TP7
Grande dimension, tests multiples

Objectifs du TP

— comprendre les problématiques spécifiques aux jeux de données de grande taille (temps de calcul et difficultés statistiques),
— mettre en œuvre la sélection de variables sur un jeu de données de grande dimension,
— comprendre pourquoi il est difficile d'interpréter le résultat d'un grand nombre de tests sans tenir compte de ce nombre,
— mettre en œuvre des corrections de p-valeurs : Bonferroni, Benjamini-Yekutieli.

Étude de données d'expression de gènes

Le jeu de données `liver` contient des mesures du niveau de toxicité du foie chez le rat (via le niveau de cholestérol) ainsi que des mesures du niveau d'expression de quelques milliers de gènes.

Lors du chargement des données (fichier `donnees_liver.rda`), le tableau `liver` est créé avec 64 lignes (les observations) et 3117 colonnes. La première colonne `cholesterol` est la variable à expliquer. Les 3116 restantes sont les expressions de 3116 gènes (plus précisément, le logarithme du rapport entre les niveaux d'expression dans deux conditions expérimentales).

Référence : *Bushel, P., Wolfinger, R. D. and Gibson, G. (2007). Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. BMC Systems Biology.*

On cherche à identifier les variables (et donc les gènes) ayant un lien avec la réponse, ainsi qu'à construire un modèle linéaire capable de bien prévoir la valeur de la réponse.

Un code R succint, qui permet de mettre en œuvre la plupart des méthodes mentionnées ci-dessous, est disponible (fichier `TP7_2021_aide.R`). L'objectif principal est donc d'interpréter les résultats.

1. Charger les données du fichier `donnees_liver.rda` et procéder à une analyse descriptive rapide (et forcément partielle). Faire bien attention à la taille du jeu de données !
2. **Sélection forward.**
   Mettre en œuvre la procédure de sélection de variables par recherche forward, en se limitant aux modèles faisant intervenir 6 variables ou moins. On prêtera attention au temps de calcul induit par cette recherche.
3. **Lasso.**
   À l'aide du package `glmnet` (que l'on installera si besoin), calculer le chemin de régularisation $(\widehat{\beta}_\lambda)_{\lambda>0}$ de l'estimateur Lasso.
   Visualiser l'ensemble du chemin (avec la fonction « plot »), puis l'ensemble des variables sélectionnées lorsque $\lambda = \sqrt{\log(p)}$ (qui n'est pas nécessairement un bon choix).
   Que proposeriez-vous pour choisir λ ?
4. **Tests multiples.**
   On souhaite mettre en évidence un ensemble de gènes reliés à la variable réponse. Pour cela, on propose de tester, pour chaque variable (considérée seule), si elle a un effet linéaire sur la réponse.

(a) Pour la j-ème covariable ($j \in \{1, \dots, p\}$), préciser quelle est l'hypothèse nulle $H_{0,j}$ et l'hypothèse alternative $H_{1,j}$ du test que l'on souhaite réaliser.

(b) Calculer une p-valeur $\widehat{p}_j$ pour chacun des tests ci-dessus ($j \in \{1, \dots, p\}$). Visualiser le résultat avec un ou deux graphiques.

(c) On décide de rejeter $H_{0,j}$ pour tous les j tels que $\widehat{p}_j \leq \alpha = 0.05$. Combien de gènes a-t-on ainsi sélectionné ?

(d) Supposons qu'en réalité toutes les hypothèses nulles $H_{0,j}$ sont correctes (c'est-à-dire, aucun gène n'a de lien avec la réponse). Dans ce cas, avec la procédure de la question (c), en moyenne, combien de gènes sélectionne-t-on (à tort) ?

(e) **Correction de Bonferroni.** On décide de rejeter $H_{0,j}$ pour tous les j tels que $\widehat{p}_j \leq \alpha/p$. Combien de gènes a-t-on ainsi sélectionné ? Comparer au résultat de la question (c).

(f) **Correction de Benjamini-Yekutieli.** Réordonner les p-valeurs par ordre croissant :

$$\widehat{p}_{\sigma(1)} \leq \widehat{p}_{\sigma(2)} \leq \cdots \leq \widehat{p}_{\sigma(p)} \,.$$

Calculer

$$\widehat{k} := \max\left\{ k \in \{1, \dots, p\} \,/\, \widehat{p}_{\sigma(k)} \leq \frac{\alpha k}{p H_p} \right\}$$

$$\text{où} \qquad H_p := \frac{1}{p} \sum_{k=1}^{p} \frac{1}{k} \,.$$

On décide de rejeter $H_{0,j}$ pour tous les j tels que $\widehat{p}_j \leq \frac{\alpha \widehat{k}}{p H_p}$. Combien de gènes a-t-on ainsi sélectionné ? Comparer aux résultats des questions (c) et (e).