Binarsity: Prédiction en grande dimension via la sparsité induite par la binarisation de variables

ElMokhtar E. Alaya [1], Stéphane Gaïffas [2], Agathe Guilloux [3]

[1] *LSTA, Université Pierre et Marie Curie-Paris VI,*
Boîte 208, Tour 15-16, 2ème étage,
4 place Jussieu, 75252 Paris Cedex 05, France
elmokhtar.alaya@upmc.fr
[2] *CMA, Ecole Polytechnique,*
Route de Sacaly, 91128 Palaiseau Cedex, France
stephane.gaiffas@cmap.polytechnique.fr
[3] *LSTA, Université Pierre et Marie Curie-Paris VI, Unité INSERM 762 "Instabilité des Microsatellites et Cancers",*
Boîte 209, Tour 15-16, 2ème étage,
4 place Jussieu, 75252 Paris Cedex 05, France
agathe.guilloux@upmc.fr

Résumé. Nous considérons le problème d'estimation d'une fonction de régression en grande dimension. Pour cela, nous nous intéressons à la construction et à la mise en œuvre d'une nouvelle notion de sparsité nommée *binarsity*. Elle compte le nombre de valeurs différentes du vecteur de paramètres à estimer dans un espace engendré par des variables binarisées. Nous introduisons une procédure d'estimation basée sur une relaxation convexe avec poids de binarsity. Nous proposons des inégalités oracles pour cette procédure et un algorithme efficace pour la résolution du problème convexe étudié.

Mots-clés. Binarisation de variables; Variation-Totale; Inégalités d'oracle; Méthodes proximales

Abstract. We consider the problem of estimation a regression function in the high dimensional setting. For this, we introduce a new notion of sparsity called *binarsity*. It counts the number of different values of the parameter extented in the space of binarized features. We focus on an estimation procedure based on a data-driven weighted convex relaxation of binarsity. We prove oracle inequalities for this procedure. We give an algorithm that efficiently solves the convex problem studied in this work.

Keywords. Binarization of features; Total-Variation; Oracle inequalities; Proximal methods

1 Introduction

Le challenge de la grande dimension survient dans des domaines divers, en biologie computationnelle, bio-informatique, marketing digital, etc. La disponibilité des bases de données massives dans ces domaines a créé de nouveaux problèmes scientifiques qui présentent des challenges pour les méthodes d'apprentissage et d'analyse statistique. Face à ces données de grande dimension, l'inférence sous hypothèse de *sparsité* est considérée comme une technique majeure pour la réduction de la dimension et la sélection de variables, voir Bühlmann and Van De Geer (2011) et Hastie et al. (2001). Il s'agit de supposer que parmi les très nombreuses variables à notre disposition, peu d'entre elles sont en fait utiles pour expliquer les observations. Dans la littérature statistique, il existe plusieurs approches, dont celle basée sur la relaxation convexe de la sparsité considérée, la plus commune étant le Lasso introduit par Tibshirani (1996), que l'on peut comprendre comme un relaxation du nombre de coefficients non nuls, le groupe Lasso Yuan and Lin (2006) qui prend en compte la structure de groupe dans les variables, et Simon et al. (2013) qui considère une combinaison de group Lasso et Lasso, et enfin le fused Lasso Tibshirani et al. (2005), qui encourage la sparsité dans le gradient discret des paramètres du modèle, parmi beacoup d'autres approches.

Dans ce travail, on introduit une nouvelle notion de sparsité nommée *binarsity* qui compte le nombre de valeurs différentes du vecteur de paramètres à estimer dans un espace engendrée par des variables binarisées. Notons $\boldsymbol{X} = [\boldsymbol{X}_{i,j}]_{1 \leq i \leq n; 1 \leq j \leq p}$ la matrice des observations d'entrée avec n exemples et p variables, et Y le vecteur des sorties observées. Nous considérons une collection $D_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ de copies *i.i.d.* d'une variable (X, Y) à valeurs dans $\mathbb{R}^p \times \mathcal{Y}$. L'espace $\mathcal{Y}$ peut être un sous ensemble de $\mathbb{R}$ ou $\{0, 1\}$ pour une réponse binaire.

2 Binarisation de variables et binarsity

On utilise la notation suivante: $\boldsymbol{X}_{\bullet,j}$ la j-ème variable en colonne et $\boldsymbol{X}_{i,\bullet}$ la i-ème variable en ligne de $\boldsymbol{X}$.

