Building genetically inspired constraints with the Parsimony framework

17 avril 2014

Résumé

1 Les données et les contraintes

1.1 Websters et résultats univariés

Données décrites dans les papiers [3, 6]. Voir aussi le rapport de V Guillemot sur l'étude que nous avions faite pour se caler sur le papier séminal de 2009 concernant l'association avec le GE.

Les données contiennent des SNPs, de l'expression de gène et le statut clinique pour 380 (TOFIX) personnes. pour l'expression, des données considérées sont les celles de l'article qui supprime les effets confondants (statut Apo, TOFIX). Pour la regression logistique on considère le statut (Alz, Ctl). Pour la régression linéaire on considère l'expression du gène KIF1B qui ressort dans le papier comme gène s'associant avec des SNPs en univarié lorsque la maladie est avérée : l'interprétation exacte du papier [6] est très difficile !

1.2 Contraintes

Les contraintes considérées sont de plusieurs natures :

1.3 Contraintes de groupes

Rendues par du GroupLasso. On considère les informations :

- GeneOntology (group avec overlap massif car en tant qu'ontologie il s'agit de groupe organisés hiérarchiquement),
- pathway Keggs ou autre base de collection de gènes : c'est aussi du group avec overlap. Cette base est utilisée par Montana et coll. [4] et aussi par Chen et coll. [1]

On pourrait considérer la prise en compte d'une contrainte de type fused lasso (flou pour ce qui me concerne).

1.4 Contraintes de régions recombinantes

Il s'agit de rendre compte de la plus ou moins grande probabilité de recombinaison observée sur le génome. Les zones à faible probabilité pourraient être traitées avec une approche Group-TV : de la cohérence serait demandée dans de telles régions.

La description des données de probabilité de recombinaison est donnée dans

2 Travaux en cours

2.1 Approches GL

Les pénalités travaillées sont Ridge + L1 dans le cas de la regression logistique.

- Montage des contraintes repose sur inst_bioresource OK pour gene et snp tedious pour les pathways
- Mise en place d'une visu des résultats
- Mise en place d'un cadre pour le choix des hyperparametres. Map Reduce et usage de Gabriel. Implémenté par VF
- Passage a l'échelle pour le nombre de pathways
- Choix des poids a appliquer à chacun des pathway. ici il y a une ambiguité explicitée dans la Figure 2 . Implémenté par FHS

2.1.1 Adaptive Group Lasso

L'idée est de fonder la choix des parametres de poids dans le terme de GroupLasso. FHS a recensé deux papier [5, 2]. Ces méthodes ont été décrites dans le cadre de la régression linéaire. Dans ce cadre, le paramétrage revient à : (i) un run en moindre carré (qd on est en version regression linéaire) et partir des beta estimé comme estimateur non biaisés (entachés de variance), (ii) calculer les poids du GL et les hyper-paramètres du $L2$, $L1$ suivant la formule Eq. 1.

$$w_{\mathbf{G}} = \frac{1}{||\beta_{\mathbf{G}}||^{-\gamma}} \tag{1}$$

Il reste deux paramtrers : le paramètre α global (lien avec SNR) et le parametre γ qui peut varier dans $[0, 1]$

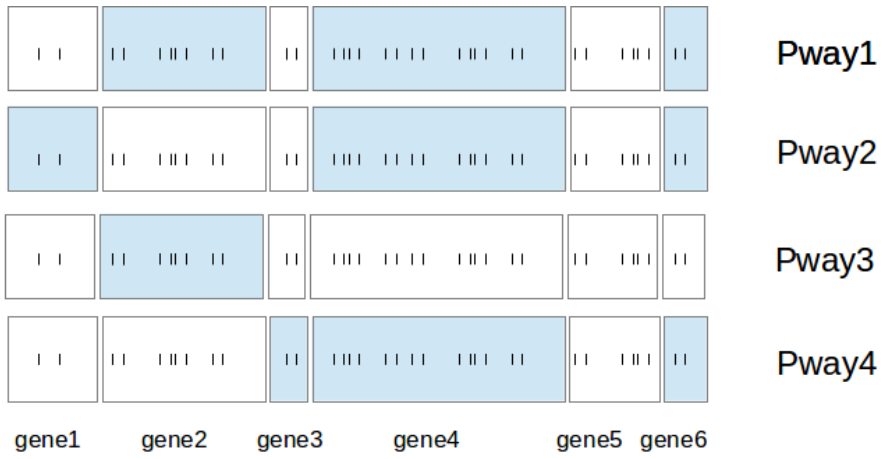

FIGURE 1 – Illustration de la difficulté du choix des poids pour un pathway donné. Les régions bleus dans un pathway sont les gènes qui sont gardés. Les SNPs sont les petits bâtons. Pour un pathway de meme longueur il peut etre consitué de gènes plus ou moins long. Il semble que les pathways contenant des gènes longs sont favorisés.

