Méthodologie Statistique

Épreuve Théorique

Examen sans document de 1h15 (INSA) et 2h30 (DEA)

Soit $X_1, \cdots, X_p$, p vecteurs de $\mathbb{R}^n$ connus (avec $p < n$) tels que $\text{rg}(X_1, \cdots, X_p) = p$ et soit $m_p = \{1, 2, \cdots, p\}$. On suppose que l'on a obtenu les résultats du modèle de régression suivant, dit “vrai modèle” :

$$Y = X^*.\beta^* + \varepsilon^*,$$

avec $Y = (y_1, \cdots, y_n)'$ un vecteur connu, $\varepsilon^* = (\varepsilon_1^*, \cdots, \varepsilon_n^*)$, où les (ε_i^*) sont des v.a.i.i.d. centrées gaussiennes et de variance σ_*^2 inconnue, $\beta^* = (\beta_1^*, \cdots, \beta_{|m^*|}^*)'$ est un vecteur de $(\mathbb{R} \setminus \{0\})^{|m^*|}$ et X^* est la matrice $(X_{m_1^*}, \cdots, X_{m_{|m^*|}^*})$ (les différents vecteurs X_i de $\mathbb{R}^n$ sont mis l'un à côté de l'autre) où $m^* = (m_1^*, \cdots, m_{|m^*|}^*) \subset m_p$ est un ensemble inconnu (on note pour tout ensemble fini m, $|m| = \text{Card}(m)$). On désire estimer m^* par une procédure de choix de modèle.

Notations : • Pour $(a_n)_n$ et $(b_n)_n$ deux suites réelles, on note $a_n = \mathcal{O}(b_n)$ lorsque $\exists\, C > 0$ tel que pour tout $n \in \mathbb{N}$, $|a_n| \leq C|b_n|$.
• Pour A une matrice carrée d'ordre k, symétrique, on rappelle que A est diagonalisable dans $\mathbb{R}$ et on note $\lambda_k(A) \leq \lambda_{k-1}(A) \leq \cdots \leq \lambda_1(A)$ ses valeurs propres.
• Pour $Z \in \mathbb{R}^k$, avec $k \in \mathbb{N}^*$, on note $||Z||^2 = Z'.Z$.

I. Soit $m = (m_1, \cdots, m_{|m|})$ un sous-ensemble de m_p. On note $X^{(m)}$ la matrice $(X_{m_1}, \cdots, X_{m_{|m|}})$. On désire identifier les paramètres inconnus du vrai modèle et pour cela on utilise le modèle de régression :

$$Y = X^{(m)}.\beta^{(m)} + \varepsilon,$$

avec $\beta^{(m)}$ un vecteur inconnu de $\mathbb{R}^{|m|}$ et ε un vecteur gaussien centré de matrice de covariance $\sigma^2.I_n$ avec σ^2 inconnu et I_n la matrice identité dans $\mathbb{R}^n$.

1. Déterminer l'expression de l'estimateur $\widehat{\beta}^{(m)}$ de β par moindre carrés ordinaires, c'est-à-dire en minimisant $||Y - X^{(m)}.\beta||^2$ pour $\beta \in \mathbb{R}^{|m|}$ (redémontrer ce résultat).

2. Par la suite, on notera $\widehat{Y}^{(m)} = X^{(m)}.\widehat{\beta}^{(m)}$, $S^{(m)}(n) = ||Y - \widehat{Y}^{(m)}||^2$ et $\widehat{\sigma}^2_{(m)}(n) = \dfrac{S^{(m)}(n)}{n - |m|}$.
Montrer que si $m^* \subset m$, alors $\widehat{\sigma}^2_{(m)}(n) \xrightarrow[n\to+\infty]{\mathcal{L}} \sigma_*^2$, puis que $\widehat{\sigma}^2_{(m)}(n) \xrightarrow[n\to+\infty]{P} \sigma_*^2$. Déterminer $\text{var}\left(\widehat{\sigma}^2_{(m)}(n)\right)$. Pour quel ensemble m cette variance est-elle minimale (avec $m^* \subset m$) ?

