Partial Least Squares A new statistical insight through orthogonal polynomials.

Mélanie Blazère
Institut de mathématiques de Toulouse
University Paul Sabatier

INSTITUT de MATHEMATIQUES de TOULOUSE

Work supervised by Fabrice Gamboa and Jean-Michel Loubes

19th European Young Statisticians Meeting, Prague, September 2

Outline

Overall framework

- **Linear regression model**

$$Y = X\beta^{*} + \varepsilon$$

Overall framework

- **Linear regression model**

$$Y = X\beta^* + \varepsilon$$

where

- $Y = (Y_1, ..., Y_n)^T \in \mathbb{R}^n$ is the response.
- $X = (X_{ij})_{i,j} \in \mathbb{M}_{n\times p}$ is the design matrix.
- $\beta^* = (\beta_1^*, ..., \beta_p^*)^T \in \mathbb{R}^p$ is the target parameter vector.
- $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T \in \mathbb{R}^n$ are unobservable i.i.d random variables which capture the noise.

Overall framework

- **Linear regression model**

$$\boxed{Y = X\beta^* + \varepsilon}$$

where

  - $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ is the response.
  - $X = (X_{ij})_{i,j} \in \mathbb{M}_{n\times p}$ is the design matrix.
  - $\beta^* = (\beta_1^*, \ldots, \beta_p^*)^T \in \mathbb{R}^p$ is the target parameter vector.
  - $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ are unobservable i.i.d random variables which capture the noise.

- **Notation and assumptions**
  - We allow $\mathbf{p} > \mathbf{n}$.
  - We denote by **r the rank of $\mathbf{X^T X}$**.

- **Goal :** to estimate β^* for future prediction.

A useful tool : Singular Value Decomposition

- **SVD of X** given by

$$\boxed{X = UDV^T}$$

where

- $U = (u_1, ..., u_n) \in \mathbb{M}_{n,n}$ and $U^T U = UU^T = I$.
- $V = (v_1, ..., v_p) \in \mathbb{M}_{p,p}$ and $V^T V = VV^T = I$.
- $D \in \mathbb{M}_{n,p}$ contains $(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r})$ on the diagonal and zero anywhere else.

A useful tool : Singular Value Decomposition

- **SVD of X** given by

$$\boxed{X = UDV^T}$$

where

  - $U = (u_1, ..., u_n) \in \mathbb{M}_{n,n}$ and $U^T U = UU^T = I$.
  - $V = (v_1, ..., v_p) \in \mathbb{M}_{p,p}$ and $V^T V = VV^T = I$.
  - $D \in \mathbb{M}_{n,p}$ contains $(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r})$ on the diagonal and zero anywhere else.

- **Assumptions**
We assume that $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_r > 0$.

- **Notations**
Two important quantities :
➟ $\mathbf{p_i} = (\mathbf{X}\beta^*)^{\mathbf{T}}\mathbf{u_i}, \; i = 1, ..., n.$
➟ $\hat{\mathbf{p}}_\mathbf{i} = \mathbf{Y}^{\mathbf{T}}\mathbf{u_i}, \; i = 1, ..., n.$

Limits of the OLS

- **Ordinary least squares**

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y = \sum_{i=1}^{n} \frac{\hat{p}_i}{\sqrt{\lambda_i}} v_i.$$

- **Limits** when some covariates are **nearly collinear**, some λ_i are small
  $\Rightarrow$ high variance of the estimator
  $\Rightarrow$ unstability and unaccurate predictions.
- **Solution** : **regularization** of the LS solution to decrease the variance.
  $\Rightarrow$ penalization method (Ridge, Lasso,...)
  $\Rightarrow$ dimension reduction method (PCR, PLS,..)

What is PLS ?

- **Main idea behind PLS**

  ➤ **The PLS method** at step k (where $k \leq r$) consists in finding $(w_l)_{1 \leq l \leq k}$ that maximize

$$[\mathrm{Cov}(Y, Xw_l)]^2 = \mathrm{Var}(Y)\mathrm{Var}(Xw_l)\mathrm{Cor}(Y, Xw_l)$$

  under the constraints

  - $\|w_l\|^2 = 1$
  - $t_l = Xw_l$ is orthogonal to $t_1, ..., t_{l-1}$.

