Agrégation externe de mathématiques, session 2019
Épreuve de modélisation, option A : Probabilités et Statistiques

(*A02-public*)

Résumé : On s'intéresse à une suite de valeurs numériques réelles dépendant du temps, par exemple les niveaux annuels maximaux de crue du Nil. Notre objectif sera de proposer une modélisation acceptable pour ces données après avoir montré que les modèles classiques (variables indépendantes ou marches aléatoires) ne sont pas adaptés. On portera ainsi notre attention sur une suite de variables aléatoires identiquement distribuées mais fortement dépendantes appelée bruit gaussien fractionnaire. On exhibera certaines propriétés de cette suite et on proposera un estimateur convergent des paramètres.

Mots clefs : Vecteur gaussien, théorème limite central, estimation, test statistique.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. La présentation, bien que totalement libre, doit être organisée et le jury apprécie qu'un plan soit annoncé en préliminaire. L'exposé doit être construit en évitant la paraphrase et mettant en lumière les connaissances, à partir des éléments du texte. Il doit contenir des illustrations informatiques réalisées sur ordinateur, ou, à défaut, des propositions de telles illustrations. Des pistes de réflexion, indicatives et largement indépendantes les unes des autres, vous sont proposées en fin de texte.*

Ce texte vous est fourni avec un fichier `Nil.txt` *permettant d'illustrer sur des données réelles certains des résultats présentés. Des instructions pour lire ce fichier à l'aide des logiciels disponibles sont données en dernière page du texte.*

Introduction

Dans de très nombreux problèmes, par exemple météorologiques, économétriques, sociologiques,..., on rencontre ce que l'on appelle des séries chronologiques, c'est-à-dire des données numériques indexées par le temps. Nous prendrons l'exemple des niveaux maxima d'un fleuve, ici le Nil, relevés chaque année entre 722 et 1281 [1]. Le but que nous nous fixons est de pouvoir modéliser "correctement" ces données par une suite de variables aléatoires sans utiliser de possibles variables "explicatives" exogènes (pour le niveau d'un fleuve cela pourrait être la quantité de pluies, l'ensoleillement, la construction d'une digue,...etc). L'intérêt d'offrir un tel modèle est essentiellement de pouvoir prédire le niveau maximal à très court terme ou bien d'évaluer des probabilités d'événements extrêmes (grande ou faible crue).

1. Les données numériques sont contenues dans le fichier `Nil.txt`

1. Premières modélisations

Commençons par formaliser un tel problème : on dispose de données (t, X_t), où t désigne le temps dans une certaine unité (dans notre exemple, des années) et X_t la variable qui nous intéresse à l'instant t. On suppose que la suite $(X_t)_{t\in\mathbf{N}}$ est une suite de variables aléatoires. On dira alors que $(X_t)_{t\in\mathbf{N}}$ est une série chronologique. On supposera de plus que pour tout $t \in \mathbf{N}$, $\mathbb{E}(X_t^2) < \infty$. Enfin, on observe $(X_0, X_1, \dots, X_n)$, où $n \in \mathbf{N}^*$.

Nous utiliserons la propriété suivante souvent associée aux séries chronologiques :

Définition 1. *On dit que $(X_t)_{t\in\mathbf{N}}$ est une série stationnaire lorsque pour tout $m \in \mathbf{N}^*$, tout $(t_1, \dots, t_m) \in \mathbf{N}^m$ et tout $c \in \mathbf{N}$, $(X_{t_1}, \dots, X_{t_m})$ a la même distribution que $(X_{t_1+c}, \dots, X_{t_m+c})$.*

Cela signifie que tout vecteur aléatoire issu de la série est invariant en distribution pour toute translation. En particulier, si $(X_t)_{t\in\mathbf{N}}$ est stationnaire, les X_t sont toutes identiquement distribuées.

Pour toute la suite on supposera la suite $(X_t)_{t\in\mathbf{N}}$ stationnaire.

