Sélection automatique de blocs de variables

Pierre Alquier

Laboratoire de Probabilités et Modèles Aléatoires
Université Paris 7, site Chevaleret, case 7012
175, rue du Chevaleret
75205 Paris Cedex 13 - FRANCE
&
Laboratoire de Statistique du CREST
Timbre J350
3, Avenue Pierre Larousse
92240 Malakoff - FRANCE

Résumé: En génomique apparaissent certains modèles de classification et de régression avec un paramètre β parcimonieux (c'est-à-dire que la plupart des coordonnées β_i sont nulles[1]) et constant par morceaux, voir par exemple les articles de Rapaport, Barillot et Vert (2008) et Huang, Salim, Lei, O'Sullivan et Pawitan (2009). Des méthodes d'estimation ont été proposées dans ce contexte, par exemple le *fused LASSO* de Tibshirani, Saunders, Rosset, Zhu et Knight (2005). L'idée de cet estimateur est d'adapter la pénalité dans l'estimateur LASSO de Tibshirani (1996) en forçant β_i à être égal à β_{i+1} dans beaucoup de cas, ce qui permet de sélectionner des groupes de variables *sans les connaître a priori*. Dans l'article Alquier (2010), on a proposé une méthode dont l'idée est d'utiliser une adaptation de la forme duale du LASSO, donnée dans Osborne, Presnell et Turlach (2000). On présente ici brièvement cette méthode, puis on donne quelques résultats d'application à des données CGH.

Abstract: In genomics applications, it makes sense to consider classification or regression estimation with a parameter β "sparse" (most of its coordinates are 0) and "blocky" (β_i and β_{i+1} are likely to be equal). We refer the reader for example to the paper of Rapaport, Barillot and Vert (2008) or Huang, Salim, Lei, O'Sullivan and Pawitan (2009). Some estimation methods taking this information into account are available, such as the fused LASSO, by Tibshirani, Saunders, Rosset, Zhu, and Knight (2005). This method relies on an adaptation of the penalty in the well known LASSO, defined by Tibshirani (1996), to selection of *unknown* blocks of features. In the paper Alquier (2010), we proposed another estimation method, that relies on an adapatation of the dual form of the LASSO provided by Osborne, Presnell and Turlach (2000). We describe shortly this method, and provide some applications to array-CGH data.

Mots-clés: sélection de variables, modèles parcimonieux, sparsité, régression linéaire, groupes de variables, tableaux CGH.

[1]Il est devenu courant d'utiliser l'anglicisme *sparse* dans ce cas, et *sparsité* pour parcimonie.

1 Introduction

On se place dans le modèle de régression linéaire gaussienne

$$Y \sim \mathcal{N}\left(X\beta, \sigma^2 I_n\right) \tag{1}$$

où pour la simplicité de la présentation, X est une matrice déterministe $n \times p$, et $\beta = (\beta_1, ..., \beta_p)'$ avec possiblement $p > n$ (on utilise la convention que si $u \in \mathbb{R}^p$, u est un vecteur colonne et u' sa transposée). On note X_1, ..., X_p les colonnes de X, et $\mathbf{P}$ la loi de Y donnée par (1). On suppose que les données sont renormalisées de façon à avoir $X_j' X_j / n = 1$ pour tout j. Soit le critère d'erreur

$$L(b) = \|X(b - \beta)\|_2^2 .$$

Sans hypothèses supplémentaires, il est impossible d'estimer correctement β. Une hypothèse courante est de supposer que le nombre de coordonnées non-nulles dans β est petit (hypothèse de sparsité). Dans ce cas, l'estimateur LASSO

$$\hat{\beta}_{LASSO}(\lambda) = \arg\min_b \left[\|Y - Xb\|_2^2 + \lambda \sum_{j=1}^{p} |b_j| \right]$$

pour un $\lambda > 0$, défini par Tibshirani (1996), ou le Dantzig selector, de Candès et Tao (2007), ont de bonnes performances en pratique et sont bien compris en théorie - voir par exemple Candès et Tao (2007) et Bickel, Ritov et Tsybakov (2009) pour les résultats théoriques. La sparsité a de plus souvent un sens pour le praticien, en terme de sélection de variables.

