MAP 553 Statistique

PC1: Fléau de la dimension - ACP

1 Le fléau de la dimension

Certaines stratégies d'estimation qui sont naturelles en "basse dimension" sont complètement mises en échec lorsque la dimension devient "grande". Voici deux petits exemples simples.

a- Estimation de densité par histogrammes réguliers.
Supposons qu'on observe $X_1, \dots, X_n \in [0,1]^p$ i.i.d. tirées selon une loi inconnue ayant une densité $f : [0,1]^p \to \mathbb{R}$ par rapport à la mesure de Lebesgue. On cherche à avoir une idée de la forme de f. Une idée naturelle est de faire un histogramme avec disons des "cases" de 0.1 de côté. Voici un exemple en dimension $p = 1$.

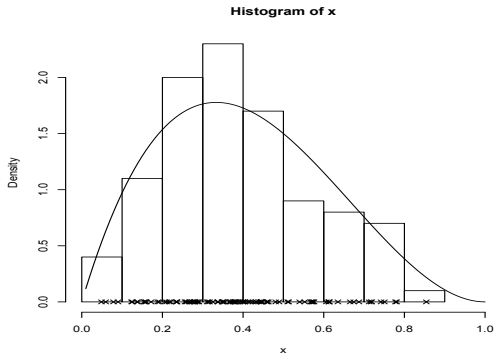

FIGURE 1 – Histogramme pour un échantillon de taille 100 en dimension $p = 1$. Les croix représentent les X_i, en ligne pleine, la vraie densité f.

Pour avoir en moyenne 10 observations par cases, quelle taille doit avoir n ? Conclusion ? Comment faire avec des échantillons plus petits ?

b- Echec des méthodes locales en régression.
Dans le modèle de régression (univarié) on observe des variables d'intérêt $Y_1, \dots, Y_n \in \mathbb{R}$ et des variables explicatives $X_1, \dots, X_n \in \mathbb{R}^p$ et on suppose que les Y_i et les X_i sont reliés par l'équation

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

où f est inconnue et les ε_i sont i.i.d. (bruit d'observation). On cherche à déterminer l'allure de la fonction f. Comme précédemment avec les histogrammes, une idée naturelle est d'estimer la valeur $f(x)$ en prenant la moyenne des valeurs des Y_i associées aux X_i dans la boule $\mathcal{B}(x, r)$ avec un r petit. Pour simplifier on va supposer que les $\varepsilon_i = 0$ (observations non bruitées). On va aussi supposer que les X_i sont i.i.d et tirés selon la loi uniforme sur la boule unité.

1. Pour $r < 1$ montrer que la probabilité qu'au moins un des X_i se trouve dans la boule $\mathcal{B}(0, r)$ vaut $1 - (1 - r^p)^n$.
2. Que doit valoir r pour que cette probabilité soit au moins de $1/2$?
3. Pour estimer $f(0)$ avec au moins un point, quel est l'ordre de grandeur du diamètre r minimal ? Conclusion ?

2 Faire plus de mesures = perdre de l'information ?

Illustrons ce problème avec l'analyse des puces à ADN. Après un traitement approprié, les données de puces à ADN correspondent à un grand 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 : $Y_j = \theta_j + \varepsilon_j$, pour $j = 1, \dots, p$ avec les ε_j i.i.d. de loi $\mathcal{N}(0, \sigma^2)$ avec σ^2 connu, disons $\sigma^2 = 1$. Les sites qui nous intéressent d'un point de vue biologique sont les gènes j où θ_j est non nul (on dit alors que le gène j est positif). En général, de 1 à 10% des gènes sont positifs.

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 (fausse découverte) soit de au plus 5%. (on veut limiter les fausses découvertes car les sites déclarés positifs donnent lieu à de nouvelles expériences biologiques qui coûtent cher).
2. Supposons que $p = 5000$ et 4% des gènes sont positifs. Quel est le nombre moyen de faux positifs que va engendrer la règle de décision ci-dessus ?
3. Sachant $\mathbb{P}\left(\max_{j=1,\dots,p} \varepsilon_j^2 > \gamma\, 2 \log p\right) \overset{p\to\infty}{\to} 1_{\gamma \geq 1}$ 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 lorsque p devient très grand ? Comment gérer ce problème ?

Pour aller plus loin :
Jiashun Jin. *Impossibility of successful classification when useful features are rare and weak.* Proceedings of the National Academy of Sciences of the USA. 106 (22) ; 2009. pp.8859-64.
`http ://www.pnas.org/content/106/22/8859.full`

3 Réduction de dimension et ACP

On dispose d'un tableau de données $X \in \mathbb{R}^{n \times p}$ avec possiblement $p > n$. On supposera que les colonnes de la matrice $X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$ sont centrées. On cherche à projeter les $X_i \in \mathbb{R}^p$ sur un espace vectoriel V de petite dimension en perdant un minimum d'information. On va rechercher des espaces V_d réalisant

$$V_d = \arg \min_{\dim(V)=d} \sum_{i=1}^{n} \|X_i - \mathrm{Proj}_V(X_i)\|^2.$$

1. Montrer que V_d vérifie
$$V_d = \arg \max_{\dim(V)=d} \sum_{i=1}^{n} \|\mathrm{Proj}_V(X_i)\|^2.$$
2. Lorsque $d = 1$ montrez que $V_1 = \mathrm{vect}\{a^{(1)}\}$ où $a^{(1)}$ est un vecteur propre associé à la plus grande valeur propre de $X^T X$.
3. Pour $d > 1$, par quels vecteurs V_d est-il engendré ?
4. Exprimez $\mathrm{Proj}_{V_d}(X_i)$ à l'aide des composantes principales de l'ACP.