A non asymptotic analysis of the first component PLS regression

Luca Castelli [1] & Irène Gannaz [2]& Clément Marteau[1]

[1] *Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France, castelli@math.univ-lyon1.fr marteau@math.univ-lyon1.fr*

[2] *Univ. Grenoble Alpes, CNRS, Grenoble INP, G-SCOP, 38000 Grenoble, France, irene.gannaz@grenoble-inp.fr*

Résumé. La régression PLS (*Partial Least Squares*) est une méthode statistique permettant de travailler dans un cadre de grande dimension. Cette méthode projette les covariables sur un sous-espace bien choisi, considérant les corrélations successives avec la variable à expliquer dans le but d'améliorer la qualité de prédiction. Nous nous focaliserons sur le cas de la projection sur une composante, qui fournit un cadre utile pour comprendre le mécanisme sous-jacent. Malgré sa simplicité apparente, ce cadre présente de nombreux défis statistiques. En particulier, la non-linéarité de l'estimateur correspondant exige une attention particulière.

Nous fournissons une borne non asymptotique sur la perte quadratique en prédiction avec grande probabilité. Nous montrons que la qualité de l'estimation PLS dépend de l'inverse de l'inertie relative de la composante par rapport à la variance des covariables. Nous étendons ces résultats à l'approche Sparse PLS. En particulier, nous présentons des bornes supérieures similaires à celles obtenues avec l'algorithme LASSO, avec une contrainte supplémentaire sur les valeurs propres restreintes de la matrice de design.

Mots-clés. Réduction de dimension, Régression, Parcimonie

Abstract. Partial Least Squares (PLS) regression is a dimension reduction technique used to handle high dimensionality. This method projects the data onto a carefully chosen subspace, considering successive correlations with the explanatory variable in order to improve the prediction quality. We focus our attention on the single component case, that provides a useful framework to understand the underlying mechanism. Despite its apparent simplicity, this scenario presents numerous statistical challenges. Specifically, the non-linearity of the corresponding estimator demands careful attention. We provide a non-asymptotic upper bound on the quadratic loss in prediction with high probability in a high dimensional regression context. The bound is attained thanks to a preliminary test on the first PLS component. In a second time, we extend these results to the sparse partial least squares approach. In particular, we exhibit upper bounds similar to those obtained with the lasso algorithm, up to an additional restricted eigenvalue constraint on the design matrix.

Keywords. Dimension reduction, regression, sparsity

1 Introduction

Nous nous intéressons au modèle linéaire classique dans un contexte de grande dimension. Nous observons un échantillon de taille n, (X_i, Y_i), $i = 1, \ldots, n$, où les $Y_i \in \mathbb{R}$ sont les variables de sortie et les $X_i \in \mathbb{R}^p$ sont les covariables p-dimensionnelles. Nous considérons une relation linéaire au sein de chaque couple (X_i, Y_i), représentée par l'équation :

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

où $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim \mathcal{N}_n\big(0, \tau^2 I_n\big)$, $X = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n \times p}$ et $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$. La matrice I_n est la matrice identité de taille n et le paramètre τ caractérise le niveau de bruit. Nous désignons par $\Sigma = X^T X/n$ la matrice de Gram associée à la matrice de design X, $\hat{\sigma} = X^T Y/n$ la covariance empirique entre X et la cible Y, et $\sigma = \mathbb{E}[\hat{\sigma}] = \Sigma\beta$ son espérance.