En génomique, dans certains modèles, cette hypothèse est naturelle si elle est combinée avec l'existence de "blocs de coefficients" dans β, c'est-à-dire que le nombre de j tels que $\beta_j \neq \beta_{j+1}$ est aussi petit. Par exemple,

$$\beta = (0, 0, 0, 0, 0, 0, 0, 0, 5, 5, 5, 5, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 2, 2, 0, 0, 0, 0)' .$$

Ceci est par exemple le cas dans les données de type CGH, où les index j sont des positions de "sondes" le long de chromosomes. On observe la quantité d'ARN correspondant à chaque sonde de façon très bruitée: $Y_j = \beta_j + \varepsilon_j$ (dans ce cas, $X = I_n$). Dans des cellules cancéreuses, des portions entières de chromosomes peuvent subir des amplifications ou des délétions, ce qui se traduira respectivement par une augmentation où une diminution de toute une série de β_j consécutifs. Pour un chromosome d'une cellule saine, $\beta_j = 0$ pour tout j.

Pour traiter ce cas, adapter la pénalité du LASSO de la façon suivante semble naturel:

$$\hat{\beta}_{F-LASSO}(\lambda, \mu) = \arg\min_b \left[\|Y - Xb\|_2^2 + \lambda \sum_{j=1}^{p} |b_j| + \mu \sum_{j=2}^{p} |b_j - b_{j-1}| \right]$$

pour $\lambda, \mu > 0$. Cet estimateur a été introduit par Tibshirani, Saunders, Rosset, Zhu, et Knight (2005) sous le nom de *fused LASSO*. On pourra cependant remarquer la ressemblance avec les méthodes de débruitage d'images utilisant la variation totale, cf. Mumford et Shah (1989) ou Rudin, Osher et Fatemi (1992). Dans des articles de bioinformatique et de biologie, de telles pénalisations ont effectivement été utilisées, cf. par exemple les articles de Rapaport, Barillot et Vert (2008) et Huang, Salim, Lei, O'Sullivan et Pawitan (2009).

On peut comparer par exemple le fused LASSO au *group LASSO* introduit par Yuan et Lin (2006): contrairement au fused LASSO, le group LASSO ne peut fonctionner si on ne connait pas a priori la position des groupes de variables. La même remarque s'applique á la méthode de Zhao, Rocha et Yu (2009). En contrepartie, les garanties théoriques sur le group LASSO sont plus générales, voir Chesnau et Hebiri (2008). Dans le cas du fused LASSO, les seuls résultats disponibles sont dans le cas $X = I_n$, prouvés par Rinaldo (2009).

En pratique, le fused LASSO peut être calculé rapidement par une optimization coordonnée par coordonnée, cf. Friedman, Hastie, Höfling et Tibshirani (2007). Plus récemment, des algorithmes capables de calculer rapidement toutes les solutions (pour toutes les valeurs de λ et μ) sont apparus. Toutes reposent sur une généralisation de l'algorithme LARS de Efron, Hastie, Johnstone, et Tibshirani (2004): Hoefling (2009) ou Tibshirani et Taylor (2009). Toujours en pratique, on pourra aussi considérer le *smooth LASSO* de Hebiri (2008), étudié en détail par Hebiri et Van de Geer (2010), comme une alternative au fused LASSO, qui, bien que conçu pour un objectif légèrement différent, est souvent plus stable numériquement:

$$\hat{\beta}_{F-LASSO}(\lambda, \mu) = \arg\min_b \left[\|Y - Xb\|_2^2 + \lambda \sum_{j=1}^{p} |b_j| + \mu \sum_{j=2}^{p} (b_j - b_{j-1})^2 \right].$$

Cet estimateur peut aussi être vu comme un cas particulier de l'estimateur *strictured elastic net* de Slawski, zu Castell et Tutz (2010).

