1}{n^a}\mathbf{S}^a\omega),$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\{\mathbf{S}+(\mathbf{I}-\hat{\mathbf{A}}(\hat{\mathbf{A}}'\mathbf{S}^{-1}\hat{\mathbf{A}})^{-}\hat{\mathbf{A}}'\mathbf{S}^{-1})\mathbf{X}\mathbf{1}_n\mathbf{1}_n' n^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{S}^{-1}\hat{\mathbf{A}}(\hat{\mathbf{A}}'\mathbf{S}^{-1}\hat{\mathbf{A}})^{-}$$

Numerical illustration

- The data set from Fearn (1983) is a classical data set with collinear variables.
- It consists of two sets, one with 24 and one with 26 wheat samples. The protein content y in wheat is the response variable and the log values of reflectance, obtained from near infrared measurements (NIR) at 6 wavelengths $L_1 - L_6$, are the predictors.
- Following Hoerl's (1985) procedure, we use the first 12 samples to perform the regression analysis and validate the prediction by the other 12 samples and the 26 samples of the other set.

Numerical illustration

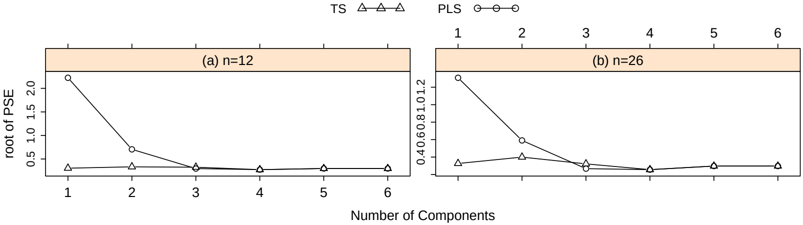
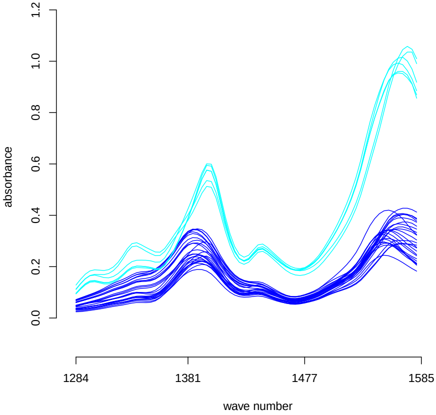


Figure: The root of the mean squared error of prediction (PSE) for PLS and the two-step method (TS) for Fearn's (1983) data: *(a): the prediction of the 12 samples of the first data set, (b): the prediction of the 26 samples of the second data set*.

Group effect



Two-step method for the grouped data

The model in the first step would be:

$$\mathbf{X} = \mathbf{ABC} + \mathbf{E}, \quad \mathbf{E} \sim N_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_n),$$

with $\mathbf{X}$: $p \times n$, $\mathbf{A} = \boldsymbol{\Sigma}\mathbf{G}_a$: $p \times q$, $\mathbf{G}_a = (\boldsymbol{\omega} : \boldsymbol{\Sigma}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a-1}\boldsymbol{\omega})$ is the Krylov matrix, $\mathbf{B}$: $q \times k$, $\mathbf{C}$: $k \times n$, k is the number of groups. For example, if there are 3 groups with 3, 5 and 4 observations in each group, we may then choose

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix},$$

if $k = 1$ then $\mathbf{C} = \mathbf{1}'_n$. The matrix $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are unknown and should be estimated. In the second step, the response $\mathbf{y}$ will be predicted using the the estimators from first step, i.e.:

$$\hat{\mathbf{y}}' = \boldsymbol{\omega}'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X} - \widehat{\boldsymbol{\mu}}_x\mathbf{C}) + \boldsymbol{\mu}'_y, \quad \widehat{\boldsymbol{\mu}}_x = \widehat{\mathbf{AB}},$$

MLE in the two-step model for group data

Theorem

Suppose that ω *in* $\mathbf{A}$ *is known, where* $\mathbf{A} = \mathbf{\Sigma G}_a = (\mathbf{\Sigma}\omega, \mathbf{\Sigma}^2\omega, \ldots, \mathbf{\Sigma}^a\omega)$ *and* $\mathbf{S} = \mathbf{X}(\mathbf{I} - \mathbf{P}_{c'})\mathbf{X}'$, *where* $\mathbf{P}_{\mathbf{c}'} = \mathbf{C}'(\mathbf{CC}')^{-}\mathbf{C}$. *Then, if* $n > p$, *the maximum likelihood estimators of* $\mathbf{\Sigma}$ *and* $\mathbf{AB}$ *are given by*

$$\widehat{\boldsymbol{AB}} = \boldsymbol{\hat{A}}(\boldsymbol{\hat{A}}'\boldsymbol{S}^{-1}\boldsymbol{\hat{A}})^{-}\boldsymbol{\hat{A}}'\boldsymbol{S}^{-1}\mathbf{XC}'(\mathbf{CC}')^{-},$$

$$\widehat{\mathbf{A}} = (\frac{1}{n}\mathbf{S}\boldsymbol{\omega}, \frac{1}{n^2}\mathbf{S}^2\boldsymbol{\omega}, \cdots, \frac{1}{n^a}\mathbf{S}^a\boldsymbol{\omega}),$$

