M2R Mathématiques appliquées
2010–2011
Modèlisation Statistique

Examen du 6 Janvier 2012 de 8h à 12h

Les notes de cours sont autorisées. Les calculatrices sont autorisées.

1 Choix de Modèles

Soit $N > 1$ et $x_1^N, x_2^N, \dots x_N^N$ avec $\sum_{j=1}^N x_j^N = 0$, $\sum_{j=1}^N \left(x_j^N\right)^2 = N/3$. Pour, $i = 1, 2, 3$ et $j = 1 \dots N$, on considère le modèle linéaire gaussien de dimension $3N$

$$Y_{i,j} = \delta_i^* + \eta^* x_j^N + \varepsilon_{i,j}$$

où $(\varepsilon_{i,j})_{i=1,2,3,j=1\dots N} \sim \mathcal{N}_{3N}(0, \sigma_*^2 I_{3N})$ avec $\sigma_*^2 > 0$. On pose

$$\theta^* := \begin{pmatrix} \delta_1^* \\ \delta_2^* \\ \delta_3^* \\ \eta^* \end{pmatrix} = \begin{pmatrix} \theta_1^* \\ \theta_2^* \\ \theta_3^* \\ \theta_4^* \end{pmatrix}.$$

Dans tout l'exercice on suppose que l'on a $\theta_1^* = \theta_3^* = 0$.

1. Montrer que le modèle précédent rentre dans le cadre du modèle linéaire gaussien régulier. Calculer les estimateurs du maximum de vraisemblance $\widehat{\theta_N}$ et $\widehat{\sigma_N^2}$ de θ^* et σ_*^2.
2. Montrer que $\widehat{\theta_N}$ converge presque sûrement, quand N tend vers l'infini, vers θ^*.
3. Quelle est la loi jointe de $\left(\sqrt{N}(\widehat{\theta_N} - \theta^*), 3N\frac{\widehat{\sigma_N^2}}{\sigma_*^2}\right)$? Décrire les régions de confiance pour θ^* et σ_*^2.
4. Pour $J \subset \{1, 2, 3, 4\}$, on pose $\theta_J^* := (\theta_j^*)_{j \in J}$. On se propose de comparer les stratégies de choix de modèles du C_p de Mallows avec BIC. Rappelons tout d'abord les fonctions d'objectifs associées. Pour $J \subset \{1, 2, 3, 4\}$, $|J|$ désigne le nombre d'éléments de J et $\widehat{\sigma_{N,J}^2}$ la variance estimée par maximum de vraisemblance dans le modèle à $|J|$ paramètres où le paramètre intervenant dans la moyenne des observations est θ_J. On a alors

$$C_p(J) := \frac{\widehat{\sigma_{N,J}^2}}{\widehat{\sigma_N^2}} + \frac{|J|}{3N},$$

$$BIC(J) := \log(\widehat{\sigma_{N,J}^2}) + \frac{\log 3 + \log N}{3N}|J|.$$

Rappelons que l'on se place dans le cas où $J^* = \{2, 4\}$. Les questions qui suivent seront traitées pour le BIC et C_p de Mallows

– Pour chacun des 3 sous-ensembles J de $\{1, 2, 3, 4\}$ qui sur-ajustent J^* calculer les probabilités de préférer le modèle J au modèle J^*. Calculer les limites, quand N tend vers l'infini, de ces probabilités.
– Que pensez-vous des critères BIC et C_p de Mallows pour le modèle étudié ici ?

2 Loi des Extrêmes-POT

1. Soit $X_1, \ldots, X_n$ des variables i.i.d. de loi exponentielle de moyenne 1
   (a) Montrer que $X_{(n)} - \log(n)$ converge en loi vers la loi de Gumbel.
   (b) Montrer que la loi exponentielle n'a pas de mémoire :
   $$\mathbb{P}(X_1 > t+s | X_1 > t) = P(X_1 > s) \; (t, s > 0).$$
   (c) Soit $t, \varepsilon > 0$, Quelle est la limite en probabilité quand ε tend vers 0 de εX_1 conditionnellement à l'événement $\{\varepsilon X_1 \geq t\}$?
   (d) Quelle est la loi de $\varepsilon^{-1}(\varepsilon X_1 - t)$ conditionnellement à l'événement $\{\varepsilon X_1 \geq t\}$?
2. Pour $a > 1$, soit $Y_1, \ldots, Y_n$ des variables i.i.d. de loi de Pareto de densité (par rapport à la mesure de Lebesgue)
   $$f_a(x) := \frac{a-1}{x^a} \mathbb{I}_{x \geq 1}.$$
   (a) Montrer que $n^{-\frac{1}{a-1}} Y_{(n)}$ converge en loi vers une loi de densité g_a à préciser.
   (b) Calculer pour $t > 1$ et $s > 0$,
   $$\mathbb{P}(Y_1 > t+s | Y_1 > t)$$
   (c) Soit $t, \varepsilon > 0$, Quelle est la limite en probabilité quand ε tend vers 0 de εY_1 conditionnellement à l'événement $\{\varepsilon Y_1 \geq t\}$?
   (d) Comparer les résultats obtenus pour la loi de Pareto à ceux obtenus pour la loi exponentielle.

3 Processus AR(1)

Soit $-1 < a < 1$ et soit $(\varepsilon_n)_{n \in \mathbb{N}^*}$ une suite de variables i.i.d. de loi normale standard. Soit X_0 une variable aléatoire gaussienne centrée de variance $(1-a^2)^{-1}$ indépendante de la suite $(\varepsilon_n)_{n \in \mathbb{N}^*}$ On considère le processus $(X_n)_{n \in \mathbb{N}}$ défini par l'équation de récurrence :

$$X_{n+1} := a X_n + \varepsilon_{n+1}, \quad (n \in \mathbb{N}).$$

1. Calculer pour $n, h \in \mathbb{N}$ $\operatorname{cov}(X_n, X_{n+h})$. Montrer que $(X_n)_{n \in \mathbb{Z}}$ est un processus gaussien stationnaire.
2. Quelle est la densité spectrale de ce processus.
3. Soit N un entier naturel non nul. Exprimer $(X_0, \varepsilon_1, \ldots, \varepsilon_N)$ en fonction de $(X_0, \ldots, X_N)$. Déterminer la densité du vecteur aléatoire $(\varepsilon_n)_{n=0,\ldots,N}$. En déduire la densité du vecteur aléatoire $(X_n)_{n=0,\ldots,N}$.
4. Soit Γ_N la matrice de variance covariance du vecteur $(X_n)_{n=0,\ldots,N}$. En utilisant la question précédente, déterminer Γ_N^{-1}.
5. On observe $(X_n)_{n=0,\ldots,N}$, déterminer très précisement la loi conditionnelle de X_{N+1} sachant cette observation. On désire estimer le paramètre a. Ecrire la vraisemblance associée aux observations $(X_n)_{n=0,\ldots,N}$. Calculer l'estimateur du maximum de vraisemblance $\widehat{a}_N$.
6. Montrer que $\widehat{a}_N$ est un estimateur qui converge en probabilité vers a.