Dans l'article Alquier (2010), on propose un estimateur alternatif dans ce contexte. Celui-ci ne repose pas sur une modification de la pénalisation du LASSO, mais de la forme duale du LASSO donnée par Osborne, Presnell et Turlach (2000). On présente ici cet algorithme (Section 2), ensuite, on rappele un résultats théoriques sur cet algorithme (Section 3). On donne enfin une illustration sur des données réelles.

2 L'algorithme

L'algorithme est itératif: il part de $\hat{\beta}^{(0)} = (0, \ldots, 0)$ et à chaque étape, il met à jour un groupe de coordonnées. On décrit maintenant cette mise à jour. On suppose donc que $\hat{\beta}^{(m)}$ est donné et on va définir $\hat{\beta}^{(m+1)}$.

On commence par quelques notations. Soit la fonction de seuillage "doux" $\forall x \in \mathbb{R}, \forall u \geq 0, \gamma(x,u) = \text{sign}(x)(|x|-u)_+$. Pour tout $j \in \{1,...,p\}$ et $k \in \{1,...,p-j+1\}$ on définit $\mathbf{1}_{j,k} \in \mathbb{R}^p$ par $(\mathbf{1}_{j,k})_i = 1$ si $i \in \{j,...,j+k-1\}$, et $(\mathbf{1}_{j,k})_i = 0$ sinon, autrement dit si i appartient au groupe de coordonnées de longueur k, commençant à la position j. Choisissons un $s > 0$ (on discute ce choix dans la section suivante).

Pour tout groupe défini par la position initiale j et la longueur k, on définit le vecteur $\tilde{\beta}^{(m+1),j,k}$ par:

$$\left\{ \begin{array}{c} \tilde{\beta}_j^{(m+1),j,k} = \hat{\beta}_j^{(m)} + \gamma\left(\frac{(Y - X\hat{\beta}^{(m)})' \sum_{h=j}^{j+k-1} X_h}{\mathbf{1}'_{j,k} X'X \mathbf{1}_{j,k}}, \frac{s}{\sqrt{\mathbf{1}'_{j,k} X'X \mathbf{1}_{j,k}}} \right) \\ \vdots \\ \tilde{\beta}_{j+k-1}^{(m+1),j,k} = \hat{\beta}_{j+k-1}^{(m)} + \gamma\left(\frac{(Y - X\hat{\beta}^{(m)})' \sum_{h=j}^{j+k-1} X_h}{\mathbf{1}'_{j,k} X'X \mathbf{1}_{j,k}}, \frac{s}{\sqrt{\mathbf{1}'_{j,k} X'X \mathbf{1}_{j,k}}} \right) \end{array} \right. \quad (2)$$

et $\tilde{\beta}_\ell^{(m+1),j,k} = \hat{\beta}_\ell^{(m)}$ pour tout $\ell \notin \{j,...,j+k-1\}$.

On doit ensuite choisir un couple (j,k), on pose simplement:

$$\hat{\beta}^m = \arg\max_{j,k} \|X(\tilde{\beta}^{(m+1),j,k} - \hat{\beta}^{(m)})\|_2^2.$$

3 Un premier résultat théorique

Le résultat suivant, prouvé dans Alquier (2010), constitue un premier pas dans l'analyse des propriétés statistiques de cet algorithme. Il garantit que l'on peut itérer autant de fois que l'on veut l'étape de mise à jour sans risquer le surapprentissage.

Théorème: *On a, pour tout* $M \in \mathbb{N}$,

$$\mathbf{P}\left[L(\hat{\beta}^{(M)}) \leq L(\hat{\beta}^{(M-1)}) \leq \cdots \leq L(\hat{\beta}^{(0)}) \right] \geq 1 - p^2 e^{-\frac{s^2}{2\sigma^2}}.$$

Ceci signifie que si l'on choisit $s = 2\sigma[\log(p/\varepsilon)]^{1/2}$, alors avec probabilité au moins $1-\varepsilon$, à chaque étape le risque de l'estimateur diminue.

