Université d'Angers
Département de Mathématiques M2 Data Science

Apprentissage Statistique en Grande Dimension-TD 2 Corrigé

Exercice 1 (Faire plus de mesures = perdre de l'information ?)**.**

$$Y_j = \theta_j + \varepsilon_j, \quad j = 1, \dots, p,$$

et (ε_j) suite de variables aléatoires *i.i.d* de loi $\mathcal{N}(0, \sigma_j^2)$. Pour simplifier, on suppose que l'on est dans un cadre *homoscédastique*, *i.e.* que $\sigma_j = \sigma$. Le but est ici de détecter les gènes "positifs", *i.e.* ceux pour lesquels

$$\theta_j \neq 0.$$

En général, de 1 à 10% des gènes sont positifs. On suppose que l'on observe n individus : chaque observation est notée

$$Y^{(i)} = (Y_1^{(i)}, \dots, Y_p^{(i)}).$$

1. Lorsque j est fixé, on est face à un test de la moyenne classique :

   $$H_0^j : \theta_j = 0 \quad \text{contre} \quad H_1^j : \theta_j \neq 0.$$

   Assurer que la probabilité que le gène j soit déclaré positif à tort soit de au plus 5% revient à effectuer le test ci-dessus au niveau $\alpha = 5\%$ (on suppose σ connu pour simplifier). Sous H_0, $\bar{Y}_n^j = \frac{1}{n}\sum_{i=1}^n Y_j^{(i)} \sim \mathcal{N}(0, \frac{\sigma^2}{n})$ de sorte que

   $$\mathbb{P}_{H_0^j}\left(|\bar{Y}_n^j| > r\right) = \mathbb{P}\left(|Z| > \frac{r\sigma}{\sqrt{n}}\right).$$

   où $Z \sim \mathcal{N}(0,1)$. Ainsi, $\mathbb{P}_{H_0^j}\left(|\bar{Y}_n^j| > r\right) \leq \alpha$ si et seulement si

   $$r \geq \frac{\sigma q_{\frac{1-\alpha}{2}}}{\sqrt{n}}$$

   où $q_{\frac{1-\alpha}{2}}$ est le quantile d'ordre $\frac{1-\alpha}{2}$ de la loi $\mathcal{N}(0,1)$.
2. Il y a donc 200 gènes positifs. Quitte à changer l'ordre, supposons que les 4800 premiers soient négatifs. Dans ce cas, le nombre de faux positifs sera égal à :

   $$B_1 = \sum_{j=1}^{4800} 1_{\{|\bar{Y}_n^j| > \frac{\sigma q_{\frac{1-\alpha}{2}}}{\sqrt{n}}\}}$$

   avec $\alpha = 0.05$. Ainsi, comme les $\bar{Y}_n^j$ sont indépendantes, B_1 suit une loi binômiale de paramètres 0.05 et $N = 4800$, de sorte que

   $$\mathbb{E}[B_1] = 240.$$

En moyenne, il y a donc 240 faux positifs (vs 200 vrais positifs). Calculons une minoration de la FDR définie par

$$FDR = \mathbb{E}[\frac{FP}{FP+TP}1_{\{FP+TP\geq 1\}}] = \mathbb{E}[\frac{B_1}{B_1+B_2}1_{\{B_1+B_2\geq 1\}}]$$

où B_2 est la somme de variables aléatoires de Bernoulli indépendantes (et indépendantes de B_1) de paramètre

$$p_j = \mathbb{P}_{H_1^j}(\bar{Y}_n^j < \frac{\sigma q_{\frac{1-\alpha}{2}}}{\sqrt{n}}).$$

On ne peut calculer p_j en l'état vu la forme de l'alternative. Comme $B_2 \leq 200$, on peut en revanche minorer la FDR par

$$\mathbb{E}[\frac{B_1}{B_1+200}1_{\{B_1\geq 1\}}] = \mathbb{E}[\frac{B_1}{B_1+200}] \approx 0.54!!$$

3. Notons $W_j = \frac{\sqrt{n}}{\sigma}\bar{Y}_n^j$. Supposons que les q première variables correspondent à des gènes négatifs. Pour $j \in \{1,\dots,q\}$, les W_j sont indépendants et de loi $\mathcal{N}(0,1)$. Pour qu'il n'y ait aucun faux positif, il faut fixer r_q de sorte que

$$\mathbb{P}(|W_1|^2 \leq r_q, \dots, |W_q|^2 \leq r_q) \xrightarrow{q\to+\infty} 1.$$

