Adaptive Sparse PLS for Logistic Regression

Ghislain DURIF [1] & Franck PICARD [1] & Sophie LAMBERT-LACROIX [2]

[1] *LBBE, UMR CNRS 5558 Univ. Lyon 1, F-69622 Villeurbanne, France*
[2] *UMR 5525 UJF-Grenoble 1/CNRS/UPMF/TIMC-IMAG, Grenoble, F-38041, France*

Résumé. Depuis quelques années, l'analyse de données rencontrent des problématiques liées à la grande dimension. Dans ce contexte, c'est-à-dire quand le nombre de variables considérées est bien supérieur au nombre d'observations dans l'échantillon, les méthodes classiques de classification supervisée sont inappropriées, ce qui appelle au développement de nouvelles méthodologies. Nous présentons une nouvelle méthode appropriée pour la classification supervisée en grande dimension. Elle utilise la régression *sparse Partial Least Squares* ou SPLS, réalisant compression et sélection de variables combinés à une régression logistique pénalisée par Ridge. Par des simulations, nous montrons la précision, la stabilité et la convergence de notre méthode, comparé à d'autres approches dans l'état de l'art. En particulier, il apparaît que la compression améliore l'exactitude de la sélection, et que notre méthode est plus stable concernant le choix des hyper-paramètres par validation croisée, contrairement aux approches réalisant la classification supervisée avec la *sparse* PLS.

Mots-clés. classification supervisée, séléction de variables, compression, réduction de dimension, modèle linéaire généralisé

Abstract. Since few years, data analysis struggles with statistical issues related to the "curse of high dimensionality". In this context, meaning when the number of considered variables is far larger than the number of observations in the sample, standard methods for classification are inappropriate, calling for the development of new methodologies. I will present a new method suitable for classification in the high dimensional case. It uses Sparse Partial Least Squares (Sparse PLS) performing compression and variable selection combined to Ridge penalized logistic regression. In particular, we developed an adaptive version of Sparse PLS to improve the dimension reduction process. I will illustrate the interest of our method by classification results on simulated and real data set, comparing to state-of-the-art approaches. The application focus on genomics where dimensions are huge, and especially on prediction of breast cancer relapse (binary) using gene expression level (quantitative).

Keywords. classification, variable selection, compression, dimension reduction, generalized linear model

Compression et sélection de variables pour la classification supervisée

La grande dimension constitue une problématique majeure pour le développement de nouvelles méthodologies statistiques (Marimont et Shapiro, 1979; Donoho, 2000). Dans le contexte de l'analyse de données génomiques par exemple, le nombre de variables p (comme l'expression des gènes) est beaucoup plus élevé que la taille de l'échantillon n. Dans ce cas, les méthodes classiques pour la régression ou la classification supervisé deviennent inappropriées (Aggarwal et al., 2001; Hastie et al., 2009). En effet, la grande dimension est souvent associée à des phénomènes de dépendances entre variables, amenant à des singularités dans les processus d'optimisation, avec des solutions non uniques ou non stables.

Ce challenge appelle au développement d'outils spécifiques, comme les approches réduisant la dimension, qui peuvent être de deux types différents. D'une part, les techniques de compression consistent à projeter les observations dans un espace de dimension inférieure pour résumer l'information contenue dans les différentes variables. Par exemple, la régression *Partial Least Squares* ou PLS (Wold et al., 1984; Helland, 1988; Tenenhaus, 1998; Wegelin, 2000, Wold et al., 2001) est appropriée à la régression linéaire, en particulier dans le cas de covariables très corrélées. Cette méthode construit des nouvelles composantes comme combinaisons linéaires des prédicteurs, lesquelles maximisent leur covariance avec la variable réponse. D'autre part, les méthodes de sélection de variables sont basées sur une hypothèse de parcimonie, signifiant que seulement quelques variables contribuent au modèle. Leur objectif est de sélectionner ces variables pertinentes et de retirer les autres du modèle. Un exemple d'une telle approche est le Lasso (Tibshirani, 1996), avec sa pénalité sur la norme ℓ_1 des coefficients, forçant les coefficients correspondants aux variables les moins importantes à être nuls. Enfin, la régression *sparse* PLS (Lê Cao et al., 2008; Chun et Keleş, 2010) combinent compression et sélection de variables pour réduire la dimension. Elle consiste à introduire une étape de sélection dans l'algorithme PLS, afin de construire des nouvelles composantes comme combinaisons linéaires parcimonieuses des prédicteurs. Son avantage par rapport au Lasso se révèle dans le cas d'un design très corrélé. Alors que le Lasso ne sélectionne qu'une seule variable parmi un groupe de variables importantes corrélées, la *sparse* PLS sélectionne tous les prédicteurs pertinents dans un groupe corrélé (Chun et Keleş, 2010). Il apparaît même que combiner compression et sélection améliore l'efficacité de la prédiction et la justesse de la sélection, comparé au Lasso ou même à l'Elastic Net (Chun et Keleş, 2010), méthode introduite par Zou et Hastie (2005).

