SÉLECTION DE VARIABLES POUR L'IMAGERIE HYPERSPECTRALE

Anthony Zullo [1,2] & Mathieu Fauvel [1] & Frédéric Ferraty [2]

[1] *Laboratoire DYNAFOR - UMR 1201 - INRA & INP Toulouse, Avenue de l'Agrobiopole, 31326 Castanet-Tolosan, France*
[2] *Institut de Mathématiques de Toulouse - UMR 5219 & Université de Toulouse, 118 route de Narbonne, 31062 Toulouse, France*
Adresses mail : anthony.zullo@toulouse.inra.fr ; mathieu.fauvel@ensat.fr ; ferraty@math.univ-toulouse.fr

Résumé. L'imagerie hyperspectrale est un domaine qui s'est développé récemment et qui nécessite le développement de nouvelles méthodes statistiques spécifiquement adaptées. Le principal problème engendré par ces images provient de la finesse de leur résolution spectrale, générant ainsi des données de grandes dimensions, et entraînant en conséquence l'apparition du problème statistique appelé "fléau de la dimension" se référant à la situation où le rapport nombre de variables sur taille d'échantillon est très grand. L'objectif de cette présentation est d'évaluer la pertinence de deux méthodes parcimonieuses, l'une linéaire et l'autre non linéaire, permettant de prédire une réponse scalaire à partir d'un petit nombre de variables explicatives. Nous nous focalisons sur la mise en œuvre de deux techniques statistiques dites "sélectives" dont l'objectif principal est de retenir un nombre raisonnable de variables explicatives tout en conservant un bon pouvoir prédictif. L'avantage de ce type de méthodes sélectives est qu'il propose des modèles plus interprétables. La première méthode sélective implémentée, appelée Lasso, permet de retenir les variables les plus explicatives dans le cadre d'un modèle de régression linéaire. La seconde est une méthode sélective non-paramétrique développée récemment qui combine un algorithme "pas à pas" de type forward avec un outil d'estimation non-paramétrique appelé régression linéaire locale. L'aspect non-paramétrique de cette méthode autorise la prise en compte de relations non linéaires. Les comportements de ces deux méthodes sont comparés sur un jeu de données hyperspectral selon un critère de validation croisée.

Mots-clés. Fléau de la dimension, imagerie hyperspectrale, Lasso, régression linéaire locale, sélection non-paramétrique de variables, validation croisée.

Abstract. Hyperspectral imaging is an area that has been developed recently and requires the development of new statistical methods specifically adapted. The main problem caused by these images comes from their fine spectral resolution, thereby generating high-dimensional data, and therefore causing the appearance of the statistical problem called "curse of dimensionality" referring to the situation where the relative number of variables on sample size is very large. The objective of this presentation is to assess the relevance of

two sparse methods, one linear and one nonlinear, for the prediction of a scalar response from a small number of explanatory variables. We focus on the implementation of two statistical "selective" techniques whose main objective is to keep a reasonable number of variables while maintaining a good predictive power. The advantage of selective methods is that it provides more interpretable models. The first implemented selective method, called Lasso, keeps the most explanatory variables in the context of a linear regression model. The second is a more recently developed nonparametric method that combines a selective step-by-step forward type algorithm with a non-parametric estimation tool called local linear regression. The nonparametric aspect of this method allows the inclusion of nonlinear relations. Behavior in practice of these two methods are compared on a set of hyperspectral data using a cross-validation criterion.

Keywords. Curse of dimensionality, hyperspectral imaging, Lasso, local linear regression, non-parametric variable selection, cross-validation.