Or,

$$\mathbb{P}(|W_1|^2 \leq r_q, \dots, |W_q|^2 \leq r_q) = \mathbb{P}(\max_{i=1}^{q} |W_i|^2 \leq r_q),$$

et d'après l'indication,

$$\mathbb{P}(\max_{i=1}^{q} |W_i|^2 > r_q) \xrightarrow{q\to+\infty} 0$$

si

$$r_q \geq 2\log q.$$

Comme on ne connaît pas le nombre de gènes positifs, on rejette donc H_0^j si

$$\frac{\sqrt{n}}{\sigma}|\bar{Y}_n^j| \geq \sqrt{2\log p}.$$

Par construction, cette règle de décision garantit qu'asymptotiquement, il n'y ait aucun faux positif, *i.e.* que la FWER (Family Wise Error Rate) est nulle.

4. Lorsque $p \to +\infty$ et n fixé, $\frac{\log p}{n} \to +\infty$ et ce test ne détecte alors plus aucun gène !! Il faut donc trouver un compromis entre les deux types de tests. On peut alors fixer une FWER petite qu'un certain seuil (c'est le principe du test de Bonferroni). Le test de Benjamini-Hochberg a lui pour objectif de contrôler la FDR. Ce type de condition est moins contraignant.

5. (a) Par indépendance,

$$\mathbb{P}(\max_{j=1,\dots,p} |\varepsilon_j| \leq q) = (1-G(q))^p.$$

(b) Par IPP,

$$\mathbb{P}(Z > z) = \sqrt{\frac{1}{2\pi}}\frac{e^{-\frac{z^2}{2}}}{z} - \sqrt{\frac{1}{2\pi}}\int_z^{+\infty} \frac{e^{-\frac{x^2}{2}}}{x^2}dx$$

d'où

$$\mathbb{P}(|Z| > z) = \sqrt{\frac{2}{\pi}}\frac{e^{-\frac{z^2}{2}}}{z} - \sqrt{\frac{2}{\pi}}\int_z^{+\infty} \frac{e^{-\frac{x^2}{2}}}{x^2}dx.$$

Or,

$$\int_z^{+\infty} \frac{e^{-\frac{x^2}{2}}}{x^2}dx \leq \frac{1}{z^3}\int_z^{+\infty} ue^{-\frac{u^2}{2}}du \leq \frac{e^{-\frac{z^2}{2}}}{z^3}$$

d'où le résultat.

(c) Pour obtenir le résultat sur le max de gaussiennes, posons $z_p = \sqrt{2\gamma \log p}$. Dans ce cas,

$$1 - G(z_p) = 1 - \sqrt{\frac{2}{\pi}}\frac{e^{-\gamma \log p}}{\sqrt{2\gamma \log p}}\left(1 + O(\frac{1}{z_p^2})\right)$$

et

$$(1 - G(z_p))^p = \exp\left(p \log\left(1 - \sqrt{\frac{2}{\pi}}\frac{e^{-\gamma \log p}}{\sqrt{2\gamma \log p}}\left(1 + O(\frac{1}{z_p^2})\right)\right)\right).$$

Un développement limité permet alors d'obtenir le résultat.

=

Exercice 2. 1. (a) $\text{rg}(A) \leq p$ puisque $\text{rg}(A)$ est la dimension de l'espace engendré par les colonnes de A mais comme $\text{rg}(A) = \text{rg}(A^T)$, on a aussi $\text{rg}(A) \leq n$.

(b) Soit $y \in \mathcal{I}m(A^T)$. Alors, il existe $x \in \mathbb{R}^n$, tel que $y = A^T x$. Soit $z \in KerA$, alors

$$\langle z, A^T x\rangle = \langle Az, x\rangle = 0.$$

Ainsi, $\mathcal{I}m(A^T)$ est orthogonal à $KerA$. Déduisons-en que $\mathcal{I}m(A) = \mathcal{I}m(AA^T)$. Comme $\mathcal{I}m(A^T)$ est orthogonal à $KerA$, on en déduit d'abord en notant f l'application linéaire associée à A (de la base canonique de $\mathbb{R}^p$ vers celle de $\mathbb{R}^n$), que f restreinte à $\mathcal{I}m(A^T)$ est une bijection. Ainsi,

$$\text{rg}(AA^T) = dim(\mathcal{I}m(AA^T)) = dim(\mathcal{I}m(A^T)) = \text{rg}(A^T) = \text{rg}(A).$$

Par ailleurs,

$$\mathcal{I}m(AA^T) \subset \mathcal{I}m(A)$$

de sorte que $\mathcal{I}m(AA^T) = \mathcal{I}m(A)$ (car leurs dimensions sont égales).