Deux exemples simples de séries chronologiques sont souvent employés dans une première tentative de modélisation. En premier lieu, on considère souvent une suite de variables aléatoires identiquement distribuées et indépendantes. Concernant notre exemple du Nil, ainsi que dans de nombreux autres cas, c'est essentiellement l'hypothèse d'indépendance qui pose problème. On s'intéresse ici au moyen de tester l'hypothèse

$$H_0 : \text{“}(X_t)_{t\in\mathbf{N}} \text{ est une suite de variables indépendantes et de même loi”}$$

contre l'hypothèse

$$H_1 : \text{“}(X_t)_{t\in\mathbf{N}} \text{ est une suite (stationnaire) de variables non indépendantes”}.$$

Une possibilité pour mettre en œuvre un tel test est d'utiliser la covariance (même si celle-ci ne caractérise pas de manière exhaustive l'indépendance). On définit donc :

$$(1) \quad r(k) := \operatorname{cov}(X_0, X_k) = \mathbb{E}\Big[\big(X_0 - \mathbb{E}[X_0]\big)\big(X_k - \mathbb{E}[X_k]\big)\Big] = \mathbb{E}[X_0X_k] - \mathbb{E}[X_0]\mathbb{E}[X_k] \quad \text{pour} \quad k \in \mathbf{N}.$$

Comme on a supposé la suite $(X_t)_{t\in\mathbf{N}}$ stationnaire, on a $r(k) = \operatorname{cov}(X_i, X_{i+k})$ pour tout $i \in \mathbf{N}$. La covariance dépendant des unités de mesure, on préférera utiliser la *corrélation*, définie par $\rho(k) := \frac{r(k)}{r(0)} \in [-1, 1]$ pour $k \in \mathbf{N}$. On a en particulier la proposition suivante :

Proposition 2. *Sous H_0, la suite $(\rho(k))_{k\in\mathbf{N}}$ est donnée par $\rho(0) = 1$ et $\rho(k) = 0$ si $k \geq 1$.*

Pour tester H_0 contre H_1, considérons les estimateurs “naturels” de $r(k)$ et de $\rho(k)$:

$$(2) \qquad \widehat{r}_n(k) := \frac{1}{n-k+1} \sum_{j=0}^{n-k} \big(X_j - \bar{X}_n\big)\big(X_{j+k} - \bar{X}_n\big) \quad \text{et} \quad \widehat{\rho}_n(k) := \frac{\widehat{r}_n(k)}{\widehat{r}_n(0)},$$

où $\bar{X}_n = \frac{1}{n+1} \sum_{j=0}^{n} X_j$. Ces estimateurs sont convergents :

Proposition 3. *Sous H_0, pour tout $k \in \mathbf{N}$, $\widehat{r}_n(k) \xrightarrow[n\to+\infty]{p.s.} r(k)$ et $\widehat{\rho}_n(k) \xrightarrow[n\to+\infty]{p.s.} \rho(k)$.*

Démonstration. On peut développer $\widehat{r}_n(k)$ en :

$$\widehat{r}_n(k) = \frac{1}{n-k+1}\sum_{j=0}^{n-k} X_j X_{j+k} - \frac{\bar{X}_n}{n-k+1}\sum_{j=0}^{n-k} X_j - \frac{\bar{X}_n}{n-k+1}\sum_{j=0}^{n-k} X_{j+k} + \bar{X}_n^2 \tag{3}$$

Sous H_0, la loi des grands nombres s'applique aux trois derniers termes, dont la somme converge vers $-\mathbb{E}[X_0]\mathbb{E}[X_k]$. Pour le premier terme, on remarque que

$$\sum_{j=0}^{n-k} X_j X_{j+k} = \sum_{q=0}^{k} S_{q,n}, \quad \text{avec} \quad S_{q,n} = \sum_{\substack{0\le j\le n-k \\ j\equiv q \mod (k+1)}} X_j X_{j+k}. \tag{4}$$

La loi des grands nombres permet ensuite de conclure. □

Pour choisir entre les hypothèses H_0 et H_1, on aimerait définir un test statistique utilisant $\widehat{\rho}_n(k)$. Une convergence presque-sûre n'étant pas utilisable à cette fin, on préférera établir un théorème limite central. Pour ce faire, on va se placer sous l'hypothèse que $(X_t)_{t\in\mathbf{N}}$ est une série chronologique gaussienne, c'est-à-dire que pour tout $m \in \mathbf{N}$ et tout $t_1, \dots, t_m \in \mathbf{R}$, le vecteur $(X_{t_1}, \dots, X_{t_m})$ est gaussien. Dans ce cadre, on a le théorème suivant, dont la démonstration est donnée plus bas.

