Partial Least Squares: une nouvelle approche au travers de polynômes orthogonaux

Mélanie Blazère [1]& Fabrice Gamboa [1] & Jean-Michel Loubes [1]

[1] *Université Paul Sabatier, Institut de Mathématiques de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, melanie.blazere@math.univ-toulouse.fr*

Résumé. La méthode des moindres carrés partiels aussi appelée PLS est très utilisée pour la prédiction en régression multivariée, notamment lorsque l'on a de fortes corrélations au sein des variables explicatives ou lorsque ces dernières dépassent en nombre les observations que l'on a à disposition. La PLS est une méthode de réduction de dimension astucieuse qui cherche à résoudre le problème de multicollinéarité en créant de nouvelles variables latentes qui maximisent la variance des variables initiales tout en restant optimales pour la prédiction. Si la PLS se révèle être un outil très utile et puissant dans de nombreux domaines (plus particulièrement en génétique et en génie chimique) elle n'en reste pas moins une procédure complexe et peu de ses propriétés théoriques sont connues. Dans cet exposé, je vous présenterai une nouvelle façon de considérer la PLS basée sur les liens étroits qu'elle a avec des polynômes orthogonaux particuliers que j'expliciterai et que nous appellerons par la suite polynômes résiduels. La théorie des polynômes orthogonaux nous permet ensuite de donner une expression analytique explicite pour ces polynômes résiduels. Nous verrons que cette expression montre clairement de quelle façon l'estimateur PLS dépend du signal et du bruit. A la suite de quoi, nous montrerons la puissance de cette nouvelle approche dans l'analyse des propriétés statistiques de la PLS en établissant de nouveaux résultats sur son risque empirique et son erreur quadratique moyenne de prédiction. Nous évoquerons aussi certaines propriétés de seuillage de cet estimateur. Nous conclurons enfin en montrant comment l'approche par polynômes orthogonaux fournit un cadre unifié permettant de retrouver directement des propriétés déja connues mais démontrées par des approches diverses et différentes de la notre.

Mots-clés. Régression PLS, moindres carrés contraints, polynômes orthogonaux, risque empirique, erreur quadratique moyenne de prédiction, estimateur de seuillage.

Abstract. Partial Least Square (PLS) is nowadays a widely used dimension reduction technique in multivariate regression, especially when the explanatory variables are highly collinear or when they outnumber the observations. Originally designed to remove the problem of multicollinearity in the set of explanatory variables, PLS acts as a dimension reduction method by creating orthogonal latent components which maximize the variance and are also optimal for predicting the output variable. If the PLS method is very helpful in a large variety of situations (especially in chemical engineering and genetics),

this iterative procedure is complex and so little is known about its theoretical properties. In this talk, I will present a new direction (based on the connections between PLS and orthogonal polynomials) to analyze some statistical aspects of this method. First, I will present the PLS method as it was originally introduced. Then, I will explain the link between PLS and some specific discrete orthogonal polynomials, called the residual polynomials. Thanks to the theory of orthogonal polynomials, we will derive a new and explicit analytical expression for the residual polynomials which clearly shows how the PLS estimator depends on the signal and noise. Based on this approach, new results will be stated for the empirical risk and the mean square error. The shrinkage properties of the PLS estimator will also be investigated. At last, we will conclude this talk by showing how this new approach through polynomials provides a unified framework to easily recover most of the already known PLS properties.

Keywords. Partial Least Squares regression, constrained least squares, orthogonal polynomials, empirical risk, mean squares prediction error, shrinkage properties.

1 Introduction

Le but de cet exposé est de vous présenter une nouvelle approche de la PLS. Cette approche est motivée par le fait que bien que cette méthode ait fait ses preuves dans la pratique elle n'en reste pas moins encore assez mystérieuse. Si ses propriétés statistiques ne sont pas encore totalement comprises, c'est en grande partie dû au fait que l'estimateur PLS dépend de façon non linéaire de la réponse. Notre travail a consisté à trouver une écriture explicite (en termes du bruit et de la décomposition en élements simples de la matrice de design) de la fonction de dépendance qui lie l'estimateur PLS à la réponse [1]. Nous avons ensuite mis ce travail à profit afin d'apporter des éléments nouveaux dans l'étude des propriétés statistiques de cet estimateur [2].

