Valeurs extrêmes pour l'oceanométéorologie
Notes de cours STA2102U

V. Monbet

14 septembre 2007

Table des matières

Chapitre 1

Introduction

Voir les transparents de Michel Olagnon (Ifremer).
Vocabulaire : Hs, Tp, ...

Chapitre 2

Introduction à la théorie des valeurs extrêmes

2.1 Généralités

La théorie des valeurs extrêmes a émergée comme étant l'une des plus importantes disciplines des sciences appliquées dans les 60 dernières années. Cette théorie est utilisé en environnement et en particulier en météorologie mais aussi en médecine et dans de nombreuses autres disciplines :

- *portfolio in industrial insurance*
- estimation du risque sur les marchés financiers
- prediction de traffic en télécommunication

Une des principales caractéristiques de l'analyse des valeurs extrêmes est que l'on cherche à caractériser un comportement stochastique à des niveaux inhabituellement élevés (ou faibles). En particulier, on doit estimer la probabilité d'évènements qui n'ont jamais (ou presque jamais) été observés. Par exemple, on cherche

- la probabilité d'une vague très haute sur un site pétrolier
- la probabilité d'observer une crue de plus de m mètres dans un centre ville
- encore la probabilité qu'un niveau de marée très élevé combiné à de fortes vagues produise un franchissement d'eau au dessus d'une digue de protection.

La question peut-être posée différement : quelle doit être la hauteur de la plate forme pétrolière pour que sur cent ans la probabilité d'avoir un franchissement d'eau soit inférieure à un risque fixé. Des données de hauteur de vague peuvent être disponibles mais sur une période beaucoup plus courte, par exemple 10 ou 20 ans. Le challenge est alors d'estimer quel niveau d'eau et hauteur de vagues risque d'être observé dans les 100 prochaines années à partir de l'historique des données. La théorie des valeurs extrême fournit un cadre pour réaliser ce genre d'extrapolation.

Dans ce cadre, nous considérons 4 points :

1. **Méthodes d'estimations** - comment inférer les paramètres du modèle à partir de l'historique des observations. Nous nous concentrerons essentiellement sur les méthodes basées sur des estimateurs du maximum de vraisemblance et des méthodes de moments. Ces techniques ont l'avantage d'être portable d'un modèle à l'autre.

2. **Quantification de l'incertitude** - Dans le contexte des valeurs extrêmes, une faible modification de l'estimation des paramètres peut engendrer de grandes différences au moment de l'extrapolation. Nous verrons que lorsque l'inférence est basée sur une fonction de vraisemblance on peut en déduire facilement des estimateurs de l'incertitude.
3. **Validation de modèle** - Il sera important de valider le choix du modèle. Nous utiliserons essentiellement des techniques basées sur des techniques de simulation.
4. **Utilisation du maximum d'information disponible** - Donnons un exemple : dans de nombreuses analyses de valeurs extrêmes on considère uniquement l'échantillon des maxima sur une période de référence (par exemple les maxima annuels) pour construire le modèle. Ainsi, si une année donnée on observe d'autres valeurs plus fortes que les extrema d'autres années, on ne conserve pas ces valeurs. On perd donc de l'information.

2.2 Exemple

Pour illustrer ce cours, nous considérerons entre autres comme exemples l'historique des températures journalières et l'intensité du vent enregistrées à Brest depuis le 1er janvier 1976 (voir Figure 2.3).

En supposant que la série de température ne présente pas de tendance pluri annuelle, les questions que l'on va se poser sont par exemple
- Quelle est la probabilité qu'on observe une température maximale annuelle de plus de 30 degrés ?
- Avec quelle fréquence peut-on s'attendre à voir revenir une telle température c'est à dire quelle est la période de retour de la température 30 degrés ?
- Quelle température maximale doit-on attendre dans les 100 prochaines années c'est à dire quel est le niveau de retour à 100 ans ?

On se posera le même type de question pour le vent avec par exemple comme application l'évaluation des risques de rupture sur une éolienne, un pont, ...

On peut aussi considérer des séries de températures en deux sites (par exemple Brest et Paris) ou la température et la pression.

Vagues et surcote.

2.3 Plan du cours (*pas à jour ?*)

1. Séquences i.i.d.
   (a) Rappels sur l'ajustement de lois
   (b) Théorie classique des valeurs extrèmes
   (c) Modèles à seuils
2. Séquences dépendantes
   (a) Rappels sur les séries temporelles
   (b) Modèles d'extrèmes pour les séries stationnaires

(c) Extremal index et modèles à seuils

(d) Cas multivarié

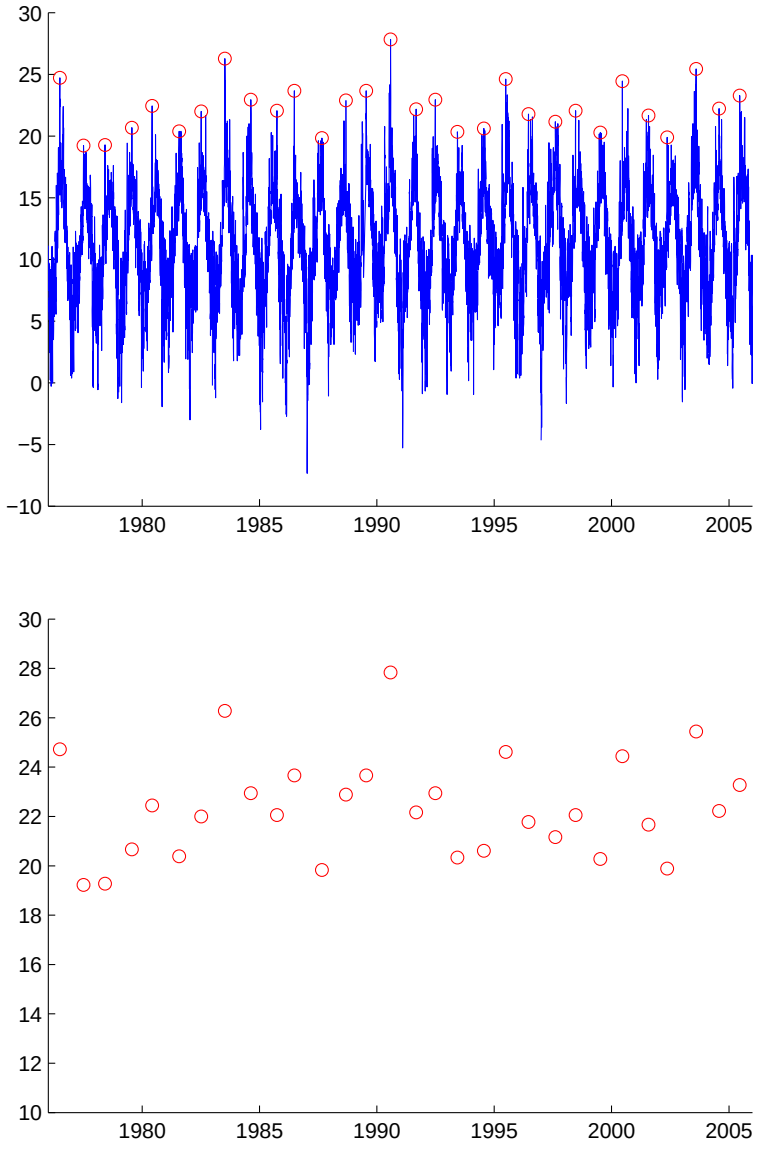

FIG. 2.1 – Températures journalières enregistrées à Brest de 1976 à 2005 (haut) et maxima annuels (bas)