1 Introduction

L'étude d'images hyperspectrales a fait l'objet d'une attention particulière au cours des dix dernières années. Nous nous intéressons à des images hyperspectrales pour lesquelles, à chaque pixel i est associé, d'une part, un hyperspectre, courbe finement échantillonnée selon d longueurs d'onde $\lambda^1, \ldots, \lambda^d$, représenté par un vecteur aléatoire $X_i = (X_i^1, \ldots, X_i^d)$ de dimension d (i.e., $X_i^j = X_i(\lambda^j),\ \forall j \in \{1, \ldots, d\}$), et d'autre part, une variable réponse quantitative Y_i. Étant donné $\{(X_i, Y_i),\ \forall i \in \{1, \ldots, n\}\}$ un échantillon d'apprentissage, le problème de "régression" consiste à associer à chaque pixel une estimation de la valeur de la variable réponse correspondante. La régression supervisée dans le cadre de l'étude d'images hyperspectrales est cependant une tâche difficile : la majorité des méthodes de régression n'est pas appropriée au traitement de telles données comme nous l'explique Jimenez et Landgrebe (1998). Le manque d'efficience de ces méthodes est principalement dû au concours de divers paramètres. En effet, les hyperspectres échantillonnés ont un nombre important d de bandes spectrales ; de plus, on se place dans le cas d'une petite taille n d'échantillon d'apprentissage (i.e., un petit nombre de pixels). D'un point de vue statistique, cela revient donc à considérer un jeu de données contenant un grand nombre de variables d pour un échantillon de petite taille n. Cette situation inconfortable est plus connue sous le nom de "fléau de la dimension", phénomène notamment expliqué par Donoho (2010).

Parmi toutes les méthodes statistiques existantes, la sélection de variables regroupe un ensemble de méthodes permettant de résoudre des problèmes en grande dimension, notamment lorsque le nombre de variables est largement supérieur au nombre d'individus. Dans une telle configuration, on choisit en général d'émettre l'hypothèse que seules quelques variables (en nombre inférieur au nombre d'individus) suffisent à la construction d'un modèle permettant une résolution convenable du problème posé. On obtient ainsi un modèle

plus interprétable qui peut même dans certains cas être "meilleur" (au sens d'un critère défini) que le modèle complet. Nous avons choisi de présenter une méthode non-linéaire de sélection de variables appelée *Most-Predictive Design Points* (MPDP), méthode récemment développée par Ferraty et al. (2010), et de comparer ses performances sur un jeu de données avec une autre méthode de sélection plus standard car linéaire appelée Lasso (Tibshirani, 1996).

Dans la suite de cet article, nous présenterons ces deux méthodes avant de les appliquer pour les comparer sur un jeu de données particulier, puis nous conclurons quant aux résultats obtenus. Une troisième méthode non sélective appelée régression Ridge (Hoerl et Kennard, 1970), sera aussi implémentée afin de souligner l'intérêt de construire des modèles parcimonieux.

2 Méthodologie statistique

Présentons brièvement les deux méthodes de sélection de variables qui seront appliquées sur notre jeu de données hyperspectral : la méthode Lasso, qui modélise de façon parcimonieuse et linéaire la relation entre la variable réponse et les variables explicatives, ainsi que la méthode MPDP, récemment développée, qui sélectionne les variables de façon non-paramétrique.

2.1 La méthode Lasso

Cette méthode est fondé sur le principe de minimisation d'un problème de moindres carrés pénalisés : $\hat{\beta}^L := \arg\min_{\beta} \{\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{d} X_i^j \beta_j)^2 + \lambda^L \sum_{j=1}^{d} |\beta_j|\}$, où λ^L est un paramètre de l'estimateur à régler. L'introduction d'une pénalité fondée sur la norme L^1 du vecteur des paramètres permet d'annuler un grand nombre de paramètres, ce qui revient à sélectionner un petit nombre de variables. Concernant la procédure d'estimation des paramètres, elle nécessite l'utilisation d'une variante de l'algorithme appelé *Least Angle Regression* (LAR), algorithme développé par Efron et al. (2004) cherchant les variables explicatives X_i les plus linéairement corrélées avec les résidus successifs de la variable réponse Y. En pratique, le réglage du paramètre λ^L est remplacé par celui d'un paramètre de saturation du modèle s compris entre 0 (correspondant à la nullité de tous les coefficients estimés) et 1 (correspondant à l'estimation des moindres carrés).