Notre objectif est d'étudier les performances en prédiction dde la régression par Moindres Carrés Partiels (Partial Least Squares, PLS). L'idée de la régression PLS est de rechercher un nombre fixe de directions - disons $K \in \{1, \ldots, p\}$ - formées par des combinaisons linéaires des coordonnées de X, qui sont fortement corrélées avec la variable cible Y (voir Mateos-Aparicio (2011) pour une introduction complète). Ces K directions sont regroupées dans une matrice de poids $W \in \mathbb{R}^{p \times K}$. Le paramètre β est ensuite estimé par une combinaison linéaire appropriée des colonnes de W. Plus formellement, l'estimateur PLS satisfait :

$$\hat{\beta}_W = \underset{w \in [W]}{\operatorname{argmin}} \|Y - Xw\|^2, \tag{2}$$

où $[W] \subset \mathbb{R}^p$ désigne l'espace engendré par les colonnes de la matrice de poids W, et $\|.\|$ est la norme ℓ^2 sur $\mathbb{R}^n$. L'objectif de la réduction de dimension donnée par W est de réduire le nombre de caractéristiques de p à K tout en conservant autant d'informations que possible. Contrairement à la réduction sur composantes principales où les directions sont construites uniquement en considérant les covariables X, la régression PLS construit les poids W de manière itérative, en tenant compte des corrélations successives avec la réponse Y, afin d'améliorer la qualité de prédiction.

Nous allons concentrer notre attention dans la suite sur le cas de la projection sur une seule composante, c'est-à-dire $K = 1$.

2 La régression PLS

La régression PLS a été principalement développée dans la communauté de la chimiométrie (Martens et Naes, 1992). Cette approche a démontré sa capacité à prédire des modèles de régression avec de nombreuses variables prédictives (Garthwaite, 1994). Elle a été largement utilisée en chimiométrie (Wold, 1995; Wold, Sjöström et Eriksson, 2001), mais aussi dans d'autres domaines tels que les sciences sociales (Sawatsky, Clyde et Meek, 2015) et la biologie (Palermo, Piraino et Zucht, 2009; Yang et al., 2017). Plusieurs extensions ont été proposées au fil des ans, comme par exemple Delaigle et Hall (2012) pour les données fonctionnelles ou Naik et Tsai (2000) pour les modèles à indice unique.

La PLS possède une structure algorithmique due à la construction du sous-espace $[W]$ constitué des poids permettant de sélectionner les variables pour favoriser la prédiction.

Algorithm 1 PLS Algorithm

Input $\mathbf{X}$,Y and K

$\mathbf{X}_1$=$\mathbf{X}$
for k=1,..., K **do**
  $\mathbf{w}_k$=$\mathbf{X}^{(k)^T}Y/\|\mathbf{X}^{(k)^T}Y\|_2$ (loadings computation)
  $\mathbf{t}_k$=$\mathbf{X}^{(k)}\mathbf{w}_k$ (component construction)
  $\mathbf{X}^{(k+1)}$=$\mathbf{X}^{(k)}$ - $\mathbf{P}_{[\mathbf{t}_k]}(\mathbf{X}^{(k)})$ (deflation step)
end for

Les composantes PLS $(t_1, \ldots, t_K)$ construites par l'algorithme sont rassemblées dans la matrice $T \in \mathbb{R}^{n\times K}$, où chaque colonne k de T correspond à t_k. En particulier on peut noter que $[T] = [XW]$ où $W = (w_1, \ldots, w_K) \in \mathbb{R}^{p\times K}$ est la matrice contenant les poids w_k. La prédiction PLS associée $\hat{Y}_{PLS} := X\hat{\beta}_{PLS}$ est donnée par

$$\hat{Y}_{PLS} = T(T^TT)^{-1}T^TY = P_{[XW]}(Y),$$

où l'exposant T désigne la transoposition. L'estimateur de β se calcule ensuite de la manière suivante :

$$\hat{\beta}_{PLS} = \hat{\beta}_W = W(W^T\Sigma W)^{-1}W^T\hat{\sigma}, \tag{3}$$

où $\hat{\sigma} = X^TY/n$ correspond à la covariance empirique entre X et Y.