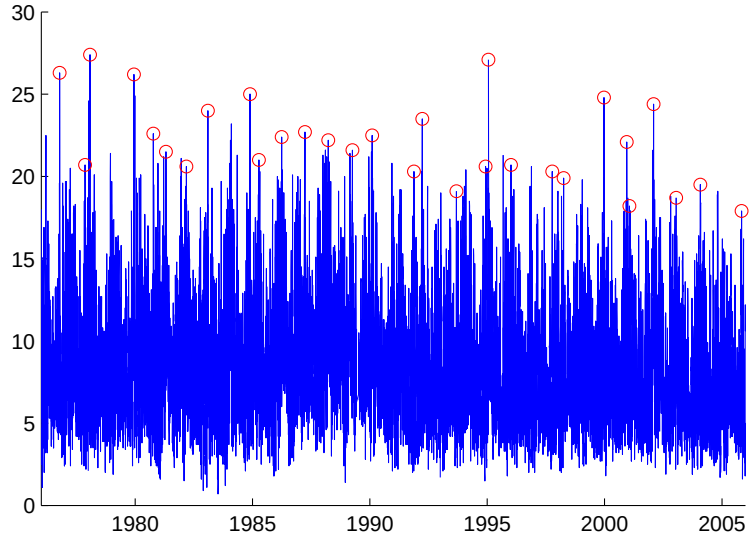

FIG. 2.2 – Intensité du vent enregistré à Brest de 1976 à 2005 et maxima annuels (ronds rouges)

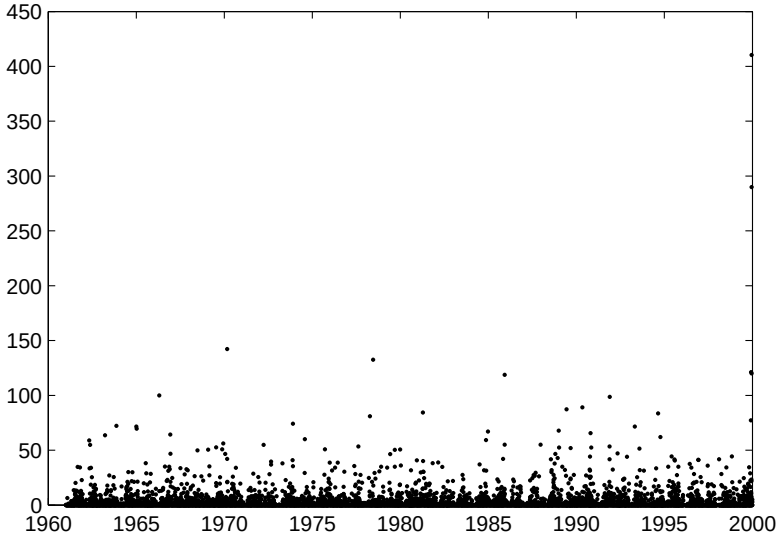

FIG. 2.3 – Précipitations journalières au Venezuela de 1960 à 2000

Chapitre 3

Rappels sur l'ajustement de lois

Soit $X \in \mathbf{R^p}$ une variable aléatoire définie sur Ω de fonction de répartition[1] $F_\theta(x) = P(X \leq x)$ pour tout $x \in \Omega$ et de densité de probabilité[2] f_θ.

En océanométéorologie, certaines lois de probabilité sont utilisées plus souvent :

- **Loi de Poisson** - $f(x) = \frac{e^{-\theta}\theta^x}{x!}$, $x \in \Omega = \{1, 2, 3, ...\}$. La loi de Poisson servira par exemple à modéliser les temps d'arrivée d'évènements extrêmes.

- **Loi de Gauss** de moyenne μ et de variance σ^2 et notée $X \sim \mathcal{N}(\mu, \sigma^2)$. On notera Φ sa fonction de répartition et ϕ la densité de probabilité associée. La loi normale est utilisée, par exemple, pour modéliser l'élévation de la surface libre en un point de l'océan (eau profonde).

- **Loi log-normale** - X est disctribuée suivant une loi log-normale si $\log(X)$ suit une loi normale.

- **Loi de Rayleigh** - Si X_1 et X_2 suivent des lois de Gauss centrées de variance σ, alors $X = \sqrt{X_1 + X_2}$ suit une loi de Rayleigh telle que $F(x) = 1 - \exp\left(-\frac{x^2}{2\sigma^2}\right)$, $x \geq 0$ et $\theta > 0$. La loi de Rayleigh est utilisée pour modéliser la hauteur des vagues dans une mer formée en eau profonde (*à préciser*).

- **Loi Gamma** - $f(x) = x^{a-1)} \exp(-x/b)/\gamma(a)/b^a$, $a, b > 0$, $x \geq 0$. La loi Gamma peut être utilisée pour modéliser l'intensité du vent.

3.1 Ajustement de loi et maximum de vraisemblance

La méthode du maximum de vraisemblance est une méthode générale et robuste pour l'estimation des paramètres d'une distribution.

Les estimateurs du maximum de vraisemblance sont consistents et, quand la taille de l'échantillon tend vers l'infini, ils convergent en loi vers une variable aléatoire de loi de Gauss ayant pour moyenne le paramètre à estimer et pour variance l'inverse de l'information de Fisher. Ce résultat

[1] En anglais : *cumulative distribution function*
[2] En anglais : *probability density function*

permet de construire des intervalles de confiance et des tests d'hypothèses pour le paramètre à estimer et pour toute fonction suffisament régulière du paramètre par la delta methode.

Delta méthode

La delta méthode est une approche générale permettant de construire des intervalles de confiance pour des fonctions d'estimatuer du maximum de vraisemblance. On construit une approximation linéaire de la fonction basée sur un développement de Taylor et on approche la variance de la fonction par la variance de l'approximation linéaire.