2.2 La méthode *Most-Predictive Design Points* (MPDP)

Il s'agit d'une méthode statistique s'appuyant sur une méthode de régression non-paramétrique particulière : la régression linéaire locale (Fan et Gijbels, 1996). L'algorithme permettant de réaliser cette sélection de variables est basé sur une procédure de type *Forward* pour sélectionner les variables les plus significatives. Cette méthode cherche le sous-ensemble de variables $\{X^{j_1}, \ldots, X^{j_p}\} \subset \{X^1, \ldots, X^d\}$ minimisant le critère de validation croisée de type LOOCV défini par $cv(X^{j_1}, \ldots, X^{j_p}) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}_{-i}(X_i^{j_1}, \ldots, X_i^{j_p}))^2$, où $\hat{g}_{-i}$ est l'estimateur de la régression linéaire locale calculé sans le $i^{ième}$ individu. Cet algorithme procède pas à pas : on cherche la variable la plus prédictive, puis parmi les variables restantes, on en déduit le couple le plus prédictif en conservant la première variable sélectionnée, et ainsi de suite jusqu'à atteindre un critère d'arrêt convenablement choisi. En pratique, la régression linéaire locale nécessite le choix d'un paramètre de lissage h obtenu par une validation croisée LOOCV.

3 Application aux données et comparaison des résultats

Le jeu de données étudié provient de relevés réalisés sur 25 prairies à l'aide d'un spectromètre pour des longueurs d'onde λ comprises entre 350 et 2400 nanomètres. Pour chaque prairie, on dispose d'une centaine d'hyperspectres environ. Ces données présentent la particularité d'être découpées en trois parties disjointes : 350-1350 nm, 1450-1800 nm et 2050-2400 nm, pour un total de 1698 variables explicatives. Ce découpage s'explique par une absorption athmosphérique des ondes sur les deux zones 1350-1450 nm et 1800-2050 nm. La variable réponse que l'on cherche à prédire, notée NV, représente le taux d'azote contenu dans chacune de ces prairies. Pour chacune des méthodes présentées précédemment, on obtient un modèle de la forme $Y_i = \mathcal{M}(X_i) + \varepsilon_i$, avec ε_i l'erreur associée au modèle, où $\mathcal{M}$ sélectionne les variables les plus pertinentes. Ces modèles sont ensuite comparés en utilisant le critère LOOCV relativement à la variance de la variable réponse : $LOOCVR(\mathcal{M}) = \dfrac{\sum_{i=1}^{n}(Y_i - \mathcal{M}_{-i}(X_i))^2}{var(Y)}$, où $\mathcal{M}_{-i}$ est le modèle construit à partir de l'échantillon d'apprentissage auquel on a enlevé le $i^{ième}$ individu. Pour des raisons de temps de calcul, nous nous sommes focalisé sur un échantillon d'apprentissage de taille modeste (250 hyperspectres, soit 10 hyperspectres aléatoirement sélectionnés dans chacune des 25 prairies). Concernant le réglage en pratique des paramètres de chacune des méthodes comparées, une sélection automatique a été réalisée pour le choix des paramètres s (équivalent à λ^L) pour la méthode Lasso et h pour la méthode MPDP.

TABLE 1 détaille les résultats obtenus pour cet échantillon ; la colonne intitulée "Ridge" correspond à la mise en œuvre de la méthode de régression Ridge sur ce même échantillon, servant de référence afin de comparer les méthodes de sélection de variables avec une méthode statistique standard non sélective.

Modèle	Ridge	Lasso	MPDP
Valeur choisie pour le paramètre	$\lambda^R = 10$	$s = 0,144$	$h = 2,193$
Nombre de variables sélectionnées	1698	83	7
LOOCVR	$0,46$	$0,41$	$0,29$

TABLE 1 – Comparaison des méthodes de sélection de variables Ridge, Lasso et MPDP sur le jeu de données "prairies"

On constate que la méthode MPDP donne de meilleurs résultats comparativement aux deux autres méthodes, tant sur le nombre de variables que sur la valeur du critère de validation croisée relative LOOCVR. En effet, alors que la régression Ridge conserve l'ensemble des variables et la méthode Lasso sélectionne 83 variables, la méthode MPDP sélectionne seulement 7 variables pour une valeur du critère LOOCVR inférieure aux deux autres.

