Examen de biostatistique : Etude du cancer de la prostate

Ce projet s'attaque à la prédiction d'échantillons normaux vs. tumoraux à partir de l'expression de plusieurs milliers de gènes (quantité d'ARN produit par chaque gène). Vous trouverez le jeu de données réelles à l'adresse suivante (attention le fichier s'auto-détruira le 07/04/2019, date à laquelle vous devrez remettre votre projet) et un papier permettant de remettre le projet dans son contexte médical :

https://filesender.renater.fr/?s=download&token=03d59813-287d-fa89-5c94-10d4846a0dcd

Le jeu de données `prostate` est caractérisé par 52 tumeurs de la prostate et 50 échantillons normaux. La classe de l'échantillon est encodée dans le vecteur $\mathbf{y}$ par la valeur 0 pour les échantillons normaux et 1 pour les échantillons tumoraux. La matrice de variables explicatives contient l'expression de 6033 gènes. Ces données ont été prélabablement normalisées, log-transformation et standardisées (i.e. centrées et réduites) par échantillon. Ces normalisations spécifiques sont justifiées dans [Dettling and Beuhlmann, 2002] et [Dettling, 2004].

On souhaite comprendre, décrire et prédire les relations complexes entre le statut de l'échantillon et l'expression des gènes. On travaillera dans toute la suite sur des **variables centrées et réduites**.

Question 1. Réaliser une analyse en composantes principales et visualiser les individus sur le premier plan principal ; commentez. Reportez les *10* gènes qui participent le plus à la construction des composantes principales 1 et 2.

Question 2. Réaliser une analyse massivement univariée pour identifier les gènes différentiellement exprimées entre patients et contrôle. On veillera à appliquer une correction de type bonferonni ou fdr. Reportez le nombres de gènes différentiellement exprimés avant et après correction.

Question 3. Nous souhaitons dans cette question résoudre le problème d'optimisation de la régression PLS parcimonieuse (sPLS) défini comme suit :

$$\max \operatorname{cov}(\mathbf{Xa}, \mathbf{y}) \text{ s.c. } \|\mathbf{a}\|_2^2 = 1 \ \& \ \|\mathbf{a}\|_1 < s \tag{1}$$

Montrer que la solution optimale du problème sPLS est donnée par :

$$\mathbf{a}^* = \frac{S(\frac{1}{n}\mathbf{X}^\top \mathbf{y}, \lambda)}{\|S(\frac{1}{n}\mathbf{X}^\top \mathbf{y}, \lambda)\|_2}$$

où $\mathbf{X}$ est le tableau de variables explicatives standardisées, $\mathbf{y}$ la variable à expliquer centrée, S l'opérateur de seuillage doux defini par $S(a, \lambda) = \text{sign}(a) max(0, |a| - \lambda)$ et λ un scalaire choisi (par binary search) de sorte à satisfaire la contrainte ℓ_1 du problème d'optimisation de sPLS. Dans la suite, on fixera le degré de sparsité en ajustant directement la valeur de λ plutôt que la valeur de s.

Question 4. Implémenter une fonction ayant pour arguments d'entrée ($\mathbf{X}$, $\mathbf{y}$ et λ) et renvoyant $\mathbf{a}^*$ et la première composantes sPLS, $\mathbf{Xa}^*$.[1]

Question 5. Il est tout à fait possible de construire les composantes sPLS suivante "par deflation" en appliquant le problème d'optimisation (1) sur une "matrice résiduelle". Construire une fonction permettant de calculer les dimensions successives de sPLS.

Question 6. Appliquer l'algorithme sPLS sur le jeu de données `prostate` et visualiser, pour une grille de valeurs de λ, les individus sur le plan défini par les deux premières composantes sPLS. Commentez.

[1] en cas d'échec, vous pourrez utiliser une bibliothèque existante (e.g. 'RGCCA', 'spls' ou 'mixOmics')

Question 7. Par analyse discriminante, il est possible de construire un modèle prédictif du statut du patient à partir des composante(s) sPLS. Tracer le taux d'erreur cross-validé en fonction du paramètre de sparsité et du nombre de composantes sPLS. Reportez le couple (paramètre de sparsité, nombre de composantes) optimal.

Question 8. Construire un modèle de régression logistique pénalisée ℓ_1 (on pourra utiliser le package `glmnet` sour R) permettant de prédire le statut du patient à partir de l'expression des gènes. On veillera à retenir le paramètre de sparsité par validation croisée. Comparer le modèle de régression logistique pénalisée obtenu au modèle sPLS obtenu à la question précédente (en termes de qualité de prédiction et de variables sélectionnées.)

Références

Dettling M (2004), "BagBoosting for tumor classification with gene expression data", Bioinformatics, Vol. 20, pp. 3583-3593.

Dettling M and Beuhlmann P (2002), "Supervised clustering of genes", Genome Biology, 3(12), research0069.1.