Statistique en Grande Dimension et Apprentissage - TD 2

Ce TD a pour but d'illustrer le chapitre 2 sur les problèmes de la grande dimension puis de faire un rappel sur l'analyse en composantes principales sur laquelle on reviendra dans la suite du cours (dans sa version sparse).

Exercice 1 (Faire plus de mesures = perdre de l'information ?)**.** On s'intéresse dans ce problème à l'analyse des puces à ADN. Après un traitement approprié, les données de puces à ADN correspondent à un vecteur $(Y_1, \dots, Y_p)$ de différences de log-intensités. Typiquement, le nombre p de gènes est de l'ordre de quelques milliers. Ces observations peuvent être modélisées comme suit :

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

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. On s'intéresse à un gène j fixé. Proposez une règle de décision telle que la probabilité que le gène j soit déclaré positif à tort (faux positif/false discoverie) soit de au plus 5%. (On veut limiter les fausses découvertes pour ne faire des expérimentations biologiques (généralement coûteuses) que sur les gènes pour lesquels on "est sûr" qu'ils sont positifs).
2. Supposons que $p = 5000$ et que 4% des gènes sont réellement positifs. Quel est le nombre moyen de faux positifs que va engendrer la règle de décision ci-dessus ?
3. Sachant que
$$\mathbb{P}(\max_{j=1,\dots,p} \varepsilon_j^2 > 2\gamma \log p) \xrightarrow{p\to+\infty} 1_{\gamma<1}, \tag{1}$$
proposez une règle de décision telle que la probabilité qu'il existe au moins un j déclaré positif à tort, soit asymptotiquement nulle.
4. Que se passe-t-il quand p devient très grand ? Comment pourriez-vous envisager de pallier ce problème ?
5. Dans cette dernière questions, on se propose de prouver le résultat (1).
   (a) Pour un seuil q fixé, exprimez
   $$\mathbb{P}(\max_{j=1,\dots,p} |\varepsilon_j| \leq q)$$
   en fonction de la fonction $G(z) = \mathbb{P}(|Z| > z)$.

(b) Montrez que pour tout $z > 1$,

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

Indication : On pourra faire une IPP (et considérer en premier lieu $\mathbb{P}(Z > z)$).

(c) En déduire le résultat.

Exercice 2 (Analyse en composantes principales)**.** L'analyse en composantes principales (ACP) est un moyen usuel de réduire la dimension en tentant d'extraire les directions qui contiennent le plus d'information. Dans cet exercice, on se propose de faire quelques rappels sur l'ACP, théoriques et pratiques et d'étudier numériquement la robustesse (ou plutôt la non-robustesse) de l'ACP à la dimension.

1. On souhaite ici redémontrer la décomposition en valeurs singulières (SVD). On considère A une matrice $n \times p$ de rang r.
   (a) Montrez que $r \leq \min(n, p)$.
   (b) Montrez que $\mathcal{I}m(A^T)$ et Ker$A$ sont orthogonaux puis que $\mathcal{I}m(A) = \mathcal{I}m(AA^T)$.
   (c) Montrez que AA^T admet une décomposition de la forme suivante :

$$AA^T = \sum_{j=1}^{r} \lambda_j u_j u_j^T$$

   où $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ et $\{u_1, \ldots, u_r\}$ est une famille orthonormale.
   (d) Posons $v_j = \lambda_j^{-\frac{1}{2}} A^T u_j$, $j = 1, \ldots, r$. Montrez que $\|v_j\|^2 = 1$ pour tout $j \in \{1, \ldots, r\}$.
   (e) Montrez également que pour tout $j \in \{1, \ldots, r\}$, v_j est vecteur propre de $A^T A$ associé à λ_j. En déduire que $\{v_1, \ldots, v_r\}$ forme une famille orthonormale de vecteurs propres de $A^T A$.
   (f) En posant $\sigma_j = \sqrt{\lambda_j}$, montrez que

$$\sum_{j=1}^{r} \sigma_j u_j v_j^T = \left(\sum_{j=1}^{r} u_j u_j^T\right) A = UU^T A$$

   où $U = (u_1, \ldots, u_r)$ (matrice ayant comme colonnes les vecteurs $u_1, \ldots u_r$).
   (g) Vérifiez que $u_1, \ldots, u_r$ forment une base de $\mathcal{I}m(AA^T)$ puis déduisez-en que UU^T est la matrice de projection sur $\mathcal{I}m(AA^T)$.
   (h) En utilisant la question 1b, en déduire que pour tout $y \in \mathcal{I}m(A)$, $UU^T y = y$ puis que $UU^T A = A$.
   (i) En déduire la SVD :

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T = UDV$$

   où $D = \text{Diag}(\sigma_1, \ldots, \sigma_r, 0 \ldots, 0)$ et $V = (v_1, \ldots, v_r)$. Les vecteurs $v_1, \ldots, v_r$ sont appelés les axes principaux.

Pour rappel, on a le théorème suivant (admis, voir *e.g.* [Giraud, p.23]) :

Théorème 0.1. Notons $\mathbf{X} = (X^{(1)}, \ldots, X^{(n)})^T$ (matrice $n \times p$). Avec les notations précédentes, $\mathbf{X}$ admet une SVD de la forme $\mathbf{X} = UDV$. Pour tout $d \leq r := \mathrm{rg}(\mathbf{X}\mathbf{X}^T)$, $V_d := \mathrm{Vect}(v_1, \ldots, v_d)$ appartient à

$$\mathrm{Argmin}_{\dim(\tilde{V}) \leq d} \sum_{i=1}^{n} \|X^{(i)} - \mathrm{Proj}_{\tilde{V}} X^{(i)}\|^2.$$

où $\mathrm{Proj}_{\tilde{V}}$ désigne la projection orthogonale sur le sous-espace vectoriel V. De plus,

$$\sum_{i=1}^{n} \|X^{(i)} - \mathrm{Proj}_{V_d} X^{(i)}\|^2 = \sum_{k=d+1}^{r} \sigma_k^2.$$

Quelques rappels : Les vecteurs $v_1, \ldots, v_r$ sont appelés les axes principaux tandis que les vecteurs $c_1 := \sigma_1 u_1, \ldots, c_r := \sigma_r u_r$ sont les composantes principales. L'intérêt de la SVD est de pouvoir lire directement la projection sur les matrices U, D et V. Plus exactement, pour tout $i \in \{1, \ldots, n\}$, on a $X^{(i)} = c_1(i)v_1 + \ldots, c_r(i)v_r$. De même,

$$\mathrm{Proj}_{V_d}(X^{(i)}) = c_1(i)v_1 + \ldots, c_d(i)v_d.$$

En pratique, l'analyse en composantes principales se fait sur la matrice des observations recentrées :

$$\tilde{X}^{(i)} = X^{(i)} - \frac{1}{n} \sum_{i=1}^{n} X^{(i)}.$$