  ➤ **Field of application :** biomedecines, chemical engineering...

Some references

☞**Helland** (2001), Some theoretical aspects of partial least squares regression , *Chemometrics and Intelligent Laboratory systems*, 58,97–107.

☞**Rosipal R. and Kramer N.** (2006), Overview and recent advances in partial least squares, *Subspace, Latent Structure and Feature selection*, 34–51, Springer.

PLS estimator and link with Krylov subspaces

- **Linear regression of Y onto $t_1, ..., t_k$**
  Define W_K the matrix whose columns are the $(w_k)_{1 \leq k \leq K}$.

The PLS estimator

$$\hat{\beta}_K^{PLS} = W_K({W_K}^T \Sigma W_K)^{-1} {W_K}^T X^T Y$$

- **Link with Krylov subspaces**

Link with Krylov subspaces

$$Span\left\{w_1, ..., w_K\right\} = Span\left\{X^T Y, (X^T X) X^T Y, ..., (X^T X)^{K-1} X^T Y\right\}.$$

The space spanned by $X^T Y, (X^T X) X^T Y, ..., (X^T X)^{K-1} X^T Y$ is called the K^{th} **Krylov subspace** with respect to $X^T X$ and $X^T Y$

PLS= LS on Krylov subspaces

- **PLS is the minimization of least squares over some Krylov subspaces.**

Link between PLS and Krylov subspaces [Helland]

Proposition :

$$\hat{\beta}_k^{PLS} = \underset{\beta \in \mathcal{K}^k(X^TX, X^TY)}{\operatorname{argmin}} \|Y - X\beta\|^2$$

where $\mathcal{K}^k(X^TX, X^TY) = \{X^TY, (X^TX)X^TY, ..., (X^TX)^{k-1}X^TY\}$.

PLS= LS on Krylov subspaces

- **PLS is the minimization of least squares over some Krylov subspaces.**

Link between PLS and Krylov subspaces [Helland]

Proposition :

$$\hat{\beta}_k^{PLS} = \underset{\beta \in \mathcal{K}^k(X^TX, X^TY)}{\operatorname{argmin}} \|Y - X\beta\|^2$$

where $\mathcal{K}^k(X^TX, X^TY) = \{X^TY, (X^TX)X^TY, ..., (X^TX)^{k-1}X^TY\}$.

- **Be careful : the constraints are random !**
  - Contrary to PCR, the PLS linear constraints are **random**.

Some references

☞ **Helland I.S.** (1988), On the structure of partial least squares regression, *Communication in statistics-Simulation and Computation*,17, 581-607.

Minimization over polynomials

✍**Blazere, M., Gamboa, F., Loubes, J. M.** (2014), PLS : a new statistical insight through the prism of orthogonal polynomials, *arXiv preprint* , arXiv :1405.5900.

- **Notation :** $\mathcal{P}_k = \mathbb{R}_k[X]$ and by $\mathcal{P}_{k,1} = \{P \in \mathcal{P}_k ; P(0) = 1\}$.
- **Another point of view**

Optimization over polynomial spaces

Proposition : For $k \leq r$ we have $\hat{\beta}_k = \hat{P}_k(X^T X)X^T Y$ where

$$\hat{P}_k \in \underset{P \in \mathcal{P}_{k-1}}{\operatorname{argmin}} \|Y - XP(X^T X)X^T Y\|^2$$

and $\|Y - X\hat{\beta}_k\|^2 = \|\hat{Q}_k(XX^T)Y\|^2$ where

$$\hat{Q}_k(t) = 1 - t\hat{P}_k(t) \in \underset{Q \in \mathcal{P}_{k,1}}{\operatorname{argmin}} \|Q(XX^T)Y\|^2.$$

Minimization over polynomials

✍**Blazere, M., Gamboa, F., Loubes, J. M.** (2014), PLS : a new statistical insight through the prism of orthogonal polynomials, *arXiv preprint* , arXiv :1405.5900.