Bien que le principe des Moindres Carrés Partiels (PLS) ait attiré beaucoup d'attention au fil des ans, peu de résultats théoriques ont été obtenus. Entre autres, on peut citer Helland (1990) qui a caractérisé l'espace [W] résultant de l'approche PLS à l'aide de Σ et σ. Cook, Li et Chiaromonte (2010) et Cook, Helland et Su (2013) ont établi un lien entre la régression PLS et les enveloppes. Alors que Chun et Keleş (2010) ont prouvé l'inconsistance de l'estimateur PLS lorsque le nombre de covariables est trop important, Cook et Forzani (2017) ainsi que Cook et Forzani (2019) ont établi - sous des contraintes fortes sur β et la matrice de design X - le comportement asymptotique de l'erreur quadratique moyenne de prédiction et ont démontré qu'elle peut tendre vers 0 lorsque le nombre d'observations tend vers l'infini. Par ailleurs, des travaux ont également proposé de prendre en compte des contraintes de parcimonie dans l'algorithme PLS. Nous renvoyons par exemple à Durif et al. (2017) et Alsouki et al. (2023).

3 Contribution

Nous résumons ici les résultats obtenus dans Castelli, Gannaz et Marteau (2023). Nous nous concentrons sur le cas de la régression PLS sur une composante, c'est à dire $K = 1$. Nous avons dans ce cadre une formule explicite de l'estimateur PLS :

$$\hat{\beta}_{PLS} = \frac{\hat{\sigma}^T\hat{\sigma}}{\hat{\sigma}^T\Sigma\hat{\sigma}}\hat{\sigma}. \tag{4}$$

Nous établissons une borne non asymptotique sur la perte de prédiction. En désignant par $\hat{\beta}_{PLS}$ l'estimateur PLS avec $K = 1$, nous obtenons le résultat suivant :

Théorème. *Soit $\delta \in (0,1)$. Supposons que*

$$\hat{\sigma}^T \Sigma \hat{\sigma} > \mathrm{t}_\delta \, p_n \text{ avec } p_n = \frac{\tau^2}{n} \rho(\Sigma) \, Tr(\Sigma) \tag{5}$$

où t_δ est une constante dépendant uniquement de δ. Alors, avec une probabilité supérieure à $1-\delta$, il existe une constante $C_\delta > 0$, dépendant uniquement de δ, telle que

$$\frac{1}{n}\|X\hat{\beta}_{PLS} - X\beta\|^2 \leq B(\beta) + C_\delta \frac{\tau^2}{n} \max\left(\frac{\mathrm{Tr}(\Sigma)}{\lambda}, \frac{\rho(\Sigma)\mathrm{Tr}(\Sigma)}{\lambda^2}\right), \tag{6}$$

avec $B(\beta) = \frac{2}{n} \inf_{v \in [\sigma]} \|X(\beta - v)\|^2$ et

$$\lambda = \frac{\sigma^T \Sigma \sigma}{\sigma^T \sigma}. \tag{7}$$

$\mathrm{Tr}(.)$ *est l'opérateur de trace sur $\mathbb{R}^{p \times p}$.*

Nous renvoyons à Castelli, Gannaz et Marteau (2023) pour la preuve de ce résultat. Les quantités t_δ et C_δ ne sont pas données de façon explicite ici, mais elles peuvent être trouvées dans la référence sus-citée.

Dans Castelli, Gannaz et Marteau (2023), nous étendons ce résultat au cas parcimonieux en utilisant une version sparse $\hat{\beta}_{sPLS}$ de l'algorithme, comprenant une contrainte ℓ_1 dans le processus d'optimisation des poids w_k. En supposant que la matrice de Gram Σ satisfait une condition de valeurs propres restreintes, similaire à Bickel, Ritov et Tsybakov (2009), nous établissons que, avec grande probabilité,

$$\frac{1}{n}\|X\tilde{\beta}_{sPLS} - X\beta\|^2 \leq B(\beta) + C\frac{\tau^2 s}{n}\ln(p),$$

où s représente le nombre de coefficients non nuls du premier axe PLS et C est une constante. En particulier, nous retrouvons, à l'exception du terme de biais, le même type de borne que celles obtenues pour la procédure Lasso (nous renvoyons à Tibshirani (1996) et Bickel, Ritov et Tsybakov (2009)). Ce résultat n'est pas détaillé ici.