2 Cadre de travail

2.1 Le modèle de régression

On considère le modèle de régression suivant

$$Y = X\beta^* + \varepsilon \tag{1}$$

où $Y \in \mathbb{R}^n$ désigne la réponse, $X \in \mathbb{M}_{n\times p}$ la matrice d'expériences, $\beta^* \in \mathbb{R}^p$ le paramètre inconnu et $\varepsilon \in \mathbb{R}^n$ désigne le bruit. On autorisera p à être plus grand que n et on notera r le rang de X^TX.

2.2 Un outil important: la décomposition en valeurs singulières

La décomposition en valeurs singulières de X est donnée par $X = UDV^T$ où $U \in \mathbb{M}_{n,n}$ et ses colonnes $u_1, ..., u_p$ forment une b.o.n de $\mathbb{R}^n$, $V \in \mathbb{M}_{p,p}$ et ses colonnes $v_1, ..., v_p$ forment une b.o.n de $\mathbb{R}^p$ et $D \in \mathbb{M}_{n,p}$ est la matrice qui contient $(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r})$ sur la diagonale et zéro ailleurs. Sans perte de généralité on supposera que $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_r > 0$. On notera $\tilde{\varepsilon}_i := \varepsilon^T u_i,\ i = 1, ..., n$ et $\tilde{\beta}_i^* := \beta^{*T} v_i,\ i = 1, ..., p$ et on définira deux quantités importantes pour la suite qui sont $p_i := (X\beta^*)^T u_i$ et $\hat{\mathbf{p}}_\mathbf{i} := \mathbf{Y^T u_i},\ i = 1, ..., n$.

2.3 La méthode PLS

La méthode PLS (cf.[3]) à l'étape k (où $k \leqslant r$) consiste à trouver $(w_k)_{1 \leq k \leq K}$ et $(t_k)_{1 \leq k \leq K}$ qui maximise $[\mathrm{Cov}(Y, Xw_k)]^2$ sous les contraintes que

$$\|w_k\|^2 = 1 \ \text{ et } \ t_k = Xw_k \text{ orthogonal à } t_1, ..., t_{k-1}.$$

L'estimateur PLS à l'étape k noté $\hat{\beta}_k$ est alors ensuite obtenu en effectuant la régression linéaire de Y sur $t_1, ...t_k$. Nous ne considèrerons pas ici la construction séquentielle de la PLS mais nous utiliserons la caractérisation équivalente suivante

Proposition 2.1. [4]. *Pour $1 \leq k \leq r$, on a*

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

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

Nous renvoyons à [5], [6], [7] et à [8] pour un aperçu des résultats importants qui existent sur la PLS.

3 Lien entre PLS et polynômes orthogonaux

Pour tout $k \in \mathbb{N}$, $\mathcal{P}_k$ désigne l'ensemble des polynômes de degrés plus petit que k et $\mathcal{P}_{k,1}$ le sous-ensemble de $\mathcal{P}_k$ des polynômes de terme constant égal à 1.

3.1 PLS: un problème de minimisation sur des polynômes

Soit $k \leq r$. La Proposition 3.1 ci-dessous montre que $\hat{\beta}_k$ est de la forme $\hat{P}_k(X^TX)X^TY$, où $\hat{P}_k \in \mathcal{P}_{k-1}$ représente une régularisation de l'inverse de X^TX.

Proposition 3.1. *Soit $k \leq r$. On a*
$\hat{\beta}_k = \hat{P}_k(X^TX)X^TY$ *où* $\hat{P}_k \in \mathcal{P}_{k-1}$ *vérifie* $\hat{P}_k \in \underset{P \in \mathcal{P}_{k-1}}{argmin} \|Y - XP(X^TX)X^TY\|^2$
et
$\|Y - X\hat{\beta}_k\|^2 = \|\hat{Q}_k(XX^T)Y\|^2$ *où* $\hat{Q}_k(t) = 1 - t\hat{P}_k(t) \in \mathcal{P}_{k,1}$ *vérifie* $\hat{Q}_k \in \underset{Q \in \mathcal{P}_{k,1}}{argmin} \|Q(XX^T)Y\|^2$.

