U.P.S. : M2R *Mathématiques Appliquées* — INSA : 5$^{\text{ième}}$ année *GMM*

**Module B5 : Modélisation Statistique**

Épreuve du 25 janvier 2007

(durée : 3 heures — Les documents ne sont pas autorisés)

## Exercice I

### Test du rapport de vraisemblance

Soit un modèle linéaire gaussien

$$Y = X\theta + \varepsilon \tag{1}$$

avec les conditions du cours. On suppose de plus ici que $\sigma^2$ est connu et, sans perte de généralité, on supposera que $\sigma^2 = 1$

1. Monter que $\widehat{\theta} = (X^T X)^{-1} X^T Y$ est l'estimateur du maximum de vraisemblance.
2. Calculer la log vraisemblance maximale du modèle
3. Calculer la log vraisemblance sous l'hypothèse $H_0 : \theta = 0$
4. Montrer que le test du rapport de vraisemblance de $H_0$ contre $H_1 : \theta \neq 0$ est une fonction monotone de $\|X\widehat{\theta}\|^2$ dont on calculera la loi.
5. Donner la région de rejet de ce test au niveau $\alpha$

On considère maintenant le modèle

$$\widetilde{Y} = \widetilde{X}\theta + \widetilde{\varepsilon} \tag{2}$$

ou $\widetilde{Y}$ est un vecteur de taille $n$ , $\widetilde{X}$ un matrice connue de plein rang et de dimension $n, p$ $p < n$ et $\widetilde{\varepsilon} \sim N(0, \Sigma)$ avec $\Sigma$ connue.

6. Montrer comment on peut se ramener au cas précédent par un changement de variable.
7. Déduisez en, précisez bien les arguments, le test du rapport de vraisemblance de $H_0 : \theta = 0$ dans le modèle (2).
8. On suppose maintenant que $p = 1$ de sorte que $\widetilde{X}$ est un vecteur. Comparer la puissance $(H_1 : \theta \neq 0)$ du test du rapport de vraisemblance avec celui basé sur la statistique

$$U = \widetilde{X}^T \widetilde{Y}$$

## Exercice II

### Estimateur par seuillage dur

On considère le modèle Gaussien séquentiel

$$y_k = \theta_k^* + \epsilon_k, \text{ avec } \epsilon_k \sim_{i.i.d.} N(0,1),\ k = 1, \ldots, n.$$

Soit $\lambda \in \mathbb{R}^+$, on rappelle que l'estimateur $\hat{\theta}^H$ par seuillage dur est défini par

$$\hat{\theta}_k^H = \begin{cases} y_k & \text{si } |y_k| > \lambda \\ 0 & \text{si } |y_k| \leq \lambda \end{cases}$$

1. Montrer que $\lim_{n\to+\infty} \mathbb{P}\left(\max_{1\leq i\leq n} |\epsilon_i| \geq \sqrt{2\log(n)}\right) = 0$.

   Indication : on pourra utiliser le fait que pour tout $\lambda \in \mathbb{R}^+$, on a que $\int_\lambda^{+\infty} \phi(x)dx \leq \frac{\phi(\lambda)}{\lambda}$ où $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
2. Soit $z \sim N(0,1)$. Montrer que pour tout $\lambda \in \mathbb{R}^+$

   $$\mathbb{E}\left(z^2 \mathbb{1}_{\{|z|>\lambda\}}\right) \leq 2(\lambda + \frac{1}{\lambda})\phi(\lambda)$$

   Indication : utiliser une intégration par partie et le fait que $\phi'(x) = -x\phi(x)$.
3. On suppose que $\theta_k^* = 0$ pour tout $k = 1, \ldots, n$. Montrer directement (sans utiliser le théorème sur l'inégalité oracle vu en cours) que pour tout $n \geq 2$ et pour $\lambda = \sqrt{2\log(n)}$

   $$\mathbb{E}\|\hat{\theta}^H - \theta^*\|^2 \leq \frac{1}{\sqrt{\pi\log(2)}}(2\log(n) + 1)$$

4. Soit $0 < \delta < 1$ et $\lambda_{n,\delta} = (1-\delta)\sqrt{2\log(n)}$. On pose $I_n = \sum_{k=1}^n \mathbb{1}_{\{|\epsilon_i|>\lambda_{n,\delta}\}}$. Montrer que

   $$\lim_{n\to+\infty} \mathbb{E}I_n = +\infty$$

   Indication : on pourra utiliser le fait que pour tout $\lambda \in \mathbb{R}^+$, on a que

   $$\int_\lambda^{+\infty} \phi(x)dx \geq \frac{\phi(\lambda)}{\lambda}\left(1 - \frac{1}{\lambda^2}\right)$$

5. Montrer que

   $$\hat{\theta}^H = \arg\min_{\theta\in\mathbb{R}^n} \|Y - \theta\|^2 + \lambda^2 \sum_{k=1}^n \mathbb{1}_{\{\theta_k \neq 0\}}$$

   où $Y = (y_1, \ldots, y_n)^t$.