Soit $\hat{\theta}$ l'estimateur du maximum de vraisemblance d'un paramètre θ et h une fonction régulière. Alors on peut montrer que

$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) \to \mathcal{N}\left(0, \nabla h(\hat{\theta})' Var(\hat{\theta}) \nabla h(\hat{\theta})\right)$$

$$h(\hat{\theta}) - h(\theta) = (\hat{\theta} - \theta)\nabla h(\hat{\theta}) + \text{reste}$$

d'où

$$\begin{aligned} Var\left(h(\hat{\theta}) - h(\theta)\right) &= E\left((h(\hat{\theta}) - h(\theta)))'(h(\hat{\theta}) - h(\theta)))\right) + \text{reste} \\ &= E(\nabla h(\hat{\theta})'(\hat{\theta} - \theta)'(\hat{\theta} - \theta)\nabla h(\hat{\theta}) + \text{reste} \\ &\approx \nabla h(\hat{\theta})' Var(\hat{\theta} - \theta)\nabla h(\hat{\theta}) \end{aligned}$$

Pour plus de détails voir http ://www.ensae.fr/ParisTech/SE203/SE203.html#documents

3.2 Choix et validation de modèle

Dans la pratique, il se pose souvent le problème de choisir un modèle de loi connaissant une séquence de réalisations d'une variable aléatoire.

Par exemple, nous avons tracé un histogramme à partir des intensités de vent pour les mois de janvier à Brest (Figure 3.1). Nous donnons aussi sur cette figure un histogramme pour les précipitations (non nulles) au Vénézuela. La question est alors : de quelles distributions sont issues ces données ?

Dans les deux cas, on peut envisager d'ajuster une loi Gamma d'après la forme de l'histogramme.

3.2.1 Probability plot

Le tracé sur du papier 'probabilisé' est un outil graphique d'aide au diagnostique. L'exemple le plus connu est celui de la droite de Henry (ou qqplot). Pour un échantillon $\{x_1, \cdots, x_n\}$ et une fonction de répartition F_θ donnés, il s'agit de tracer $F_\theta^{-1}(\hat{F}(x_{(i)}))$ en fonction de $x_(i)$ sous la forme d'un nuage de point. $x_(i)$ est la statistique de rang i pour l'échantillon et $\hat{F}$ un estimateur empirique de la fonction de répartition.

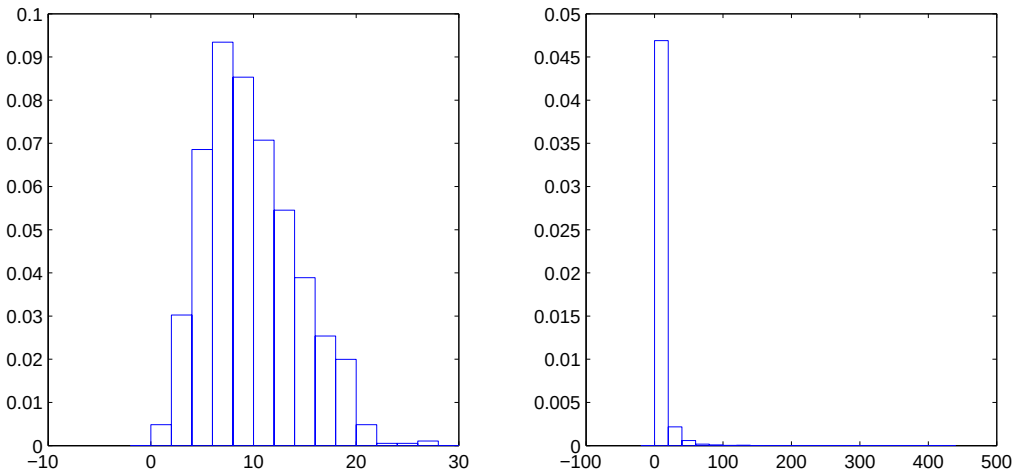

FIG. 3.1 – Histogrammes des données de vent à Brest (mois de janvier) à gauche et des précipitations journalières (valeurs strictement positives) au Vénézuela

Si la distribution dont est issu l'échantillon suit la loi F_θ alors $F_\theta^{-1}(\hat{F})$ est proche de l'identité et le nuage de points forme donc approximativement une droite.

On choisira par exemple pour l'estimateur empirique :

$$\hat{F}(x_{(i)}) = \frac{i - 0.5}{n}$$

Ce choix permet de ne pas avoir $\hat{F}(x_{(1)}) = 0$ ni $\hat{F}(x_{(n)}) = 1$ c'est à dire qu'on ne sous entend pas que le domaine de définition de la variable est borné.

La figure 3.2 représente les données de l'intensité du vent à Brest et les données de pluviométrie à Vénézuela sur du papier de distribution Gamma. On observe que, sur le graphique correspondant aux données de vent, les points sont approximativement alignés. On peut en déduire que la loi Gamma est un bon modèle pour l'intensité du vent. En revanche on remarque que ce n'est pas le cas pour les données de pluie. La forme de la courbe montre que la distribution gamma ne parvient pas à modéliser l'occurence des valeurs les plus fortes. On peut alors essayer de modéliser le log ou la racine des données. La figure 3.3 montre que la transforamtion en log est trop forte. En revanche la transformation en racine semble être assez satisfaisante, en particulier pour la queue de la loi.

Cependant, cet outil de diagnostique est insuffisant bien que très utile et il faut en général confirmer les résultats par des tests d'adéquation.

3.2.2 Tests d'adéquation

Les tests d'adéquation les plus usuels pour les distributions continues sont
- le test de Kolmogorov-Smirnov (KS)
- le test de Cramer-von Mises

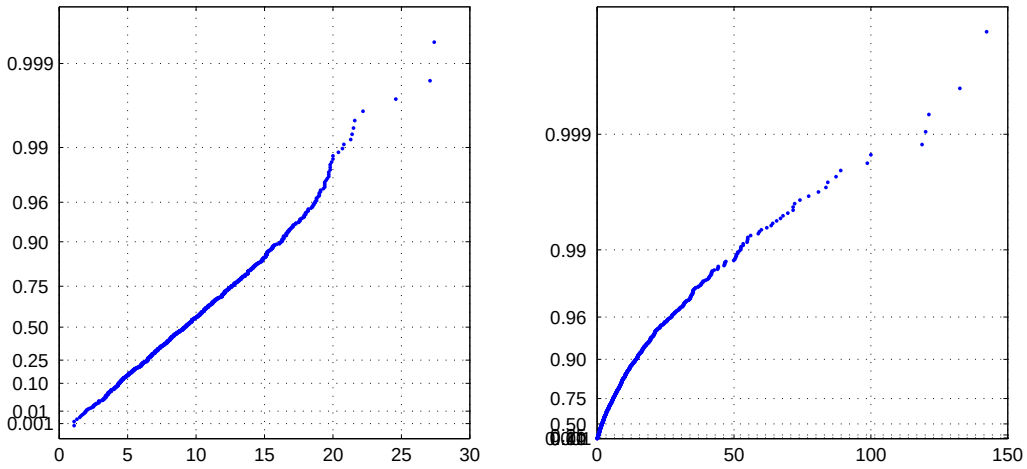

FIG. 3.2 – Tracé des données de vent à Brest (mois de janvier) à gauche et de précipitation journalière (valeurs strictement positives) au Vénézuela sur du papier de distribution Gamma

Pour l'hypothèse nulle *l'intensité du vent à Brest suit une loi Gamma de paramètres (4.44,2.20)*, le test de KS retourne un degré de signification de l'ordre de 0.52. Ce qui confirme l'intuition donnée par le tracé sur du papier de distribution Gamma.