Théorème 4. *Sous H_0 et si $(X_t)_{t\in\mathbf{N}}$ est une série chronologique gaussienne, on a pour tout $k \in \mathbf{N}^*$ les convergences en loi*

$$\sqrt{n}\,\widehat{r}_n(k) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0, r^2(0)), \quad \textit{et} \quad \sqrt{n}\widehat{\rho}_n(k) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathcal{N}(0,1). \tag{5}$$

On pourra ainsi tracer sur un graphe les points $(k, \widehat{\rho}_n(k))$ pour $k = 0, \dots, K$: c'est ce que l'on appelle le corrélogramme ($K = 30$ permet d'avoir déjà une bonne information sur la corrélation). On déduit également un test de niveau quelconque sur la corrélation pour un $k \geq 1$ fixé, en testant $\{\rho(k) = 0\}$ contre $\{\rho(k) \neq 0\}$. Sur la série des données du Nil (voir corrélogramme en Figure 1), on montre ainsi que H_0 est refusée pour $k = 1, \dots, 30$ avec un niveau de confiance de 95%. La modélisation par une suite de variables aléatoires indépendantes et de même loi est donc inadaptée.

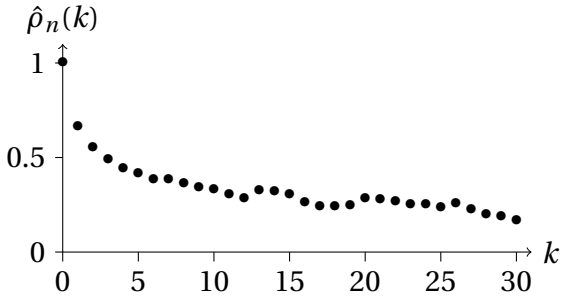

FIGURE 1. Valeurs des $\hat{\rho}_n(k)$ pour $k = 1, \dots, 30$, pour les données du Nil.

Preuve du Théorème 4. On considère la matrice $M_n(k) = (m_{ij}(k))_{0\le i,j\le n}$ où $m_{ij}(k) = 1/2$ si $|i-j| = k$ et 0 sinon (on travaille ici avec $k \ge 1$). Ainsi, on montre que, pour $\mu = \mathbb{E}(X_0)$,

$$\sum_{j=0}^{n-k} \big(X_j - \mu\big)\big(X_{j+k} - \mu\big) = (X - \mu I)' \cdot M_n(k) \cdot (X - \mu I) = r(0) \times G' \cdot M_n(k) \cdot G, \tag{6}$$

où $I = (1, \cdots, 1)'$ ($'$ désigne la transposition), et $G = (G_0, \ldots, G_n)'$, $(G_j)_{j\in\mathbf{N}}$ étant une suite de variables gaussiennes centrées réduites indépendantes (sous H_0). Comme $M_n(k)$ est symétrique on peut la diagonaliser sous la forme $Q' \cdot D \cdot Q$, avec Q une matrice orthogonale et D la matrice diagonale des valeurs propres $(\lambda_i(n))_{0\le i\le n}$. Donc, on a pour chaque n, avec $Z = (Z_0, \ldots, Z_n)' = QG$ un vecteur de variables gaussiennes centrées réduites indépendantes :

$$\tilde{r}_n(k) := \frac{1}{n-k+1} \sum_{j=0}^{n-k} \big(X_j - \mu\big)\big(X_{j+k} - \mu\big) = \frac{r(0)}{n-k+1} \sum_{i=0}^{n} \lambda_i(n) Z_i^2 \,. \tag{7}$$