La condition (5) peut être interprétée comme une condition de ratio signal sur bruit qui doit être suffisamment élevé pour que l'estimateur $\hat{\beta}_{PLS}$ contienne de l'information. En effet, on peut montrer que cette condition revient à supposer que la norme de la composante t_1 obtenue par l'algorithme PLS est suffisamment grande. Si cette norme est proche de 0, la regression PLS n'est pas pertinente. Si le signal contenu dans la première composante est plus grand qu'un seuil donné, cela assure la qualité de l'estimation.

Ensuite, la quantité $B(\beta)$ est une mesure du biais induit par l'algorithme. Le deuxième terme de la borne (6) peut être considéré comme un terme de variance. Il mesure essentiellement l'impact du bruit ε sur l'algorithme PLS. Il s'agit d'un rapport entre la trace de Σ et le terme λ introduit dans (7). La quantité λ correspond à la norme théorique de la première

composante PLS t_1. En d'autres termes, le terme $\frac{\mathrm{Tr}(\Sigma)}{\lambda}$ peut être considéré comme l'inverse de l'inertie relative. Il fournit une sorte de rapport signal/bruit inverse qui contrôle la précision de la composante PLS unique. Si ce rapport est proche de 1, la première composante PLS capture la majeure partie de l'inertie des données et nous obtenons un terme de variance avec un taux paramétrique τ^2/n. En revanche, si ce rapport est élevé, on ne peut pas s'attendre à obtenir des résultats précis pour la PLS à une seule composante : la quantité de signal capturée dans la première composante n'est pas assez importante.

Cook et Forzani (2016) ont établi un résultat similaire. Leur résultat s'exprime en fonction de $\frac{1}{\sigma^T\sigma}$ qui, sous leurs hypothèses, est équivalent à $1/\lambda$ (voir leur section 3.2 page 12). Notre résultat apporte deux principales contributions par rapport à leur travail :

- Nous ne supposons pas, contrairement à Cook et Forzani (2016), que β est colinéaire à la première composante t_1 avec notamment une hypothèse d'inversibilité sur la matrice Σ. Ceci induit notamment un terme de biais dans l'étude de la qualité de prédiction. Nous nous plaçons donc dans un cadre plus général.

- Notre étude est non asymptotique. En effet Cook et Forzani (2016) ont établi que la prédiction de l'estimateur $\hat{\beta}_{PLS}$ est asymptotiquement de l'ordre de $1/n$, mais nous montrons que ce résultat est valable en grande probabilité.

Remarquons enfin que nous considérons des covariables X déterministes, tandis que Cook et Forzani (2016) considèrent des covariables X gaussiennes. Ceci implique notamment que notre condition (5) s'exprime en fonction de X (via Σ et $\hat{\sigma}$) et que notre borne fasse intervenir le ratio $\frac{\mathrm{Tr}(\Sigma)}{\lambda}$. Nous pensons que le fait de conserver X dans l'étude permet de mieux comprendre les mécanismes de la régression PLS, notamment par la condition de borne inférieure de la norme de la composante et la borne pouvant s'interpréter comme l'inverse de l'inertie relative de l'axe.

Conclusion

Notre travail concerne l'estimateur PLS sur une composante. Nous fournissons des bornes non-asymptotiques avec covariables fixes pour la qualité de prédiction de cet estimateur. Ces bornes sont obtenues sous la condition que la quantité de signal de cette composante est suffisamment importante.