Pour l'hypothèse nulle *la racine de la pluviométrie au Vénézuela suit une loi Gamma de paramètres (1.77,1.02)*, le test de KS retourne un degré de signification de de l'ordre de 10^{-13}.

3.2.3 Tests d'ajustement de Monte Carlo

Lorsque les distributions étudiées sont complexes, il peut être avantageux de faire des tests basés sur des simulations. Le principe consiste à simuler un grand nombre de données selon la loi théorique et de tester si les données observées appartiennent à l'intervalle de fluctuation (ou intervalle de Pari) des simulations.

Intervalle de fluctuation

Pour estimer des intervalles de fluctuation pour un paramètre θ d'une distribution F correspondant à un échantillon de taille n, on peut utiliser l'algorithme suivant.

1. `Simuler` B `échantillons de taille` n `selon la distribution` F
2. `Pour chaque échantillon` $b = 1, \cdots, B$, `calculer l'estimation empirique` $\hat{\theta}^{(b)}$
3. `Un intervalle de fluctuation empirique au risque` α `a pour bornes les statistiques d'ordre partie entière de` $n\alpha/2$ `et partie entière de` $n(1-\alpha/2)$ `de l'échantillon` $\{\theta^{(1)}, \cdots, \theta^{(B)}\}$.

Un exemple est donné Figure 3.4 pour les données d'intensité du vent. On remarque que l'histogramme et la fonction de répartition empirique de l'intensité du vent sont comprises dans l'intervalle

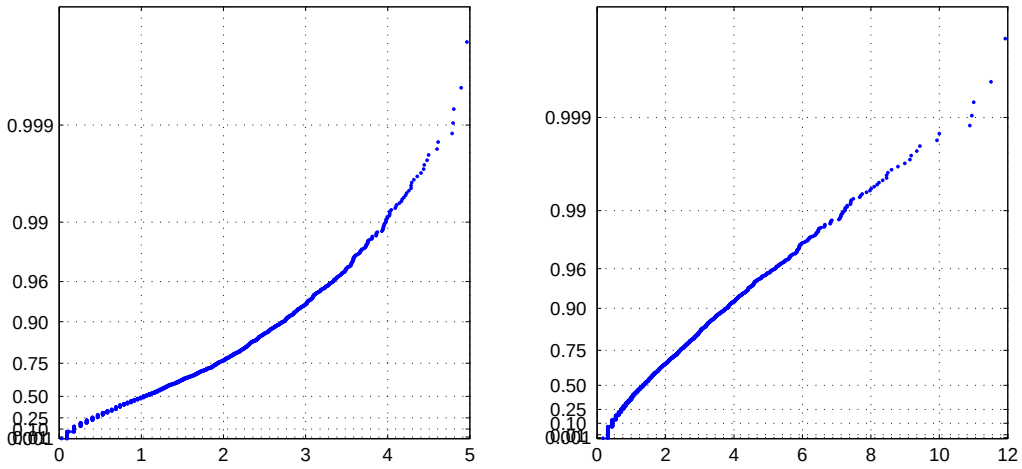

FIG. 3.3 – Tracé des données de précipitation journalière (valeurs strictement positives) au Vénézuela sur du papier de distribution Gamma pour le logarithme des données (à gauche) et pour la racine des données (à droite)

de fluctuation. On en déduit que l'observation est un échantillon vraisemblable de la loi Gamma considérée.

La figure **??** montre comme le test de KS que la racine carré des données de pluies n'est pas issue d'une loi Gamma. Mais le graphique montre aussi que le manque d'adéquation est surtout important pour les valeurs inférieures à 3. Il apporte ainsi une information plus précise que le test de KS qui se concentre sur le point où la distance entre la fonction de répatition théorique et la fonction de répartion empirique est maximum.

Test

De la même façon qu'on construit des intervalles de fluctuation par simulation on peut construire des tests d'hypothèses. Notons $\theta^{(\text{obs})}$ l'estimation de θ obtenue pour les observations.

1. `Simuler` B `échantillons de taille` n `selon la distribution` F
2. `Pour chaque échantillon` $b = 1, \cdots, B$, `calculer l'estimation empirique` $\hat{\theta}^{(b)}$
3. `Le degré de signification du test est donné par` $\min\{i$ `tel que` $\hat{\theta}^{(i)} > \theta^{(\text{obs})}\}$ `(ou` $\max\{i$ `tel que` $\hat{\theta}^{(i)} < \theta^{(\text{obs})}\}$ `selon le contexte.`

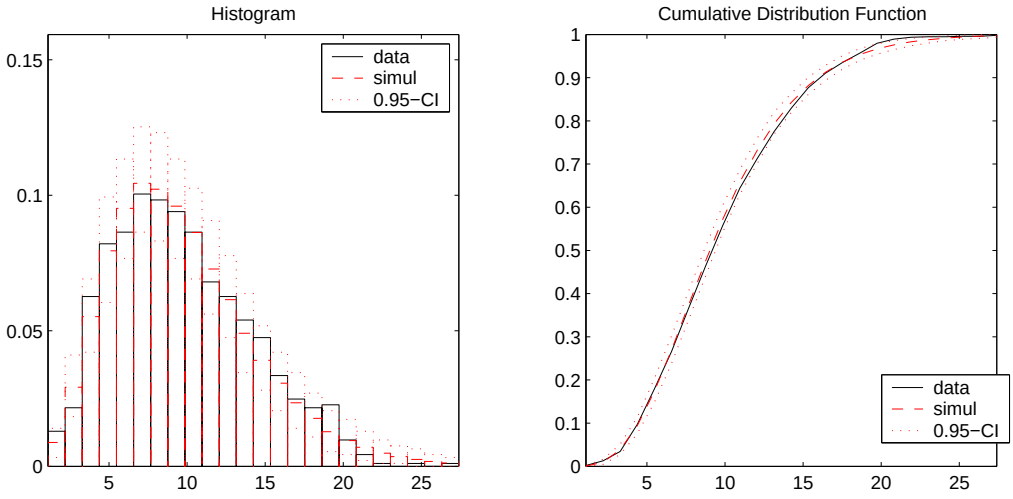

FIG. 3.4 – Exemples d'intervalles de fluctuation (pointillés rouges) pour l'histogramme (à gauche) et la fonction de répartition (à droite) d'une loi Gamma et comparaison aux statistiques (trait plein noir) des données de vent à Brest.