Or $\sum_{i=0}^{n} \lambda_i(n) = \mathrm{Trace}(M_n(k)) = 0$, donc :

$$\tilde{r}_n(k) = \frac{r(0)}{\sqrt{n-k+1}} \sum_{i=0}^{n} \Big(\frac{\lambda_i(n)\sqrt{2}}{\sqrt{n+1-k}}\Big) W_i, \tag{8}$$

avec $W_i := (Z_i^2 - 1)/\sqrt{2}$, vérifiant $\mathbb{E}(W_i) = 0$ et $\mathbb{E}(W_i^2) = 1$. On peut maintenant utiliser le Lemme 5 ci-dessous (admis) pour en déduire un théorème limite central pour $\tilde{r}_n(k)$, puisque

$$\sum_{i=0}^{n} \lambda_i^2(n) = \mathrm{Trace}(M_n^2(k)) = \frac{1}{2}(n+1-k) \tag{9}$$

et que, d'après un résultat classique d'algèbre linéaire,

$$\max_{0\le i\le n} |\lambda_i(n)| \le \max_{0\le i\le n} \sum_{j=0}^{n} |m_{ij}(k)| \le 1. \tag{10}$$

Enfin, comme

$$\widehat{r}_n(k) = \tilde{r}_n(k) + (\bar{X}_n - \mu)^2 + (\mu - \bar{X}_n) \frac{1}{n-k+1} \sum_{j=0}^{n-k} (X_j + X_{j+k} - 2\mu) \tag{11}$$

avec $\sqrt{n} \times \mathbb{E}(\bar{X}_n - \mu)^2 \underset{n\to+\infty}{\longrightarrow} 0$, on obtient le résultat. □

Lemme 5 (admis)**.** *Soit $(Y_i)_{i\in\mathbf{N}}$ une suite de variables aléatoires indépendantes et de même loi admettant un moment d'ordre 2 telle que $\mathbb{E}(Y_i) = 0$ et $\mathrm{var}(Y_i) = 1$ pour tout $i \in \mathbf{N}$. Soit la suite doublement indicée $(a_i^{(N)})_{(N\in\mathbf{N}^*, 1\le i\le N)}$ telle que pour tout $N \in \mathbf{N}^*$, $\sum_{i=1}^{N} (a_i^{(N)})^2 = 1$. Alors si $\max_{1\le i\le N} |a_i^{(N)}| \underset{N\to+\infty}{\longrightarrow} 0$, la suite $\sum_{i=1}^{N} a_i^{(N)} Y_i$ converge en loi vers $\mathcal{N}(0,1)$.*

2. Modélisation par une suite de variables aléatoires fortement dépendantes

Le corrélogramme nous donne plus de renseignements que la non indépendance entre les données. Il permet d'observer comment se comporte approximativement la corrélation en fonction de l'écart k choisi. Ici, $r(k)$ semble décroître avec k comme k^{-D} où $0 < D < 1$.

Un modèle possible pour ce type de comportement est fourni par le *bruit gaussien fractionnaire*. Il s'agit d'une série chronologique gaussienne stationnaire $(X_t)_{t\in\mathbf{N}}$ telle que :

$$r(k) = \frac{1}{2}\sigma^2\big(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}\big), \tag{12}$$

avec $\sigma^2 > 0$ et $H \in]0,1[$, avec $\mathbb{E}(X_0) = \mu$, fixé. Nous admettrons qu'une telle suite $(X_t)_{t\in\mathbf{N}}$ existe. En utilisant un développement de Taylor, on montre que :

$$r(k) \sim \sigma^2 H(2H-1)\frac{1}{k^{2-2H}} \quad \text{pour } k \to \infty. \tag{13}$$

On a donc bien un modèle possible pour nos données pour $H \in]1/2,1[$. Pour $H = 1/2$, on remarque que le bruit gaussien fractionnaire est une suite de variables aléatoires indépendantes et de même loi.

On peut aussi définir le *processus agrégé* de $(X_t)_{t\in\mathbf{N}}$, noté $(Y_t)_{t\in\mathbf{N}}$ et défini par $Y_t = \sum_{i=0}^{t-1}(X_i - \mu)$ pour tout $t \in \mathbf{N}$. On montre alors que $(Y_t)_{t\in\mathbf{N}}$ est une série chronologique à accroissements stationnaires, centrée et de covariance :

$$\mathbb{E}\big[Y_t Y_s\big] = \frac{1}{2}\sigma^2\big(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\big) \quad \text{pour tout } (s,t) \in \mathbf{N}^2. \tag{14}$$