- **Notation :** $\mathcal{P}_k = \mathbb{R}_k[X]$ and by $\mathcal{P}_{k,1} = \{P \in \mathcal{P}_k ; P(0) = 1\}$.
- **Another point of view**

Optimization over polynomial spaces

Proposition : For $k \leq r$ we have $\hat{\beta}_k = \hat{P}_k(X^T X)X^T Y$ where

$$\hat{P}_k \in \underset{P \in \mathcal{P}_{k-1}}{\operatorname{argmin}} \|Y - XP(X^T X)X^T Y\|^2$$

and $\|Y - X\hat{\beta}_k\|^2 = \|\hat{Q}_k(XX^T)Y\|^2$ where

$$\hat{Q}_k(t) = 1 - t\hat{P}_k(t) \in \underset{Q \in \mathcal{P}_{k,1}}{\operatorname{argmin}} \|Q(XX^T)Y\|^2.$$

- **PLS= regularization** by **polynomials approximation**

Minimization over polynomials

✍Blazere, M., Gamboa, F., Loubes, J. M. (2014), PLS : a new statistical insight through the prism of orthogonal polynomials, *arXiv preprint* , arXiv :1405.5900.

- **Notation :** $\mathcal{P}_k = \mathbb{R}_k[X]$ and by $\mathcal{P}_{k,1} = \{P \in \mathcal{P}_k ; P(0) = 1\}$.
- **Another point of view**

Optimization over polynomial spaces

Proposition : For $k \leq r$ we have $\hat{\beta}_k = \hat{P}_k(X^T X)X^T Y$ where

$$\hat{P}_k \in \underset{P \in \mathcal{P}_{k-1}}{\operatorname{argmin}} \|Y - XP(X^T X)X^T Y\|^2$$

and $\|Y - X\hat{\beta}_k\|^2 = \|\hat{Q}_k(XX^T)Y\|^2$ where

$$\hat{Q}_k(t) = 1 - t\hat{P}_k(t) \in \underset{Q \in \mathcal{P}_{k,1}}{\operatorname{argmin}} \|Q(XX^T)Y\|^2.$$

- **PLS= regularization** by **polynomials approximation**
- **Key idea= Cayley-Hamilton** theorem

The residuals polynomials

- **Definition**
  The polynomials $\hat{Q}_k$ are called the **residual polynomials**.
- **Interest of the residual polynomials**
  Most PLS objects can be written in terms of the residual polynomials.

Dependance of the PLS objects on the residual polynomials

- $\hat{\beta}_k = \hat{P}_k(X^TX)X^TY = \sum_{i=1}^r \left(1 - \hat{Q}_k(\lambda_i)\right) \frac{\hat{p}_i}{\sqrt{\lambda_i}} v_i.$

  $\Rightarrow$ PLS estimator= **shrinkage estimator** with **filter factor**$=1 - \hat{Q}_k(\lambda_i)$
- $X\hat{\beta}_k = (I - \hat{Q}_k(XX^T))Y = \sum_{i=1}^r \left(1 - \hat{Q}_k(\lambda_i)\right) \hat{p}_i u_i.$
- $Y - X\hat{\beta}_k = \hat{Q}_k(XX^T)Y = \sum_{i=1}^r \hat{Q}_k(\lambda_i)\hat{p}_i u_i + \begin{cases} 0 & \text{if} \quad r = n \\ \sum_{i=r+1}^n \hat{p}_i^2 & \text{if} \quad r < n \end{cases}.$

Residual polynomials= Discrete orthogonal polynomials

- **Discrete measure associated to $(\hat{Q}_k)_{1 \leq k \leq r}$**

Discrete orthogonal polynomials

Proposition :
$\hat{Q}_0 := 1, \hat{Q}_1, \ldots, \hat{Q}_r$ is a sequence of orthonormal polynomials with respect to the measure

$$d\hat{\mu} = \sum_{i=1}^{r} \lambda_i \hat{p}_i^2 \delta_{\lambda_i},$$

where we recall that $\hat{p}_i := u_i^T Y$.

Main result

- **An explicit analytical expression**
  Let $k \leq r$ and $I_k^+ = \{(j_1, ..., j_k) : r \geq j_1 > ... > j_k \geq 1\}$.