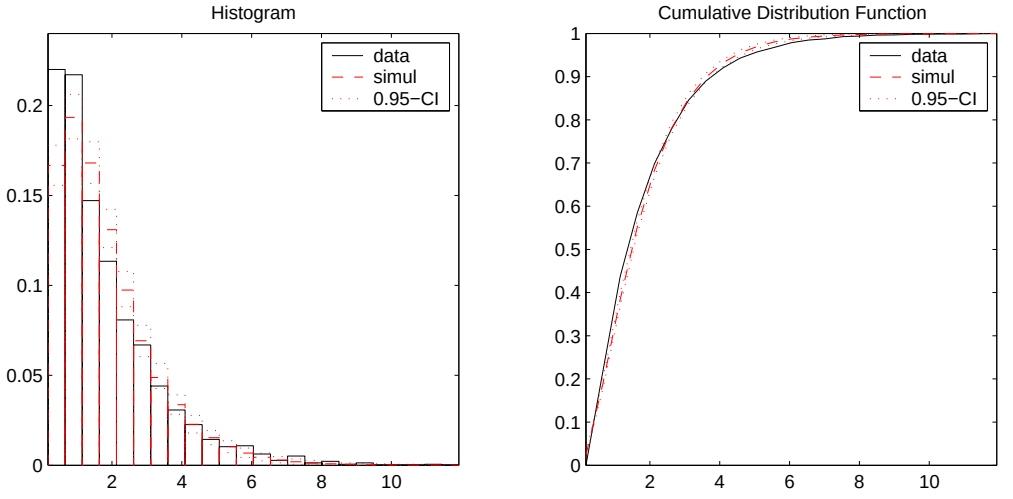

FIG. 3.5 – Exemples d'intervalles de fluctuation (pointillés rouges) pour l'histogramme (à gauche) et la fonction de répartition (à droite) d'une loi Gamma et comparaison aux statistiques (trait plein noir) des données de pluie au Vénézuela.

Chapitre 4

Théorie classique des valeurs extrêmes

4.1 Modèles asymptotiques

4.1.1 Formulation

Nous donnons ici le modèle de base pour la distribution des valeurs extrêmes. Ce modèle décrit la loi de

$$M_n = \max\{X_1, \cdots, X_n\}$$

où $X_1, \cdots, X_n$ est une suite de variables aléatoires indépendantes de même fonction de répartition F. En pratique M_n représente généralement un maximum sur une période d'observation donnée, par exemple un maximum annuel.

En théorie, la distribution de M_n est donnée pour toute valeur de n par

$$P(M_n \leq z) = P(X_1 \leq x, \cdots, X_n \leq x) = F^n(x) \tag{4.1}$$

Cependant, on ne peut pas utiliser cette formule directement car en pratique la fonction F est inconnue. Une solution est alors d'estimer F par une technique standard et de substituer l'estimateur dans (4.1). Mais une faible erreur d'estimation sur F peut conduire à une forte erreur sur F^n pour n grand...

Une approche alternative consiste à accepter que F est inconnue et à chercher à modéliser directement F^n à partir des valeurs extrêmes. Nous étudions donc le comportement de F^n quand n tend vers l'infini. Mais il faut remarquer que pour tout $z < z^+$, où z^+ est la borne supérieure du domaine de définition des X_i, on a

$$0 \leq F(z) < 1$$

et ainsi $F^n(z)$ tend vers 0 quand n tend vers l'infini et la distribution de M_n dégénère en un "point de masse" en z^+.

Cette difficulté est résolue en appliquant une renormalisation linéaire à M_n :

$$M_n^* = \frac{M_n - b_n}{a_n}$$

où $\{a_n > 0\}$ et $\{b_n\}$ sont des suites de constantes. Ces constantes permettent de stabiliser la localisation (mode) et l'échelle (variance) de M_n^*. On cherche alors une distribution limite pour M_n^* avec

des choix approppiés des suites $\{a_n\}$ et $\{b_n\}$.

Remarque - on utilise le même type de procédé quand on écrit un théorème de limite centrale pour un estimateur dont la variance dépend de la taille de l'échantillon.

4.1.2 Théorème des lois d'extrêmes

Le théorème des lois d'extrêmes donne les lois limites possibles pour M_n^*.

Théorème 1 *Théorème de Fisher-Tippett - Soit $X_1, \cdots, X_n$ une séquence de variables aléatoires indépendantes et identiquement distributées de fonction de répartition F et $M_n = \max_{i=1,\cdots,n} X_i$. S'il existe des suites de constantes $\{a_n > 0\}$ et $\{b_n\}$ telles que*

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \text{ quand } n \to \infty$$

où G est une fonction de répartition non dégénérée, alors G appartient à une des trois fammilles suivantes :

$$\text{I - Gumbel : } G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \quad -\infty < z < \infty$$

$$\text{II - Fréchet : } G(z) = \begin{cases} 0, & z \leq b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b \end{cases}$$

$$\text{III - Weibull : } G(z) = \begin{cases} \exp\left\{-\left(-\frac{z-b}{a}\right)^{\alpha}\right\}, & z < b \\ 1, & z \geq b \end{cases}$$

avec des paramètres $a > 0$,b et dans le cas des familles II et III, $\alpha > 0$.

Le théorème dit que le maximum normalisé $(M_n - b_n)/a_n$ converge en loi vers une variable de loi non dégénérée, alors cette variable a sa distribution dans une des trois familles. Il n'y a pas d'autre limite non dégénérée possible pour M_n^*. Et en ce sens le théorème 4.1.2 forme un analogue au théorème de limite centrale.

On appelle les distributions I, II et III, **distributions des valeurs extrêmes**[1]. Le paramètre b est un paramètre de localisation, a un paramètre d'échelle et α un paramètre de forme.

Historiquement, la distribution de Gumbel a été utilisée par Gumbel en 1954 pour modéliser des crues. En 1960 Thom a utilisé les distributions de Gumbel et de fréchet pour modéliser des vents extrêmes.

Exercice 1 *Montrer que la distribution du maximum d'un grand nombre N de variables aléatoires de distribution exponentielle $F(x) = 1 - \exp(-x), x > 0$ a approximativement la fonction de répartition*

$$\exp\left(-\exp(-(x - log(N)))\right)$$

Voir aussi figure 4.2.

Indication : choisir $a_n = 1$ et $b_n = n$ (ou remarquer que $ne^x = e^{\log(n)}e^x$).