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\{\mathbf{S} + (\mathbf{I} - \boldsymbol{\hat{A}}(\boldsymbol{\hat{A}}'\boldsymbol{S}^{-1}\boldsymbol{\hat{A}})^{-}\boldsymbol{\hat{A}}'\boldsymbol{S}^{-1})\mathbf{XP}_{\mathbf{c}'}\mathbf{X}'$$
$$\times(\mathbf{I} - \boldsymbol{S}^{-1}\boldsymbol{\hat{A}}(\boldsymbol{\hat{A}}'\boldsymbol{S}^{-1}\boldsymbol{\hat{A}})^{-}\boldsymbol{\hat{A}}')\}.$$

Numerical illustration

A simulation study comparing the relative performance of PLS, the two-step method(TS), ridge regression (Ridge) and the lasso regression (Lasso):

- In all cases, $N = 100$ for the training-sample and $N_t = 20$ for the test-sample.
- The simulations mirror the following situations: group effect (4 groups) or no group effect (i.e. for a single group data); low (all off-diagonal elements in correlation matrix were randomly chosen between 0 and 1) or high-collinear $\mathbf{x}$ structure (all off-diagonal elements were chosen to be larger than 0.9).

Numerical illustration

For each situation, 100 repetitions of the following procedure were performed.

1. Randomly generate $N = 100$ training observations with a joint normal distribution with specified population parameters.
2. Apply PLS and TS, for $a = 1, 2, \cdots, p$, Ridge and Lasso to the training-sample.
3. Compute the mean squared error (MSE) of the training observations.
4. Generate $N_t = 20$ independent test observations from 1.
5. Compute the mean of the predicted squared error (PSE) of the test observations.

Numerical illustration

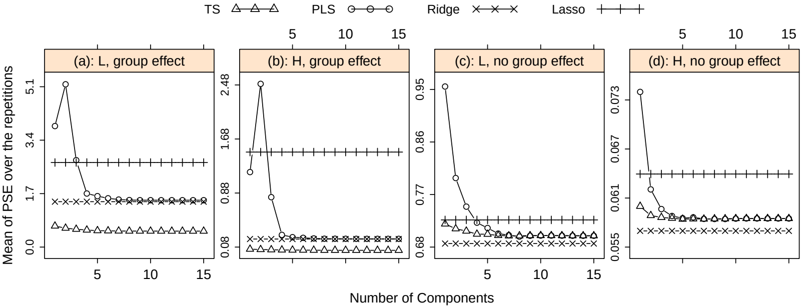


Figure: Simulation results reflecting prediction error for PLS, TS, Lasso and Ridge with 15 explanatory variables: *for the cases with (a) low-collinear and group data, (b) high-collinear and group data, (c) low-collinear and no group data, (d) high-collinear and no group data.*

Grouped data with a nested mean structure

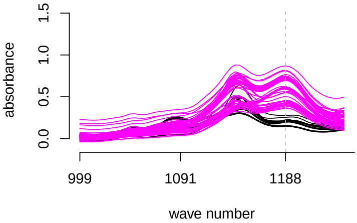
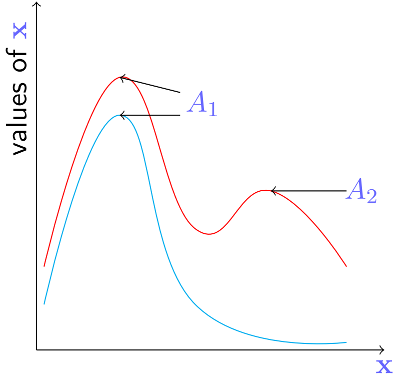


Figure: Spectra of the silage samples from two different experiments. (*Each experiment is represented by in one colour. There is an extra peak in one experiment as indicated by the dashed line.*)



Model in the first step:

$$\mathbf{X} = \mathbf{A}_1\mathbf{B}_1\mathbf{C}_1 + \mathbf{A}_2\mathbf{B}_2\mathbf{C}_2 + \mathbf{E},$$

$$\mathbf{E} \sim N_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_n)$$

where

$$\mathbf{A}_1 = (\boldsymbol{\Sigma}\boldsymbol{\omega} : \boldsymbol{\Sigma}^2\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a_1}\boldsymbol{\omega})$$

$$\mathbf{A}_2 = (\boldsymbol{\Sigma}^{a_1+1}\boldsymbol{\omega} : \cdots : \boldsymbol{\Sigma}^{a_1+a_2}\boldsymbol{\omega})$$

Summary

- **PLS has been formulated as a multivariate linear model.**
- **A new two-step method for linear predictions has been proposed.**
- **The two-step method has been extended to handle data:**
  - with group effect.
  - with a nested mean structure.
- **The explicit maximum likelihood estimators have been derived for the two-step method and its extensions.**

Future works

- Stopping rules, i.e. how many terms should be included in $\mathbf{A} = \Sigma\mathbf{G}_a$ is one of the most interesting questions should be investigated in the future.
- Multi-response prediction i.e. there is a multivariate response variable which should be predicted.

Reference

1. Inge S Helland (1990), Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, **17**, 97–114.
2. Ying Li and Dietrich von Rosen, (2012), Maximum likelihood estimators in a two step model for PLS. *Communications in Statistics - Theory and Methods*, **41**, 2503-2511.
3. Ying Li, Peter Udén and Dietrich von Rosen, (2013), A two-step PLS-inspired method for linear prediction with group effect. *Sankhyā A*, **75**, 96-117.
4. Ying Li, Peter Udén and Dietrich von Rosen, (2014), A two-step method for group data with connections to extended growth model and partial least squares regression. *submitted*.