Les polynômes $\hat{Q}_k$ sont appelés les polynômes résiduel.

3.2 Les polynômes résiduels

Nous pouvons ensuite montrer que $\left(\hat{Q}_k\right)_{0\leqslant k\leq r}$ définit une suite de polynôme orthogonaux par rapport à une mesure discrète particulière.

Proposition 3.2. *$\hat{Q}_0 := 1, \hat{Q}_1, ..., \hat{Q}_r$ sont des polynômes orthogonaux par rapport à*

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

4 Résultat principal: une expression analytique explicite pour les polynômes résiduels

Nous sommes maintenant en mesure d'établir une expression exacte et explicite pour les polynômes résiduels. Cette expression montre clairement comment le bruit sur les observations et la distribution des valeurs propres de la matrice d'expériences ont un impact sur les polynômes résiduels.

Théorème 4.1. *Soit $k \leq r$ et $I_k^+ = \{(j_1, ..., j_k) : r \geq j_1 > ... > j_k \geq 1\}$. On a*

$$\hat{Q}_k(x) = \sum_{(j_1,..,j_k)\in I_k^+} \left[\hat{w}_{(j_1,..,j_k)} \prod_{l=1}^{k} (1 - \frac{x}{\lambda_{j_l}}) \right] \tag{2}$$

où

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

$V(\lambda_{j_1}, ..., \lambda_{j_k})$ désigne le déterminant de Vandermonde associé à $\lambda_{j_1}, ..., \lambda_{j_k}$.

Lors de l'exposé, nous expliquerons et donnerons une interprétation de cette formule.

5 Application à l'étude des propriétés statistiques

Nous nous intéressons maintenant aux propriétés statistiques de l'estimateur PLS et nous allons voir comment l'expression établie précedemment est bien adaptée à cet effet.

5.1 Propriétés d'approximation

Nous présentons ci-dessus une nouvelle expression pour le risque empirique de la PLS.

Théorème 5.1. *Pour $k < r$*

$$\| Y - X\hat{\beta}_k \|^2 =$$

$$\sum_{r>j_1>\ldots>j_k\geq 1} \left[\hat{w}_{j_1,\ldots,j_k} \sum_{i=j_1+1}^{r} \left(\prod_{l=1}^{k} \left(1 - \frac{\lambda_i}{\lambda_{j_l}} \right)^2 \hat{p}_i^2 \right) \right] + \sum_{i=r+1}^{n} \hat{p}_i^2. \tag{3}$$

où par convention $\sum_{i=r+1}^{n} \hat{p}_i^2 = 0$ si $r \geq n$

Corollaire 5.2. *Soit $k < r$.*

$$\| Y - X\hat{\beta}_k \|^2 \leq \left(1 - \frac{\lambda_n}{\lambda_1} \right)^{2k} \sum_{i=k+1}^{r} \hat{p}_i^2 + \sum_{i=r+1}^{n} \hat{p}_i^2.$$

A noter que si $\frac{\lambda_r}{\lambda_k} > 1 - \delta$ alors $\sum_{i=k+1}^{r} \left[\prod_{l=1}^{k} \left(1 - \frac{\lambda_i}{\lambda_l} \right)^2 \hat{p}_i^2 \right] \leq \delta \sum_{i=k+1}^{r} \hat{p}_i^2$.

Par ailleurs, le Corollaire 5.2 permet de montrer que la PLS diminue la partie résiduelle beaucoup plus vite que la régression sur composantes principales dans le sens où $\| Y - X\hat{\beta}_k \|^2 < \sum_{i=k+1}^{n} \hat{p}_i^2 := \| Y - X\hat{\beta}_{PCR}^k \|^2$.

5.2 Propriétés de prédiction

Nous allons maintenant nous intéresser aux propriétés prédictives de la PLS.