Un premier résultat sur les données Websters en régression logistique avec les 10 premiers pathways (en fait 9 car un pathway avait 0 gènes) tirés du terme GO *synapse*. 539 SNPs TOFIX

genes. Comme il n'y a pas de Lasso aucun beta n'est nul. Un seuillage des petites contributions est faite. Les traits longs correspondent aux SNPs qui passent le seuil ; les autres sont donnés pour information en trait court. Le choix du seuil et fait de telle manière que les teme du vecteur beta qui sont seuillés représentent mois de 1% de la norme2 du vecteur beta.

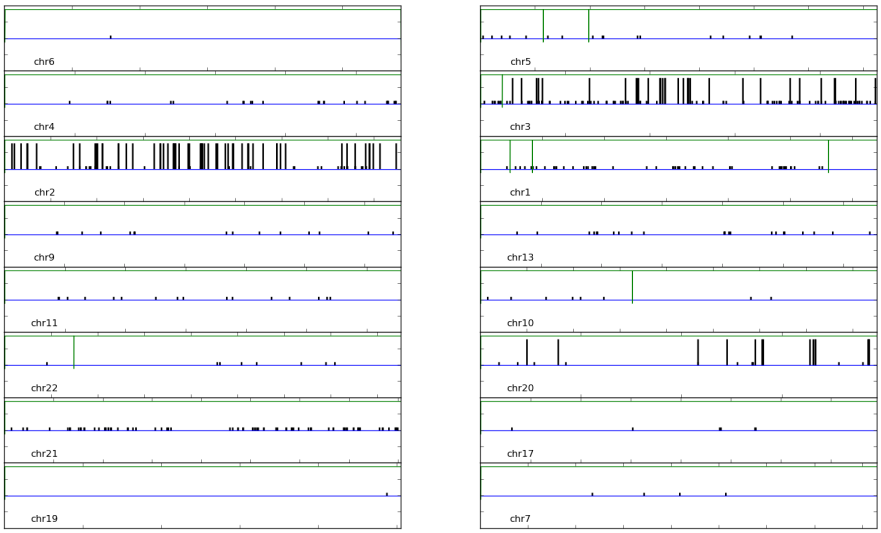

FIGURE 2 – Résultats d'un run de rRidgeGL.

Références

[1] Lin S Chen, Carolyn M Hutter, John D Potter, Yan Liu, Ross L Prentice, Ulrike Peters, and Li Hsu. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *American journal of human genetics*, 86(6) :860–71, June 2010.

[2] Jian Huang, Patrick Breheny, and Shuangge Ma. A Selective Review of Group Selection in High-Dimensional Models. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 27(4), January 2012.

[3] Amanda J Myers, J Raphael Gibbs, Jennifer A Webster, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Doris Leung, Leslie Bryden, Priti Nath, Victoria L Zismann, Keta Joshipura, Matthew J Huentelman, Diane Hu-Lince, Keith D Coon, David W Craig, John V Pearson, Peter Holmans, Christopher B Heward, Eric M Reiman, Dietrich Stephan, and John Hardy. A survey of genetic human cortical gene expression. *Nature genetics*, 39(12) :1494–9, December 2007.

[4] Matt Silver, Eva Janousova, Xue Hua, Paul M Thompson, and Giovanni Montana. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *NeuroImage*, 63(3) :1681–1694, November 2012.

[5] Hansheng Wang and Chenlei Leng. A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12) :5277–5286, August 2008.

[6] Jennifer A Webster, J Raphael Gibbs, Jennifer Clarke, Monika Ray, Weixiong Zhang, Peter Holmans, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Donald S McCorquodale, Cindy Cuello, Doris Leung, Leslie Bryden, Priti Nath, Victoria L Zismann,

Keta Joshipura, Matthew J Huentelman, Diane Hu-Lince, Keith D Coon, David W Craig, John V Pearson, Christopher B Heward, Eric M Reiman, Dietrich Stephan, John Hardy, and Amanda J Myers. Genetic control of human brain transcript expression in Alzheimer disease. *American journal of human genetics*, 84(4) :445–58, April 2009.