La preuve repose sur les considérations géométriques sur le LASSO donnés dans Osborne, Presnell et Turlach (2000) et Alquier (2008a,2008b).

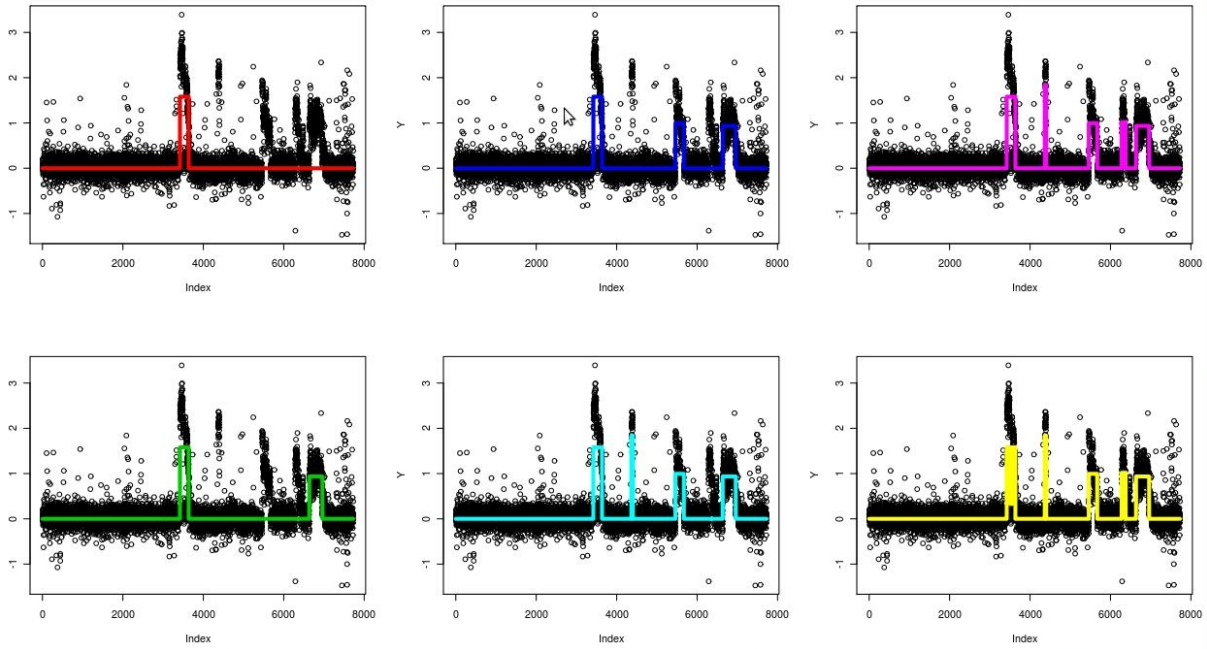
4 Test sur des données simulées

Pour des raisons de place, on ne présente pas ici en détail les simulations effectuées dans Alquier (2010). Notons que si cette étude expérimentale est assez brève, il en ressort que dans certains cas, les performances de l'algorithme de sélection itétarive de blocs de variables présenté ici sont une réelle amélioration par rapport à celles du fused LASSO.

5 Applications sur des données CGH

La Figure 1 montre un exemple d'itération sur des données réelles: le chromosome 17 d'un patient atteint d'un type particulier de cancer. On est ici dans le cas particulier présenté en introduction $Y_i = \beta_i + \varepsilon_i$, $1 \leq i \leq n = 7728$.

Figure 1: Un exemple d'itération de l'algorithme.



Tous les calculs ont été effectués avec le logiciel R, cf. R Development Core Team (2008). Les codes sont disponibles auprès de l'auteur. Les données ont été fournies par le Dr. Jean-Paul Feugeas (INSERM U944, Institut Universitaire d'Hématologie) que je tiens a remercier particulièrement pour avoir motivé ce travail.