5.2.1 Une nouvelle décomposition de la MSPE

Dans cette section nous nous intéressons à l'erreur quadratique moyenne de prédiction (MSPE) définie par

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

La proposition suivante fournit une décomposition intéressante de $\| X\beta^* - X\hat{\beta}_k \|^2$.

Proposition 5.3.

$$\| X\beta^* - X\hat{\beta}_k \|^2 = \sum_{i=1}^{r} \hat{Q}_k(\lambda_i) p_i^2 + \sum_{i=1}^{r} \left(1 - \hat{Q}_k(\lambda_i) \right) \tilde{\varepsilon}_i^2. \tag{4}$$

A noter que $\hat{\beta}_k = \sum_{i=1}^{r} \left(1 - \hat{Q}_k(\lambda_i) \right) \frac{\hat{p}_i}{\sqrt{\lambda_i}} v_i$ et donc l'estimateur PLS peut être vu comme un estimateur de seuillage. Cependant les facteurs de filtrage de la PLS sont aléatoires et pas toujours dans $[0, 1]$. En regardant plus en détails la Proposition 5.3, nous verrons pourquoi une expansion dans certaines directions ne conduit pas forcément dans le cas de la PLS à une augmentation de l'erreur MSPE.

5.2.2 Une borne supérieure pour l'erreur MSPE sous une faible variance du bruit

Le but de cette partie est d'obtenir un contrôle de $\frac{1}{n} \parallel X\beta^* - X\hat{\beta}_k \parallel^2$. Nous considérons ici que les $(\varepsilon_i)_{1\leq i\leq n}$ sont i.i.d de loi gaussienne centrée et de variance σ_n^2 et nous supposerons

- (H.1): $\sigma_n^2 = \mathcal{O}(\frac{1}{n})$ et (H.2): $\min_{1\leq i\leq n}\{p_i^2\} \geq L_n := \frac{\log n}{n}$.

On a alors le théorème suivant

Théorème 5.4. *Soit $k \leq r$. Supposons (H.1) et (H.2).*
Avec une probabilité plus grande que $1 - n^{1-C}$ où $C > 1$, on a

$$\frac{1}{n} \parallel X\beta^* - X\hat{\beta}_k \parallel^2 \leq$$

$$\frac{1}{n}\left(1 - \frac{\lambda_n}{\lambda_1}\right)^{2k} \sum_{i=k+1}^{r} p_i^2 + \frac{\log(n)}{n^2} \sum_{i=1}^{n} |1 - Q_k^*(\lambda_i)| + A.\frac{k}{n}\sqrt{\frac{\log n}{nL_n}} \sum_{i=1}^{n} \left[\underset{I_k^+}{max} \left(\prod_{l=1}^{k} \left| \frac{\lambda_i}{\lambda_{j_l}} - 1 \right| \right)^2 p_i^2 \right],$$

avec $A > 0$ qui est une constante et Q_k^* la version de $\hat{Q}_k$ non bruitée.

Lors de l'exposé, nous détaillerons dans les grandes lignes les idées qui nous ont permis d'aboutir à ce résultat.

Bibliographie

[1] Blazère M., Gamboa F. et Loubes J.M. (2014), PLS: a new statistical insight through the prism of orthogonal polynomials, *arXiv preprint arXiv:1405.5900*.
[2] Blazère M., Gamboa F. et Loubes J.M. (2014), A unified framework for the study of the PLS estimator's properties, *arXiv preprint arXiv:1411.0229*.
[3] Wold, S. et al. (1984), The collinearity problem in linear regression. The partial least squares (pls) approach to generalized inverses, *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.
[4] Helland (1988), On the structure of partial least squares regression , *Communications in statistics-Simulation and Computation* , 58,97–107.
[5] Butler N.A et 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] Lingjaerde O.C et Christophersen N. (2000), Shrinkage structure of partial least squares, *Scandinavian Journal of Statistics*, 27,249–273.
[7] Phatak A. et de Hoog F. (2002), Exploiting the connections between pls, lanczos methods and conjugate gradients, *Journal of Chemometrics*, 16(7), 361–367.
[8] Rosipal R. et Kramer N. (2006), Overview and recent advances in partial least squares, *Subspace, Latent Structure and Feature selection*, 34–51, Springer.