FIGURE 1 représente les hyperspectres moyens des 25 prairies ainsi que les variables sélectionnées par la méthode MPDP. On constate que certaines variables sélectionnées sont situées dans une zone spectrale où les hyperspectres sont nettement distincts, contrairement aux autres, localisées dans des zones où les courbes se différencient difficilement.

4 Conclusion

Ces premiers résultats indiquent clairement la pertinence des méthodes sélectives dans le contexte de l'imagerie hyperspectrale. La méthode non-paramétrique MPDP améliore sensiblement les résultats obtenus par la méthode linéaire Lasso par l'obtention d'un modèle plus parcimonieux et possédant un pouvoir prédictif plus important. Cette étude est cependant incomplète ; des comparaisons plus fines devront être menées pour explorer la stabilité des résultats obtenus notamment en répétant plusieurs fois la construction d'échantillons-test et d'apprentissage.

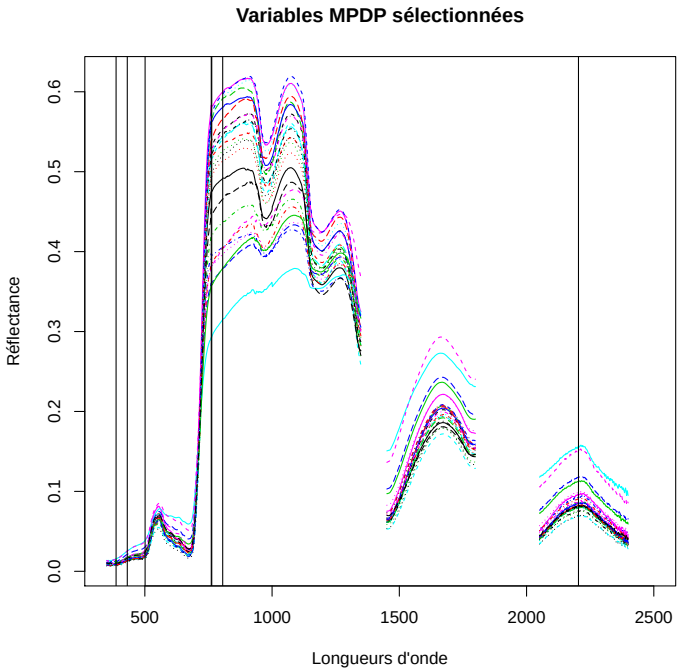

FIGURE 1 – Hyperspectres moyens des 25 prairies ; les lignes verticales localisent les longueurs d'onde correspondant aux variables sélectionnées par la méthode MPDP

Bibliographie

[1] Jimenez, L. O. et Landgrebe, D. A. (1998), Supervised classification in high-dimensional space : geometrical, statistical, and asymptotical properties of multivariate data, *Systems, Man, and Cybernetics, Part C : Applications and Reviews, IEEE Transactions on*, 28, 1, 39–54.

[2] Donoho, D. L. (2010), High-dimensional data analysis : the curses and blessing of dimensionality, *AMS Mathematical challenges of the 21st century.*

[3] Ferraty, F., Hall, P. et Vieu, P. (2010), Most-predictive design points for functional data predictors, *Biometrika*, 97, 4, 807–824.

[4] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, B58, 267–288.

[5] Hoerl, A. E. et Kennard, R. (1970), Ridge regression : biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67.

[6] Efron, B., Hastie, T., Johnstone, I. et Tibshirani, R. (2004), Least Angle Regression, *The Annals of Statistics*, 32, 2, 407–499.

[7] Fan, J. et Gijbels, I. (1996). *Local Polynomial Modeling and its Applications.* Chapman and Hall, London.