3. On considère maintenant le critère de sélection de modèle pour $m \subset m_p$:

$$D_n(m) = S^{(m)}(n) + |m|.c_n.\widehat{\sigma}^2_{(m_p)}(n),$$

avec $(c_n)_{n\in\mathbb{N}}$ une suite de réels telle que $\lim_{n\to\infty} n^{-1}c_n = 0$ et $\lim_{n\to\infty} c_n = +\infty$, et on sélectionne le modèle $\widehat{m}_n$ défini par :

$$\widehat{m}_n = \text{Argmin}_{m\subset m_p} \left\{D_n(m)\right\}.$$

Rappeler la définition du Cp de Mallows en utilisant les notations présentes. A-t-on asymptotiquement tendance à choisir un modèle plus grand (au sens de l'inclusion) avec le critère D_n ou avec le Cp de Mallows ?

II. On veut montrer la convergence de $\widehat{m}_n$ pour la famille $\mathcal{M}$ de modèles hiérarchique et sous certaines conditions. On suppose donc ici que $m^* = \{1, \cdots, |m^*|\}$ et que $m = \{1, \cdots, k\} \in \mathcal{M}$.

1. Montrer que pour A matrice symétrique d'ordre k alors $\lambda_1(A) = \max_{Z \in \mathbb{R}^k,\ ||Z||^2=1} \{Z'.A.Z\}$ et $\lambda_k(A) = \min_{Z \in \mathbb{R}^k,\ ||Z||^2=1} \{Z'.A.Z\}$.

2. Soit la matrice $G = (X^{(m_p)})'.X^{(m_p)}$. On se place désormais sous l'hypothèse suivante :

   Hypothèse : On suppose qu'il existe deux constantes réelles positives η_1 et η_2 telles que $0 < \eta_1.n \leq \lambda_p(G) \leq \cdots \leq \lambda_1(G) \leq \eta_2.n$

   Montrer qu'alors :

   $$\begin{aligned} 0 < \eta_1.n \leq (X_i)'.X_i \leq \eta_2.n \quad & \text{pour tout } i = 1, \cdots, p \qquad (1) \\ 0 < \eta_1.n \leq (X_i)'.Q_i.X_i \leq \eta_2.n \quad & \text{pour tout } i = 2, \cdots, p, \qquad (2) \end{aligned}$$

   avec Q_i la matrice du projecteur orthogonal (pour le produit scalaire usuel de $\mathbb{R}^n$) sur le sous-espace vectoriel orthogonal à $\text{Vect}(X^1, X^2, \cdots, X^{i-1})$.

3. On suppose que $(a_n)_n$ est une suite réelle telle que $b_n = \left(\sum_{i=1}^n a_i^2\right)^{1/2} \underset{n\to+\infty}{\longrightarrow} \infty$. On pose T_n la variable aléatoire $T_n = \sum_{i=1}^n a_i.\varepsilon_i^*$. Montrer que $T_n = \mathcal{O}(b_n)$ presque sûrement (on pourra montrer que $b_n^{-1}T_n$ converge presque sûrement à l'aide d'un argument de martingale). En déduire que $(X_i)'.\varepsilon^* = \mathcal{O}(n^{1/2})$ presque sûrement pour tout $i = 1, \cdots, p$.

4. Montrer que $(\varepsilon^*)'.P_{(m)}.\varepsilon^* = \mathcal{O}(1)$ presque sûrement, où $P_{(m)} = X^{(m)}.\left[(X^{(m)})'.X^{(m)}\right]^{-1}(X^{(m)})'$.

5. Montrer que $\widehat{\sigma}^2_{(m_p)}(n) \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} \sigma^2_*$.

6. On suppose que $k < |m^*|$. Montrer que

   $$\begin{aligned} D_n(m) - D_n(m^*) &= (X^*.\beta^*)'.Q_{m+1}.(X^*.\beta^*) + 2(\varepsilon^*)'.Q_{m+1}.(X^*.\beta^*) \\ &\quad -(\varepsilon^*)'.P_{(m)}.\varepsilon^* - (|m^*| - k)c_n.\widehat{\sigma}^2_{(m_p)}(n). \end{aligned}$$

   En déduire que presque sûrement $D_n(m) - D_n(m^*) > 0$ et donc que $\widehat{m}_n \geq m^*$.

7. On suppose maintenant que $k > |m^*|$. Montrer que :

   $$D_n(m) - D_n(m^*) = (k - |m^*|)c_n.\widehat{\sigma}^2_{(m_p)}(n) - (\varepsilon^*)'.(P_{(m)} - P_{(m^*)})\varepsilon^*.$$

   En utilisant les résultats précédents, en déduire que $D_n(m) - D_n(m^*) > 0$ presque sûrement et donc que $\widehat{m}_n \leq m^*$, puis que $\widehat{m}_n = m^*$ p.s. pour n assez grand.