La PLS *sparse* a montré d'excellentes performances dans le cadre de la régression traitant une réponse continue. Cependant, il s'avère que son adaptation à la classification supervisée est difficile. Chung et Keleş (2010) ou Lê Cao et al. (2011) ont proposé

d'utiliser la *sparse* PLS comme une étape de réduction de dimension préliminaire avant d'utiliser une méthode de classification standard comme l'analyse discriminante, comme l'avait proposé précédemment Nguyen et Rocke (2002 a, b) ou Boulesteix (2004) pour la PLS classique. Une autre solution consiste à utiliser la régression logistique, une méthode de classification dérivée des modèles linéaires généralisés ou GLM (Nelder et Wedderburn, 1972; McCullagh et Nelder, 1989), qui peuvent traiter des réponses aux distributions variées (binaire, multicatégoriel, comptage) via l'estimation par maximum de vraisemblance. Cette optimisation est accomplie par l'algorithme *Iteratively Reweighted Least Squares* ou IRLS (Green, 1984). Cependant, sa convergence peut être problématique (Albert et Anderson, 1984), particulièrement dans le cas de la grande dimension.

La principale difficulté quand on combine régression logistique et (S)PLS réside dans le fait que ces méthodes sont itératives, et peuvent s'avérer compliquées à associer, spécifiquement concernant l'algorithme IRLS dont la convergence est un point critique en grande dimension. Réaliser la compression par (S)PLS (Wang et al., 1999; Chung et Keleş) sur la réponse discrète comme étape préalable à la régression logistique reste contre-intuitif, puisque la (S)PLS est conçue pour manipuler une réponse continue dans un modèle homoscédastique. Marx (1996) a proposé d'utiliser la PLS à chaque itération de l'algorithme IRLS, afin de résoudre les moindres carrés pondérés à chaque itération, Chung et Keleş (2010) ont également suivi cette idée avec la *sparse* PLS, néanmoins les problèmes de convergence persistent avec une telle approche (Fort et Lambert-Lacroix, 2005). Notre méthode va donc d'abord reposer sur l'usage d'une régression logistique pénalisée par un contrainte de type ℓ_2 ou *Ridge* (Eilers, 2001) pour s'assurer de la convergence de l'algorithme IRLS. Dans ce contexte, une pseudo-réponse continue est générée, ce qui rend la régression PLS appropriée pour estimer les coefficients des prédicteurs, comme proposé par Fort et Lambert-Lacroix (2005). En particulier, les coefficients dans le modèle logistique sont estimés par sparse PLS sur les prédicteurs et cette pseudo-réponse. Cette étape de sélection de variables et de compression permet d'éviter les écueil liés à la grande dimension, tout en facilitant l'interprétation du modèle. Ainsi, contrairement à certaines des approches précédemment proposées, la *sparse* PLS est ici appliquée sur des variables continues, cadre pour lequel elle est adaptée. De plus, l'intégration de la *sparse* PLS à l'extérieur de la boucle *Ridge* IRLS évite de perturber sa convergence.