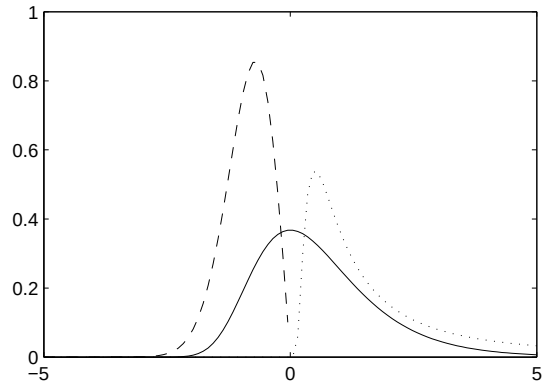

FIG. 4.1 – Densité de probabilité des distributions de valeurs extrêmes. Gumbel (trait plein), Fréchet de paramètre 1 (pointillés), Weibull de paramètre 2 (tirets)

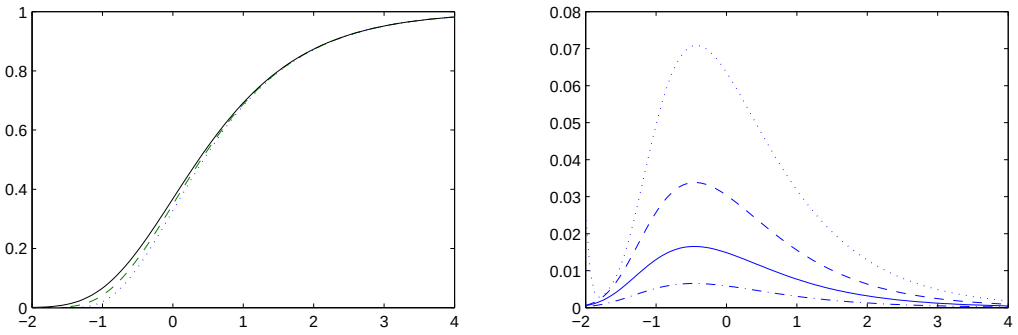

FIG. 4.2 –

Exercice 2 *Supposons que les paramètres a et b sont respectivement égaux à 1 et 0. Montrer que les assertions suivantes sont équivalentes :*

1. *X suit une loi de Fréchet de paramètre α*
2. *$log(X)^\alpha$ suit une loi de Gumbel*
3. *$-X^{-1}$ suit une loi de Weibull*

Remarques

1. La loi d'attraction G est unique à une transformation affine près. C'est à dire que si $\frac{M_n - b_n}{a_n}$ converge en loi vers une v.a. $Z_{a,b}$ de loi $H(ax+b)$ alors Z de loi $H(x)$ est la limite de $\frac{M_n - \tilde{b}_n}{\tilde{a}_n}$ avec $\tilde{a}_n = a_n/a$ et $\tilde{b}_n = b_n/b$.
2. Il existe une loi de Weibull (non extrême) dont la fonction de répartition est $F(x;a,b) = 1 - \exp(-(x/a)^b)$. Sa loi d'attraction est la loi de Gumbel.
3. Si X suit une des trois distributions du théorème de Fisher-Tippett, on dit qu'elle a une loi *max-stable* c'est à dire que M_n a même distribution que $a_n X + b_n$ ce qu'on peut formaliser de la façon suivante :

$$G^n(a_n z + b_n) = G(z)$$

[1] En anglais : extreme value distributions

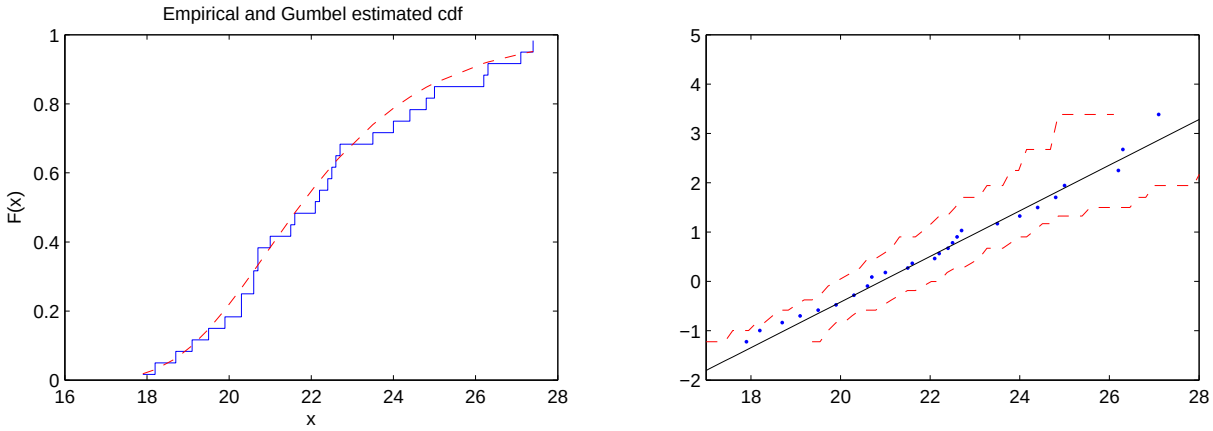

FIG. 4.3 –

On peut montrer que

$$\begin{aligned} Gumbel: &\quad b_n = log(n), a_n = n^{1/\alpha} \\ Frchet: &\quad b_n = 0, a_n = n^{1/\alpha} \\ Weibull: &\quad b_n = 0, a_n = n^{-1/\alpha} \end{aligned}$$

Les trois distributions du théorème de Fisher-Tippett sont les seules distributions max-stable.

4.1.3 GEV

Les trois types de distribution du théorème de Fisher-Tippett ont des comportements distincts (voir figure 4.1), correspondants aux formes des queues des lois F des variables X_i. Par exemple, on note que pour la distribution Weibull z^+ est finie alors que pour les deux autres distributions z^+ est infini. De plus, la distribution de Gumbel décroit de façon exponentielle alors que la distribution de Fréchet décroit à une vitesse polynomiale. Cependant, en pratique, il n'est pas toujours facile de choisir parmi ces trois distributions et, le choix induit des résultats spécifiques pour lesquels on ne tient pas compte du fait qu'il peut y avoir une incertitude sur le choix de la loi d'extrême.

On vérifie facilement, que les trois distributions d'attraction peuvent être combinées sous la forme d'un modèle unique : la famille de distributions **Generalized Extreme Value (GEV) distribution**

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \text{ avec } 1+\xi(z-\mu)/\sigma > 0, \sigma > 0 \tag{4.2}$$

Quand $k = 0$ la distribution GEV correspond à un modèle de Gumbel. Si $\xi < 0$, elle correpond à Weibull et si $\xi > 0$ à Fréchet.