Expression for the residuals polynomials

$$\hat{Q}_k(x) = \sum_{(j_1,..,j_k)\in I_k^+} \hat{w}_{(j_1,..,j_k)} \prod_{l=1}^{k} (1 - \frac{x}{\lambda_{j_l}}).$$

where

Definition of the weights

$$\hat{w}_{j_1,..,j_k} := \frac{\hat{p}_{j_1}^2 \dots \hat{p}_{j_k}^2 \lambda_{j_1}^2 \dots \lambda_{j_k}^2 V(\lambda_{j_1}, ..., \lambda_{j_k})^2}{\sum_{(j_1,..,j_k)\in I_k^+} \hat{p}_{j_1}^2 \dots \hat{p}_{j_k}^2 \lambda_{j_1}^2 \dots \lambda_{j_k}^2 V(\lambda_{j_1}, ..., \lambda_{j_k})^2}.$$

with $V(\lambda_{j_1}, ..., \lambda_{j_k}) =$ Vandermonde determinant of $\lambda_{j_1}, ..., \lambda_{j_k}$
and $\hat{p}_{j_k} = Y^T u_{j_k}$.

A new insight on PLS

$$\hat{Q}_k(x) = \sum_{(j_1,..,j_k)\in I_k^+} \hat{w}_{(j_1,..,j_k)} \prod_{l=1}^{k}(1 - \frac{x}{\lambda_{j_l}})$$

- **Interest**
  - Expression depends explicitly on **the observations noise** and on **the eigenelements of X**
  - Contains all the information.
- **Weigths**
  Notice that $0 < \hat{w}_{(j_1,..,j_k)} \leq 1$ and $\sum_{(j_1,..,j_k)\in I_k^+} \hat{w}_{(j_1,..,j_k)} = 1$.
  Be careful : the weights are **random**
- **Interpretation**
  Residual polynomial $\hat{Q}_k$= **convex combinaison** of all the polynomials in $\mathcal{P}_{k,1}$ whose roots are subsets of $\{\lambda_1, ..., \lambda_n\}$.

PLS statistical properties

Upper bound for the empirical risk

An upper bound for the empirical risk

$$\| Y - X\hat{\beta}_k \|^2 \leq \sum_{i=k+1}^{r} \left[\prod_{l=1}^{k} \left(1 - \frac{\lambda_i}{\lambda_l}\right)^2 \hat{p}_i^2 \right] + \sum_{i=r+1}^{n} \hat{p}_i^2 .$$

Notice that if $\frac{\lambda_r}{\lambda_k} > 1 - \delta$ then $\sum_{i=k+1}^{r} \left[\prod_{l=1}^{k} \left(1 - \frac{\lambda_i}{\lambda_l}\right)^2 \hat{p}_i^2 \right] \leq \delta \sum_{i=k+1}^{r} \hat{p}_i^2$.

In particular, $\| Y - X\hat{\beta}_k \|^2 \leq \sum_{i=k+1}^{n} \hat{p}_i^2 := \| Y - X\hat{\beta}_{PCR}^k \|^2$.

Corollary

Let $(\varepsilon_i)_{1 \leq i \leq n}$ be i.i.d centered random variables with commmon variance σ^2.

$$\mathbb{E}\left(\frac{1}{n} \| Y - X\hat{\beta}_k \|^2\right) \leq$$

$$\frac{1}{n}\left(1 - \frac{\lambda_n}{\lambda_1}\right)^{2k} \left[\sum_{i=k+1}^{r} \lambda_i (\beta_i^*)^2 + (r-k)\sigma^2 \right] + \frac{1}{n} \sum_{i=r+1}^{n} \left(\lambda_i (\beta_i^*)^2 + \sigma^2 \right)$$

A new insight onto the PLS filter factors

PLS= **SHRINKAGE ESTIMATOR**

$$\hat{\beta}_k = \sum_{i=1}^{r}(1 - \hat{Q}_k(\lambda_i))\frac{\hat{p}_i}{\sqrt{\lambda_i}}v_i.$$

- **New expression for the PLS filter factor**

$$f_i^{(k)} := 1 - \hat{Q}_k(\lambda_i) = \sum_{(j_1,..,j_k)\in I_k^+} \hat{w}_{(j_1,..,j_k)}\left[1 - \prod_{l=1}^{k}(1 - \frac{\lambda_i}{\lambda_{j_l}})\right]$$

- **Interest**
  It clearly and explicitely shows how the filter factors depend on the error terms and on the eigenelements of X.
  We easily recover that
  - The PLS filter factors are not always in $[0, 1]$.
  - They oscillate below and above one.