(c) AA^T est une matrice symétrique positive $n \times n$. Elle induit donc une base orthonormale de vecteurs propres $(u_1, \ldots, u_n)$. Quitte à changer l'ordre, on peut supposer que

$$\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_n.$$

Ainsi, en notant $\bar{U}$ la matrice de colonnes $u_1, \ldots, u_n$ (ou plus exactement des coordonnées de ses vecteurs dans la base canonique), on a par la formule de changement de base,

$$AA^T = \bar{U} D \bar{U}^{-1} = \bar{U} D \bar{U}^T$$

car $\bar{U}$ est orthogonale ($D = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$). Finalement, on vérifie bien que

$$\bar{U} D \bar{U}^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T = \sum_{i=1}^{r} \lambda_i u_i u_i^T,$$

car $\lambda_i = 0$ pour $i > r$. Dans la suite, on notera $U = (u_1, \ldots, u_r)$ (matrice $n \times r$).

(d) Si $v_j = \lambda_j^{-\frac{1}{2}} A^T u_j$, $j = 1, \ldots, r$, alors

$$\|v_j\|^2 = \langle \lambda_j^{-\frac{1}{2}} A^T u_j, \lambda_j^{-\frac{1}{2}} A^T u_j \rangle = \lambda_j^{-1} \langle u_j, AA^T u_j \rangle = \lambda_j^{-1} \lambda_j \|u_j\|^2 = 1.$$

(e) On a :

$$A^T A v_j = \lambda_j^{-\frac{1}{2}} A^T (AA^T u_j) = \lambda_j \left(\lambda_j^{-\frac{1}{2}} A^T u_j \right) = \lambda_j v_j.$$

Les v_j sont des vecteurs propres normés associés aux λ_j pour la matrice AA^T. Vérifions qu'ils sont orthogonaux.

$$\langle v_i, v_j \rangle = (\lambda_i \lambda_j)^{-\frac{1}{2}} \langle A^T u_i, A^T u_j \rangle = (\lambda_i \lambda_j)^{-\frac{1}{2}} \langle u_i, AA^T u_j \rangle = \left(\frac{\lambda_j}{\lambda_i} \right)^{\frac{1}{2}} \langle u_i, u_j \rangle = 0$$

car les u_i sont orthogonaux. Il s'agit donc bien d'une base orthonormée de $\mathcal{I}m(A^T A)$.

(f) Posons $\sigma_j = \sqrt{\lambda_j}$. Alors,

$$\sum_{j=1}^{r} \sigma_j u_j v_j^T = \sum_{j=1}^{r} \sigma_j \sigma_j^{-1} u_j u_j^T A = UU^T A.$$

(g) $(u_1, \ldots, u_r)$ en tant que famille libre de vecteurs contenus dans $\mathcal{I}m(AA^T)$ de dimension r forme clairement une base de ce sous-espace vectoriel. Pour tout $x \in \mathbb{R}^p$, $U^T x = (\langle u_1, x \rangle, \ldots, \langle u_r, x \rangle)^T$, de sorte que

$$UU^T x = \sum_{i=1}^{r} \langle u_i, x \rangle u_i.$$

Il s'agit bien d'une matrice de projection sur $\mathcal{I}m(AA^T)$.

(h) Ainsi, pour tout $y \in \mathcal{I}m(A) = \mathcal{I}m(AA^T)$, l'image de y par la projection sur $\mathcal{I}m(AA^T)$ est donc y, ce qui signifie que $UU^T y = y$ d'après la question précédente. De même pour tout $x \in \mathbb{R}^p$, $y = Ax \in \mathcal{I}m(A)$ de sorte que $UU^T Ax = Ax$. Ainsi, $UU^T A = A$.

(i) Il s'agit d'une conséquence directe des questions (f) et (h).