Théorème 2 *Soit $X_1, \cdots, X_n$ une séquence de variables aléatoires indépendantes et identiquement distribuées de fonction de répartition F et $M_n = \max_{i=1,\cdots,n} X_i$. S'il existe des suites de constantes*

$\{a_n > 0\}$ et $\{b_n\}$ telles que

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z) \text{ quand } n \to \infty$$

où G est une fonction de répartition non dégénérée, alors G appartient à la famille GEV définie par l'équation (4.2).

On montre que sous les hypothèses du théoreme 2, on a équivalence entre les assertions 1 et 2 :

1. $P((M_n - b_n)/a_n \leq z) = G_n(z) \to G(z)$
2. $P(M_n \leq z) = G_n((z - b_n)/a_n) \to G*(z)$

où G et G^* sont deux membres de la famille GEV. En conséquence, on peut s'affranchir en pratique d'identifier les suites a_n et b_n.

4.2 Quantiles extrêmes et période de retour

En pratique, les lois d'extrèmes sont utilisées pour estimer des périodes de retour.

Soit un échantillon de variables aléatoires indépendantes $\{X_1, \cdots, X_n\}$. On forme des blocs d'observations de longueur n, avec n grand et on en déduit une suite de maxima $M_{n,1}, \cdots, M_{n,m}$ où $M_{n,i}$ est le maximum du bloc i. Le plus souvent les blocs sont choisis de façon à correspondre à une période temps. Par exemple, un an. Les **quantiles extrêmes** des maximas des blocs sont obtenus en inversant la fonction de répartition extrême G :

$$z_p = G^{-1}(1 - 1/p) \tag{4.3}$$

On dit que z_p est le **niveau de retour**[2] associée à la **période de retour**[3] p. On peut considérer qu'en espérance, z_p est dépassé en moyenne une fois sur la période p. Plus précisement, z_p est dépassé par le maximum de bloc pour tout bloc particulier avec la probabilité $1/p$.

Considérons l'exemple des vitesses de vent à Brest et estimons une valeur de retour à 100 ans. C'est la valeur qu'on s'attend à observer en moyenne une seule fois sur une période de 100 ans. Dans cet exemple le bloc est l'année, on travaille avec des maxima annuels. Et, la période considérée $p = 100$. La valeur de retour est donc z_{100} définie par

$$G(z_{100}) = P(Z < z_{100}) = 1 - 1/100$$

soit encore

$$z_{100} = G^{-1}(1 - 1/100)$$

Si le modèle est construit à partir de données observées N fois par an et non plus les maxima annuels, on obtient

$$z_{100} = G^{-1}(1 - 1/(N * 100))$$

L'inverse de la fonction de répartition G^{-1} est aussi appelée *fonction quantile*.

[2] En anglais : return level
[3] En anglais : return period

Exercice 3 *Estimer les valeurs de retour à 50 ans, 100 ans et 1000 ans de maximas annuels distribués suivent une loi de Gumbel de paramètre 1.*

Dans le cas de l'équation (4.2), on a par exemple

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{\log(1-\frac{1}{p})\}^{-\xi}\right] & \text{si } \xi \neq 0 \\ \mu - \sigma \log\{-\log(1-\frac{1}{p})\} & \text{si } \xi = 0 \end{cases} \tag{4.4}$$

L'équation 4.4 permet de remarquer que si on trace z_p en fonction de $-\log(1-1/p)$ on obtient une droite si $\xi \neq 0$. On appelle ce graphe **tracé en niveau de retour**[4].

Exercice 4 *En fonction de quelle quantité peut-on tracer z_p pour obtenir une droite dans le cas ou $\xi = 0$. Que devient cette représentation si les données sont issues d'une loi de d'extrême de Weibull (resp. de Fréchet) ?*

4.3 Inférence pour les lois d'extrême

4.3.1 Généralités

On a vu dans la section précédente que les lois d'extrêmes sont en général utilisées pour estimer la distribution des maxima de blocs. En pratique, le choix de la taille des blocs est critique. Si les blocs sont trop petits, l'approximation par les lois limites peut être mauvaise ce qui induit du biais dans les estimations. Si les blocs sont grands, on observe peu de maxima et on induit de la variance dans les estimations.

Des considérations pratiques amènent souvent à retenir des blocs de un an. Ce choix a plusieurs avantages. On remarque en particulier que

- les maxima annuels peuvent être considérés comme des observations issues de variables aléatoires indépendantes ;
- les maxima annuels peuvent être considérés comme des observations issues de variables aléatoires identiquement distribuées. En effet, les effets de saison ne sont pas visibles à cette échelle.

De nombreuses méthodes ont été proposées pour estimer les paramètres des lois d'extrêmes :

- méthodes graphiques
- méthodes des moments (et leurs généralisations)
- méthodes basées sur les statistiques d'ordre
- méthodes basées sur le maximum de vraisemblance
- méthodes mixtes

Chaque technique a ses défenseurs et ses détracteurs. La méthode du maximum de vraisemblance est traditionnellement utilisée (et implémentée).

Exercice 5 *Soit $Z_1, \cdots, Z_n$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant une loi GEV de paramètre (μ, σ, ξ). Ecrire la log vraisemblance associée. Distinguer les cas $\xi \neq 0$ et $\xi = 0$.*

[4] return level plot

Cepedant il y a des arguments qui vont à l'encontre de ce choix. Pour les lois de Weibull, Fréchet et pour la GEV, le domaine de défnition dépend des paramètres. En conséquence les propriétés usuelles des estimateurs du maximum de vraisemblance ne sont plus vérifiées. En particulier, on peut montrer que (Smith, 1995)

- si $\xi > -.5$, les estimateurs du maximum de vraisemblance des paramètres de la GEV ont les propriétés usuelles ;
- si $-1 < \xi < -0.5$ on peut calculer les estimateurs du maximum de vraisemblance, mais on perd les propriétés asymptotiques ;
- si $\xi < -1$ on ne peut pas obtenir les estimateurs du maximum de vraisemblance.

Le cas $\xi \leq -0.5$ correspond à des distributions dont le domaine de définition est borné supérieurement par des valeurs assez faibles. Ce cas arrive assez peu en pratique. On peut vérifier que pour des valeurs positives de ξ les estimateurs du maximum de vraisemblance ont une variance importante qui induit une forte imprécision sur les estimateurs des valeurs de retour. Dans ce cas, on préfère souvent utiliser des méthodes basées sur les L-moments qui induisent pourtant un biais ou des méthodes mixtes.

Exercice 6 *Exercice à faire en TD - Simuler B échantillons de taille $n = 30$ selon la loi GEV de paramètre $\theta = (0, 1, \xi = 0.3)$. Estimer θ par la méthode du maximum de vraisemblance pour chaque échantillon et en déduire une estimation empirique de la variance de θ.*

4.3.2 Méthode des Lmoments

La méthode des L-moments est similaire à la méthode des moments mais elle repose sur les moments basés sur les statistiques d'ordre. Dans un cadre plus général, les L-statistiques sont des combinaisons linéaires des statistiques d'ordre.