Nous avons développé une méthode utilisant la *sparse* PLS afin de combiner compression et sélection de variables dans le contexte des modèles linéaires généralisés. Nous proposons également une version adaptative de la *sparse* PLS, basée sur le Lasso adaptatif (Zou, 2006), pour améliorer la pertinence de la sélection de variables. Elle consiste à ajuster la pénalité sur le coefficient de chaque prédicteurs afin de pénaliser davantage les variables les moins pertinentes pour expliquer la réponse. Par des simulations, nous montrons la précision, la stabilité et la convergence de notre méthode, par rapport à d'autres approches dans l'état de l'art. Nous avons notamment comparé notre méthode à

celles précédement citées combinant (*sparse*) PLS et classification supervisée, ainsi qu'à la mt́hode GLMNET (Friedman et al., 2010) basée sur l'optimisation d'une vraisemblance pénalisée par *Elastic Net*. En particulier, il apparaît que la compression améliore l'exactitude de la sélection, et que notre méthode est plus stable concernant le choix des hyper-paramètres par validation croisée, contrairement aux autres approches utilisant la *sparse* PLS, tout en gardant un niveau de prédiction similaire voire supérieur (erreur de prédiction plus faible). Nous avons également réalisé des comparaisons sur des données réelles. À partir de données d'expression pour des milliers de gènes concernant moins de trois cent patients (Guedj et al., 2012), les différentes méthodes ont été testées pour prédire la rechute pour des cancers du sein, confirmant les résultats obtenues sur les simulations, notamment la performance de notre méthode en terme de compression. Nous proposons enfin une version complétée du package `R plsgenomics` qui sera bientôt disponible sur le `CRAN` (http://cran.r-project.org/).

Bibliographie

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. Springer.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika, 71(1), 1-10.

Boulesteix, A.-L. (2004). PLS dimension reduction for classification with microarray data. Statistical Applications in Genetics and Molecular Biology, 3(1).

Chun, H., & Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society. Series B (Methodological), 72(1), 3-25. doi:10.1111/j.1467-9868.2009.00723.x

Chung, D., & Keleş, S. (2010). Sparse partial least squares classification for high dimensional data. Statistical Applications in Genetics and Molecular Biology, 9, Article17. doi:10.2202/1544-6115.1492

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture, 1-33.

Eilers, P. H. C. (2001). Classification of microarray data with penalized logistic regression. BiOS 2001 The International Symposium on Biomedical Optics (2001), 187-198. doi:10.1117/12.427987

Fort, G., & Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. Bioinformatics (Oxford, England), 21(7), 1104-11. doi:10.1093/bioinformatics/bti114

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized

linear models via coordinate descent. Journal of Statistical Software, 33(1), 1-20.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. Journal of the Royal Statistical Society. Series B (Methodological), 46(2), 149-192.

Guedj, M., Marisa, L., de Reynies, A, Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., Bertheau, P., Bertrand, F., Bonnet, F., Martin, A. L., Feugeas, J. P., Bièche, I., Lehmann-Che, J., Lidereau, R., Birnbaum, D., Bertucci, F., de Thé, H., Theillet, C. (2012). A refined molecular taxonomy of breast cancer. Oncogene, 31(9), 1196-206. doi:10.1038/onc.2011.301

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second., p. 767). Springer.

Helland, I. S. (1988). On the structure of partial least squares regression. Communications in Statistics - Simulation and Computation, 17(2), 581-607. doi:10.1080/03610918808812681

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., & Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. Statistical Applications in Genetics and Molecular Biology, 7(1), Article 35. doi:10.2202/1544-6115.1390

Lê Cao, K.-A., Boitard, S., & Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics, 12, 253. doi:10.1186/1471-2105-12-253

Marx, B. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. Technometrics, 38(4), 374-381.

Marimont, R. B., & Shapiro, M. B. (1979). Nearest Neighbour Searches and the Curse of Dimensionality. IMA Journal of Applied Mathematics, 24(1), 59-70. doi:10.1093/imamat/24.1.59

McCullagh, P., & Nelder, J. (1989). Generalized linear models. Chapman & Hall.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370-384.

Nguyen, D. V, & Rocke, D. M. (2002). Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics (Oxford, England), 18(9), 1216-26.

Nguyen, D. V, & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics (Oxford, England), 18(1), 39-50.

Tenenhaus, M. (1998). La régression PLS: Théorie et pratique. Technip.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.

Wang, C. Y., Chen, C. T., Chiang, C. P., Young, S. T., Chow, S. N., & Chiang, H. K. (1999). A probability-based multivariate statistical algorithm for autofluorescence spectroscopic identification of oral carcinogenesis. Photochemistry and Photobiology, 69(4), 471-7.

Wegelin, J. A. (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Tech Rep 371, Department of Statistics, University of Washington, Seattle.

Wold S., Ruhe A., Wold H. & Dunn III W.J., (1984) The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses, SIAM J. Sci. Stat. Comput., 5, nÂ°3, 735-743

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