PLS statistical properties

Mean Square Prediction Error

Blazere, M., Gamboa, F., Loubes, J. M. (2014), A unified framework for the study of the PLS estimator's properties, *arXiv preprint* , arXiv :1411.0229.

- **Definition**
  The Mean Square Prediction Error (MSPE) is defined by

$$MSPE(\hat{\beta}_k) := \mathbb{E}\left[\| X(\beta^* - \hat{\beta}_k) \|^2\right].$$

  - **Question :** Is the PLS factors not in $[0,1]$ a problem ?
  - **Answer :**

Decomposition of the MSPE

$$\| X\beta^* - X\hat{\beta}_k \|^2 = \sum_{i=1}^{r} \hat{Q}_k(\lambda_i) p_i^2 + \sum_{i=1}^{r} \left(1 - \hat{Q}_k(\lambda_i)\right) \varepsilon_i^2.$$

➠A filter factor larger than one not necessarily implies an increase of the MSE

PLS always shrinks for some specific directions

- PLS shrinks OLS in some of the eigenvectors directions but also **expands** in others.
- However PLS **globally shrinks** the OLS i.e. $\| \hat{\beta}_{k-1} \|^2 \leq \| \hat{\beta}_k \|^2 \leq \| \hat{\beta}_{LS} \|^2$.
- For all $0 \leq l \leq r$, let $\hat{s}_l = \sum_{i=1}^{r} \sqrt{\lambda_i} \hat{Q}_l(\lambda_i) \hat{p}_i v_i$. We have

$$\hat{\beta}_{LS} = \sum_{l=0}^{r-1} \left(\sum_{i=1}^{r} \hat{Q}_l(\lambda_i) \hat{p}_i^2 \right) \frac{\hat{s}_l}{\| \hat{s}_l \|^2}$$

and

$$\hat{\beta}_k = \sum_{l=0}^{k-1} \left(\sum_{i=1}^{r} (\hat{Q}_l(\lambda_i) - \hat{Q}_k(\lambda_i)) \hat{p}_i^2 \right) \frac{\hat{s}_l}{\| \hat{s}_l \|^2}.$$

But

$$0 \leq \sum_{i=1}^{r} (\hat{Q}_l(\lambda_i) - \hat{Q}_k(\lambda_i)) \hat{p}_i^2 \leq \sum_{i=1}^{r} \hat{Q}_l(\lambda_i) \hat{p}_i^2 .$$

➠PLS **always shrinks the OLS in the $\hat{s}_l$ directions**.

Conclusion

- We have proposed a **new approach** to study PLS
- We have established **exact analytical expressions** for the main PLS objects (filter factors, empirical risk, MSPE)
- This approach is useful to provide new interpretations, to shed lights on the behaviour of PLS and to **prove important properties** of the PLS
- This approach provides a **unified framework** to recover well known properties of the PLS estimator
- But this is not the end of the road.
  The expression of the residuals should be explored further to completely understand the PLS method.

Thank you for your attention

References

[1] **Blazere, M., Gamboa, F., Loubes, J. M.** (2014), PLS : a new statistical insight through the prism of orthogonal polynomials, *arXiv preprint* , arXiv :1405.5900.

[2] **Blazere, M., Gamboa, F., Loubes, J. M.** (2014), A unified framework for the study of the PLS estimator's properties, *arXiv preprint* , arXiv :1411.0229.

[3] **Butler N.A and Denham M.C** (2000), The peculiar shrinkage properties of partial least squares , *Journal of the Royal Statistical Society : Series B*, 62(3),585-593.

[6] **Helland** (2001), Some theoretical aspects of partial least squares regression , *Chemometrics and Intelligent Laboratory systems*, 58,97–107.

[7] **Kramer N.** (2007), An overview on the shrinkage properties in partial least squares, *Subspace, Latent Structure and Feature selection*, 34–51, Springer.

[8] **Lingjaerde O.C and Christophersen N.** (2000), Shrinkage structure of partial least squares , *Scandinavian Journal of Statistics*, 27,249–273.

[10] **Phatak A. and de Hoog F.** (2002), Exploiting the connections between pls, lanczos methods and conjugate gradients, *Journal of Chemometrics*, 16(7), 361–367.

[11] **Rosipal R. and Kramer N.** (2006), Overview and recent advances in partial least squares, *Subspace, Latent Structure and Feature selection*, 34–51,