Bibliographie

[1] Alquier, P. (2008a), Iterative feature selection in least square regression estimation, *Annales de l'IHP* B (Probas. Stats.), 44(1), 47–88.
[2] Alquier, P. (2008b), LASSO, iterative feature selection and the correlation selector: oracle inequalities and numerical perfomances, *Elec. Journ. of Statistics*, 2, 1129–1152.
[3] Alquier, P. (2010), An algorithm for iterative selection of blocks of features, *Proceedings of ALT'10*, M. Hutter, F. Stephan, V. Vovk & T. Zeugmann Edts., LNAI, Springer, 35–49.
[4] Bickel, P., Ritov, Y. and Tsybakov, A. (2009), Simultaneous analysis of LASSO and Dantzig selector, *The Annals of Statistics*, 37(4), 1705–1732.
[5] Candès, E. and Tao, T. (2007), The Dantzig selector: statistical estimation when p is much larger than n, *The Annals of Statistics*, 35(6), 2313–2351.

[6] Chesnau, C. and Hebiri, M. (2008), Some theoretical results on the grouped variables LASSO, *Mathematical Methods in Statistics*, 17(4), 317–326.
[7] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), Least angle regression, *The Annals of Statistics*, 32(2), 407–499.
[8] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007), Pathwise coordinate optimization, *The Annals of Applied Statistics*, 1(2), 302–332.
[9] Hebiri, M. (2008), Regularization with the smooth LASSO procedure, Preprint, arXiv: 0803.0668.
[10] Hebiri, M. and Van de Geer, S. (2010), The smooth LASSO and other $\ell_1+\ell_2$-penalized methods, Preprint, arXiv: 1003.4885.
[11] Hoefling, H. (2009), A path algorithm for the fused LASSO signal approximator, Preprint, arXiv: 0910.0526.
[12] Huang, J., Salim, A., Lei, K., O'Sullivan, K. and Pawitan, Y. (2009), Classification of array CGH data using smoothed logistic regression model, *Statistics in Medicine*, 8(30), 3798–3810.
[13] Mumford, D. and Shah, J. (1989), Optimal approximations by piecewise smooth functions and associated variational problems, *Communications on Pure and Applied Mathematics*, 42(5), 57–685.
[14] Osborne, M., Presnell, B. and Turlach, B. (2000), On the LASSO and its dual, *Journal of Computational and Graphical Statistics*, 9(2), 319–337.
[15] R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, *http://www.R-project.org*
[16] Rapaport, F., Barillot, E. and Vert, J.-P. (2008), Classification of array-CGH data using fused SVM, *Bioinformatics*, 24(13), 1375–1382.
[17] Rinaldo, A. (2009), Properties and Refinements of the fused LASSO, *The Annals of Statistics*, 37(5B), 2922–2952.
[18] Rudin, L., Osher, S. and Fatemi, E. (1992), Nonlinear total variation based noise removal algorithm, *Physica* D (Nonlinear Phen.), 60, 259–268.
[19] Slawski, M., zu Castell, W. and Tutz, G. (2010), Feature selection guided by structural information, *The Annals of Applied Statistics*, 4(2), 1056–1080.
[20] Tibshirani, R. (1996), Regression shrinkage and selection via the LASSO, *JRSS* B (Method.), 58(1), 267–288.
[21] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005), Sparsity and smoothness via the fused LASSO, *JRSS* B (Method.), 67(1), 91–108.
[22] Tibshirani, R. J. and Taylor, J. (2009), Regularization path for least squares problems with generalized ℓ_1 penalties, Preprint.
[23] Yuan, M. and Lin, Y. (2006), Model selection and estimation in regression with grouped variables, *JRSS* B (Method.), 68(1), 49–67.
[24] Zhao, P. and Rocha, G. and Yu, B. (2009), The composite absolute penalties for grouped and hierarchical variable selection, *The Annals of Statistics*, 37(6A), 3468–3497.