Cette série présente la propriété intéressante suivante, appelée *autosimilarité* : pour tout $m \in \mathbf{N}^*$, tout $(t_1,\dots,t_m) \in \mathbf{N}^m$ et tout $c \in \mathbf{N}$, les vecteurs $\big(Y_{ct_1},\dots,Y_{ct_m}\big)$ et $c^H\big(Y_{t_1},\dots,Y_{t_m}\big)$ ont même loi. Ainsi, si on considère la trajectoire de $(Y_t)_{t\in\mathbf{N}}$, que l'on "zoome" sur une partie de cette trajectoire, cette dernière va présenter le même type de distribution aléatoire que la trajectoire initiale : il y a invariance (en distribution) par changement d'échelle.

3. Estimation du paramètre H du modèle

Une première idée pour estimer le paramètre H du modèle serait d'utiliser les corrélations empiriques. En effet, puisque asymptotiquement $\rho(k) \simeq Ck^{2H-2}$ (où $C > 0$), on pourrait espérer que $\widehat{\rho}_n(k)$ donne une bonne approximation de Ck^{2H-2}. En considérant $\log(\widehat{\rho}_n(k))$ avec plusieurs valeurs de k, on aboutirait à un modèle linéaire. Une régression linéaire permettrait d'estimer $2-2H$, donc H. Ceci est effectivement possible, mais n'est concrètement pas intéressant, car la vitesse de convergence de l'estimateur n'est pas optimale, et dépend même de H. Ceci se retrouve par le calcul de $\mathrm{var}\big(\widehat{r}_n(k)\big)$ qui est, à une constante près, en n^{4H-4}.

On va plutôt considérer les *variations quadratiques* de $(Y_t)_{t\in\mathbf{N}}$. Pour $a \in \mathbf{N}^*$, on définit $(V_t^{(a)})_{t\in\mathbf{N}^*}$ par $V_t^{(a)} := (Y_{a(t+2)} - 2Y_{a(t+1)} + Y_{at})^2$ pour $t \in \mathbf{N}^*$. On montre :

Proposition 6. *Si $(X_t)_{t\in\mathbf{N}}$ est un bruit gaussien fractionnaire de paramètres $H \in]1/2,1[$ et $\sigma^2 > 0$, il existe $C(H) > 0$ tel que pour tout $t \in \mathbf{N}^*$*

$$\mathbb{E}(V_t^{(a)}) = \sigma^2\big(4 - 2^{2H}\big) \times a^{2H} \text{ et } \big|\mathrm{cov}(V_t^{(a)}, V_{t+k}^{(a)})\big| \le \frac{\sigma^4 C(H)}{k^{8-4H}} a^{4H} \text{ pour tout } |k| \ge 1. \tag{15}$$

Démonstration. La partie espérance se déduit de (14). La partie covariance utilise le fait que lorsque (Z_1,Z_2) est un vecteur gaussien centré, alors $\mathrm{cov}(Z_1^2,Z_2^2) = 2\big(\mathrm{cov}(Z_1,Z_2)\big)^2$, et également un développement de Taylor. □

Le fait que l'espérance précédente est une fonction puissance en $2H$ nous amène à penser que la moyenne empirique $S_n(a) = ([n/a]-1)^{-1}\sum_{t=0}^{[n/a]-2} V_t^{(a)}$ associée à cette espérance permettra, après passage au logarithme, d'estimer H. Or, comme la covariance décroît suffisamment vite en k (voir (15)), on peut en déduire que $S_n(a)$ vérifie bien un théorème limite central :

Théorème 7. *Si $(X_t)_{t\in\mathbf{N}}$ est un bruit gaussien fractionnaire de paramètres $H\in]1/2,1[$, $\mu\in\mathbf{R}$ et $\sigma^2>0$, pour tout $a\in\mathbf{N}^*$,*

$$\sqrt{n}\big(S_n(a)-\sigma^2\big(4-2^{2H}\big)a^{2H}\big) \xrightarrow[n\to+\infty]{\mathscr{L}} \mathscr{N}\Big(0,\gamma(a)\Big) \tag{16}$$

avec $\gamma(a)=\frac{1}{2}\sigma^4 a^{4H+1}\sum_{\ell=-\infty}^{\infty}\big(|\ell+2|^{2H}+|\ell-2|^{2H}-4|\ell+1|^{2H}-4|\ell-1|^{2H}+6|\ell|^{2H}\big)^2$.