2.1 Binarisation

La matrice binarisée de $\boldsymbol{X}$ notée $\boldsymbol{X}^B$ est la matrice à d colonnes, avec d beaucoup plus grand que p, et est telle que la j-ème colonne $\boldsymbol{X}_{\bullet,j}$ est remplacée par d_j colonnes $\boldsymbol{X}^B_{\bullet,j,1}, \ldots, \boldsymbol{X}^B_{\bullet,j,d_j}$ ne contenant que des 0 ou 1. Si la colonne $\boldsymbol{X}_{\bullet,j}$ prend des valeurs discretes dans un ensemble de modalités $\{1, \ldots, M_j\}$, alor son pose $d_j = M_j$ et

$$\boldsymbol{X}^B_{i,j,k} = \begin{cases} 1 & \text{if } \boldsymbol{X}_{i,j} = k \\ 0 & \text{sinon} \end{cases}$$

pour $i = 1, \dots, n$ et $k = 1, \dots, d_j$. Si la colonne $\boldsymbol{X}_{\bullet,j}$ est quantitative, alors on considère une partition d'intervalles formés par des quantiles de cette dernière, $I_{j,1}, \dots, I_{j,d_j}$, tels que pour tout $k = 1, \dots, d_j$, $I_{j,k} = \left[q_j(\frac{k-1}{d_j}), q_j(\frac{k}{d_j})\right)$ avec $q_j(\alpha)$ le quantile d'ordre α de $\boldsymbol{X}_{\bullet,j}$ et on pose

$$\boldsymbol{X}^B_{i,j,k} = \begin{cases} 1 & \text{if } \boldsymbol{X}_{i,j} \in I_{j,k} \\ 0 & \text{sinon.} \end{cases}$$

Ce passage de la matrice $\boldsymbol{X}$ à la matrice $\boldsymbol{X}^B$ s'appelle binarisation, et est une technique très utilisée dans de nombreux domaines, notamment en marketing digital. L'idée est qu'en "éclatant" une variable en plusieurs variables binaires, on obtient une meilleure attache au données, par une réponse non-linaire par rapport aux variables d'origine. Nous introduisons alors une pénalisation forcant les coefficients associé à une variable binarisée à ne pas prendre un trop grand nombre de valeurs différentes.

2.2 Binarsity

A chaque variable binarisée $\boldsymbol{X}^B_{\bullet,j,k}$ correspond un coefficient $\theta_{j,k}$. Le paramètre de la binarisation de la j-ème variable est un vecteur noté $\theta_{j,\bullet} = [\theta_{j,1} \cdots \theta_{j,d_j}]^\top$, et on considère la concatenation de ces vecteurs en un vecteur de taille $d = \sum_{j=1}^p d_j$:

$$\theta = [\theta_{1,\bullet}^\top \cdots \theta_{p,\bullet}^\top]^\top = \left[\theta_{1,1} \cdots \theta_{1,d_1}\, \theta_{2,1} \cdots \theta_{2,d_2} \cdots \theta_{p,1} \cdots \theta_{p,d_p}\right]^\top.$$

La notion de sparsité que l'on cherche à induire est alors la suivante: tout bloc $\theta_{j,\bullet}$ peut être nul ou contient un nombre assez petit de valeurs différentes, ce qui est mesuré par la notion de *binarsity*:

$$\text{binarsity}(\theta) = \sum_{j=1}^{p} \left(\mathbb{1}_{\theta_{j,1} \neq 0} + \sum_{k=2}^{d_j} \mathbb{1}_{\theta_{j,k} \neq \theta_{j,k-1}} \right).$$

Si la j-ème variable est statistiquement non pertinente pour la prédiction, alors le bloc $\theta_{j,\bullet}$ qui lui correspond est nul, et dans ce cas sa contribution à la binarsité est égale à zéro. Si la j-ème variable est pertinente alors le nombre de valeurs différentes dans le bloc $\theta_{j,\bullet}$ doit être assez petit pour un bon compris biais-variance.

3 Résumé des résultats obtenus

Nous introduisons une relaxation convexe de binarsity(θ) en suivant l'approche introduite dans Chandrasekaran et al. (2012), qui consiste à considérer la norme atomique obtenue par l'ensemble des atomes décrivant binarsity(θ). Cela mène à une relaxation convexe qui dépend en partie de la variation totale de chaque bloc de coefficients. Puis nous

analysons cette pénalisation dans un cadre d'apprentissage supervisé général: pour une fonction de perte $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty]$, nous considérons le risque empirique $R_n(\theta) = 1/n \sum_{i=1}^n \ell\big(Y_i, \langle \boldsymbol{X}^B_{i,\bullet}, \theta \rangle\big)$, que nous pénalisons par la relaxation convexe obtenue pour binarsity(θ), et que nous réglons avec des pondérations observables. Nous établissons des inégalités d'oracle dans ce cadre. En utilisant des algorithmes proximaux Bauschke and Combettes (2011), nous donnons un algorithme efficace et rapide pour calculer l'operateur proximal de la pénalité induite par binarsity.

Bibilographie

Bauschke, H. H. and P. L. Combettes (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York. With a foreword by Hédy Attouch.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Heidelberg: Springer. Methods, theory and applications.

Chandrasekaran, V., B. Recht, P. A. Parrilo, and A. S. Willsky (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics 12*(6), 805–849.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The elements of statistical learning*. Springer Series in Statistics. New York: Springer-Verlag. Data mining, inference, and prediction.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22*(2), 231–245.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*(1), 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B 68*, 49–67.