## Exercice III

### Estimation par lissage Spline

On considère le modèle :

$$Y = f_n + \epsilon, \text{ avec } f_n = K_n\theta_n^* \text{ et } \epsilon \sim N(0, \sigma^2 I_n),$$

où $I_n$ est la matrice identité, $\sigma$ est un niveau de bruit inconnu, $K_n$ une matrice connue symmétrique définie positive de taille $n \times n$ et $\theta_n^*$ un vecteur de $\mathbb{R}^n$ inconnu que l'on cherche à estimer. On suppose de plus qu'il existe une constante $C^* > 0$ telle que pour tout $n \geq 1$ :

$$(\theta_n^*)^t K_n \theta_n^* \leq C^*,$$

et que les valeurs propres de $K_n$ notées $\mu_{1,n} \geq \ldots \geq \mu_{n,n}$ vérifient

$$\alpha \frac{k^{2m}}{n} \leq \frac{1}{\mu_{k,n}}, \ k = 1, \ldots, n \tag{3}$$

où $\alpha \in \mathbb{R}$ et $m \in \mathbb{N}^*$ sont des constantes inconnues.

Pour $\lambda \in \mathbb{R}^+$, on considère l'estimateur de type Spline défini par

$$\hat{\theta}_\lambda = \arg\min_{\theta\in\mathbb{R}^n} \frac{1}{n}\|Y - K_n\theta\|^2 + \lambda\theta^t K_n \theta,$$

et on note $\hat{f}_\lambda = K_n\hat{\theta}_\lambda$. Le but de cet exercice est d'étudier la convergence du risque

$$R_n(\lambda) = \frac{1}{n}\|\hat{f}_\lambda - f_n\|^2$$

1. Montrer que $\hat{f}_\lambda = A_\lambda Y$ où $A_\lambda = K_n(K_n + n\lambda I_n)^{-1}$
2. En déduire que $\mathbb{E}R_n(\lambda) = \frac{1}{n}\|(I_n - A_\lambda)f_n\|^2 + \frac{\sigma^2}{n}Tr(A_\lambda^2)$
3. Montrer que $\hat{R}(\lambda) = \frac{1}{n}\|(I_n - A_\lambda)Y\|^2 - \frac{\sigma^2}{n}Tr((I_n - A_\lambda)^2) + \frac{\sigma^2}{n}Tr(A_\lambda^2)$ est un estimateur sans biais de $\mathbb{E}R_n(\lambda)$.
4. Montrer que pour tout $\theta \in \mathbb{R}^n$

$$\frac{1}{n}\|f_n - A_\lambda f_n\|^2 \leq \frac{1}{n}\|f_n - K_n\theta\|^2 + \lambda\theta^t K_n \theta$$

   En déduire que

$$\frac{1}{n}\|(I_n - A_\lambda)f_n\|^2 \leq C^*\lambda$$

5. Montrer que $\lambda^{1/2m}Tr(A_\lambda^2) \leq \frac{1}{\alpha^{1/2m}}\int_0^{+\infty}\frac{1}{(1+x^{2m})^2}dx$
   Indication : utiliser l'hypothèse (3) pour majorer $\frac{1}{1+\frac{n\lambda}{\mu_{k,n}}}$
6. Soit $(\lambda_n)_{n\geq 1}$ une suite de paramètres. Donner des conditions suffisantes sur $\lambda_n$ pour garantir que

$$\lim_{n\to+\infty}\mathbb{E}R_n(\lambda_n) = 0$$

7. Proposer un choix pour $\lambda_n$ de sorte qu'il existe une constante $C > 0$ telle que pour tout $n \geq 1$

$$\mathbb{E}R_n(\lambda_n) \leq Cn^{-\frac{2m}{2m+1}}$$