Suggestions et pistes de réflexion

► *Les pistes de réflexion suivantes ne sont qu'indicatives et il n'est pas obligatoire de les suivre. Vous pouvez choisir d'étudier, ou non, certains des points proposés, de façon plus ou moins approfondie, mais aussi toute autre question à votre initiative. Vos investigations comporteront une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats. À défaut, si vos illustrations informatiques n'ont pas abouti, il est conseillé d'expliquer ce que vous auriez souhaité mettre en œuvre.*

— *Modélisation.*
Expliquer et commenter les différents choix de modélisation considérés (en particulier quelles sont leurs limites ?). Comment aurait-on pu montrer qu'une modélisation des données par une marche aléatoire n'était pas satisfaisante ? Quels auraient pu être d'autres modèles ?

— *Développements mathématiques.*
À partir des indications données, reconstituer les preuves des propositions 2, 3 ou 6, et des théorèmes 4 ou 7. Montrer que $(Y_t)_{t\in\mathbf{N}}$ vérifie bien la propriété d'autosimilarité.

— *Étude numérique.*
  — Le fichier `Nil.txt` contient les données de crues du Nil. Tracer le corrélogramme pour ces données (on choisira $K=30$) et vérifier que l'indépendance des données n'est pas plausible. Tracer la trajectoire des $(Y_t)_{t\in\mathbf{N}}$. Tracer les points $(\log a, \log S_n(a))$ pour $1\leq a\leq m$. Est-ce bien linéaire ? Calculer alors l'estimateur de H.
  — Simuler la trajectoire ($n=1000$ par exemple et $H=0.9$) d'un bruit gaussien fractionnaire (on pourra par exemple utiliser une racine carrée de sa matrice de covariance). Calculer également $S_n(a)$. Comment pourrait-on faire numériquement pour vérifier le théorème limite central sur $S_n(a)$?

Lecture des fichiers de données

Indications pour la session 2019

Ce document est agrafé au présent texte car ce dernier est associé à un jeu de données réelles `Nil.txt`, qui vous est fourni sous format texte.

La première ligne de ce fichier de données est une ligne de titre, contenant du texte indiquant ce qu'on trouve dans chacune des colonnes ; les lignes suivantes contiennent des données numériques.

Les instructions suivantes expliquent comment charger ce fichier sous différents logiciels de sorte que les données se retrouvent dans une matrice `A`.

Commencez par déplacer `Nil.txt` dans votre répertoire personnel :

— ouvrir le répertoire `Données` se trouvant sur le bureau en double-cliquant sur la dernière icône (bleue) de la colonne de gauche

— choisir le fichier `Nil.txt` et le déplacer à la souris (ou le copier) dans le `Répertoire personnel` (première icône du bureau de la colonne de gauche).

Lecture sous Scilab.

```
[A,text] = fscanfMat("Nil.txt")
```

La matrice `A` contient les donnés numériques, `text` contient la ligne de texte.

Lecture sous R. La commande

```
B<-read.table("Nil.txt", header = TRUE)
```

crée une variable `B` qui n'est pas exactement une matrice numérique, mais plutôt une structure. Les lignes suivantes transforment cette structure en matrice numérique si besoin est :

```
D<-dim(B); n<-D[1]; p<-D[2];
A<-matrix(0,n,p);
for (i in 1:p) A[,i]<-B[,i];
```

Lecture sous Octave. Commencer par détruire la première ligne du fichier `Nil.txt`, puis exécuter :

```
A = load("Nil.txt")
```

Lecture sous Python. On utilise la fonction `loadtxt` de la bibliothèque `numpy` :

```
from numpy import loadtxt
A = loadtxt("Nil.txt", skiprows=1)
```

Le résultat obtenu est de type `array`. L'option `skiprows=1` sert à ignorer la ligne de commentaire du fichier.

Description du fichier `Nil.txt` : Ce fichier a 560 lignes (ligne de titre non comptée) et 2 colonnes. La première colonne est l'année, la seconde le niveau de crue du Nil.