Notre résultat peut être interprété comme une décomposition biais-variance reflétant l'importance de la prise en compte du biais dans nos hypothèses. Nous explicitons la borne de la prédiction, avec une interprétation en inertie relative inverse. Ceci permet de mettre en lumière le rôle des composantes dans la régression PLS : plus la composante PLS explique la variance des covariables X, meilleure sera la prédiction.

La principale perspective de ce travail est d'étendre les résultats obtenus au cadre multidimensionnel, c'est-à-dire au cas général où $1 \leq K \leq p$. Des résultats asymptotiques quant à la qualité de prédiction dans le cas multidimensionnel ont été obtenus par Cook

et Forzani (2019). Notre objectif est de fournir des résultats non asymptotiques, avec des bornes explicites, et des hypothèses peu restrictives.

Bibliographie

Alsouki L., Duval L., Marteau C., El Haddad R. and Wahl F. (2023) Dual-sPLS: a family of Dual Sparse Partial Least Squares regressions for feature selection and prediction with tunable sparsity; evaluation on simulated and near-infrared (NIR) spectra, *Chemometrics and Intelligent Laboratory system*, 237, 104813.

Bickel, Peter J. and Ritov, Ya'acov and Tsybakov, Alexandre B. (2009) Simultaneous analysis of lasso and Dantzig selector *The Annals of Statistics*, 37, 1705–1732.

Castelli L., Gannaz I. and Marteau M. (2023) A non asymptotic analysis of the single component PLS regression, arXiv preprint arXiv:2310.10115.

Chun H. and Keles S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1), 3-25.

Cook R.D., Li B., and Chiaromonte F. (2010) Envelope models for parsimonious and efficient multivariate linear regression, *Statistica Sinica*, 20(3), 927-960.

Cook, R. D., Helland I. and Su Z. (2013) Envelopes and partial least squares regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, 75(5), 851-877.

Cook, R. D. and Forzani L. (2017) Big data and partial least-squares prediction, *Canadian Journal of Statistics*, 46(1), 62-78.

Cook, R.D. and Forzani L. (2019) Partial least squares prediction in high-dimensional regression, *The Annals of Statistics*, 47(2), 884-908..

Delaigle A. and Hall P. (2012) Methodology and theory for partial least squares applied to functional data, *The Annals of Statistics*, 40(1), 322-352.

Durif G., Modolo L., Michaelsson J., Mold J.E., Lambert-Lacroix S. and Picard F. (2017) High dimensional classification with combined adaptive sparse PLS and logistic regression, *Bioinformatics*, 34(3), 485-493.

Garthwaite P.H. (1994) An interpretation of partial least squares, *Journal of the American Statistical Association*, 89(425), 122-127.

Helland I. (1990) Partial least squares regression and statistical models, *Scandinavian journal of statistics*, 17(2), 97-114.

Martens H. and Naes T. (1992) Multivariate calibration, *John Wiley & Sons*.

Mateos-Aparicio G. (2011) Partial least squares (PLS) methods: origins, evolution, and application to social sciences, *Communications in Statistics - Theory and Methods*, 40(13), 2305-2317.

Naik P. and Tsai C. (2000) Partial least squares estimator for single-index models, *Journal*

of the Royal Statistical Society: Series B (Statistical Methodology), 62(4), 763-771..

Palermo G., Piraino P. and Zucht H.D. (2009) Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data, *Advances and Applications in Bioinformatics and Chemistry*, 57-70.

Sawatsky M., Clyde M. and Meek F. (2015) Partial least squares regression in the social sciences, *The Quantitative Methods for Psychology*, 11(2), 52-62.

Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Wold S. (1995) Chemometrics; what do we mean with it, and what do we want from it?, *Chemometrics and Intelligent Laboratory Systems*, 30(1), 109-115.

Wold S., Sjöström M. and Eriksson L. (2001), PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109-130.

Yang T.C., Aucott L.S., Duthie G.G. and Macdonald H.M. (2017) An application of partial least squares for identifying dietary patterns in bone health, *Archives of osteoporosis*, 12, 1-8.