On vérifie facilement que, pour un échantillon $X_1, \cdots, X_n$, la statistique d'ordre $X_{[nq]:n}$ est un estimateur du quantile d'ordre q. Si on considère une famille de distributions paramétrée par une moyenne μ et un écart-type σ, on a

$$X_{[nq]:n} \sim \mu + \sigma F^{-1}(q)$$

. Cette relation peut être utilisée pour estimer μ et σ.

L-moments

Définition 1 *Soit X une v.a. définie sur un espace E et de fonction de répartition F les moments pondérés (probability weighted moments) sont donnés par*

$$\beta_r = \int_E x(F(x))^r dF(x)$$

Les L-moments sont construits à partir des moments pondérés. Ces moments ont des avantages théoriques sur les moments ordinaires :

- Pour que les L-moments d'une distribution de probabilité soient définis, il suffit que la distribution ait une moyenne finie [J. R. M. Hosking, J. R. Statist. Soc. B, 52 (1990), Theorem 1].

- Pour que la variance des estimateurs empiriques des L-moments soit finie, il suffit que la distribution ait une variance finie [Hosking, 1990, Theorem 3].
- Bien que les rapport des L-moments puissent être arbitriarement grands, les rapports des L-moment empiriques sont bornés [J. Dalen, Statistics and Probability Letters, 5 (1987)] ; sample L-moment ratios can take any values that the corresponding population quantities can [Hosking, 1990, page 115].

D'autre part, on a les priopriétés suivantes dans un grand nombre de situations pratiques :

- Les approximation asypptotic des distributions empiriues sont meilleurs pour les L-moments que pour les moments ordianires [Hosking, 1990, Figure 4].
- Les L-moments sont moins sensibles aux outliers [P. Royston, Statistics in Medicine, 11 (1992), Figure 7 ; R. M. Vogel and N. M. Fennessey, Water Resources Research, 29 (1993), Figures 3 and 4].
- Les L-moments permettent d'obtenir une meilleure identification de la distribution dont est issu l'échantillon. [Hosking, 1990, Figure 6]. Voir le diagramme des L-moments (http ://www.research.ibm.com/p

On montre que

$$\lambda_1 = \beta_0, \quad \lambda_2 = 2\beta_1 - \beta_0, \quad \lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0$$

Le moment λ_1 est un paramètre de localisation, λ_2 est un paramètre de dispersion et λ_3 est un paramètre de symétrie.

Exercice 7 *1. Calculer les trois premiers L-moments pour la loi uniforme.*

2. Calculer les trois premiers L-moments pour la loi GEV.

Pour la GEV,

$$\beta_r = \int_0^1 [\mu + \sigma(1 - 1 - (-\log F)^{\xi}\}/\xi] f F^r dF$$

On déduit facilement les estimateurs empiriques des moments pondérés. Soit un échantillon $X_1, \cdots, X_n$ de v.a.i.i.d.

$$\begin{aligned}
\hat{\beta}_0 &= \bar{X} \\
\hat{\beta}_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i \frac{\text{card}(X \leq X_i)}{n} \\
\hat{\beta}_r &= \frac{1}{n}\sum_{j=r+1}^{n} \frac{(j-1)(j-2)\cdots(j-r)}{(n-1)(n-2)\cdots(n-r)} X_{(j)}
\end{aligned}$$

où $X_{(j)}$ est la $j$ème statistique d'ordre.

4.3.3 Méthodes mixtes

Dans les méthodes mixtes, on calcule les estimateurs du maximum de vraisemblance mais sous des contraintes de moments. En pratique, on substitue à un ou deux paramètres leur expression en fonction des L-moments et on maximise la vraisemblance comme une fonction des paramètres non remplacés.

Méthode MIX1

$$\text{Maximiser } \log L(\theta|x)$$
$$\text{sous les contraintes } \begin{cases} \mu = \hat{\lambda}_1 + \frac{\sigma}{\xi}[1-\Gamma(1-\xi)] \\ \xi(\mu - x_i) \leq \sigma \quad i = 1, \cdots, n \\ \sigma > 0 \end{cases}$$

avec $\hat{\lambda}_1$ l'estimateur empirique du premier L-moment.

Méthode MIX2

$$\text{Maximiser } \log L(\theta|x)$$
$$\text{sous les contraintes } \begin{cases} \mu = \hat{\lambda}_1 + \frac{\sigma}{\xi}[1-\Gamma(1-\xi)] \\ \sigma = \frac{\hat{\lambda}_2 \xi}{(2^\xi - 1 - 1)\Gamma(1-\xi)} k(\mu - x_i) \leq \sigma \quad i = 1, \cdots, n \end{cases}$$

avec $\hat{\lambda}_1$ et $\hat{\lambda}_2$ les estimateurs empiriques des 2 premiers LMOM.

On montre par simulation que pour les valeurs positives de ξ (Morrison and Smith),

1. Les méthodes mixtes produisent de meilleurs estimateurs des paramètres de la GEV que le MV ou LMOM
2. Les estimateurs des quantiles donnés par les méthodes mixtes ont une erreur en moyenne quadratique proche de celle de LMOM
3. Les méthodes mixtes préservent les propriétés asymptotiques des estimateurs du maximum de vraisemblance ; on peut montrer que ces estimateurs sont consistents sous certaines hypothèses de régularité et que leur loi linmite est gausienne.

4.3.4 Inférence pour les valeurs de retour

En substituant les estimations des paramètres de la GEV dans l'expression de la valeur de retour, on obtient une estimation de la valeur de retour.

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - \{\log(1-\frac{1}{p})\}^{-\hat{\xi}}\right] & \text{si } \xi \neq 0 \\ \hat{\mu} - \hat{\sigma}\log\{-log(1-\frac{1}{p})\} & \text{si } \hat{\xi} = 0 \end{cases}$$

De plus, on caractérise la variance de l'estimateur de la valeur de retour. On utilise les mêmes arguments que dans la Delta method.

La delta method est une approche générale permettant de construire des intervalles de confiance pour des fonctions d'estimatuer du maximum de vraisemblance. On construit une approximation linéaire de la fonction basée sur un développement de Taylor et on approche la variance de la fonction par la variance de l'approximation linéaire.

Soit $\hat{\theta}$ l'estimateur du maximum de vraisemblance d'un paramètre θ et h une fonction régulière. Alors on peut montrer que

$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) \to \mathcal{N}\left(0, \nabla h(\hat{\theta})' Var(\hat{\theta}) \nabla h(\hat{\theta})\right)$$

$$h(\hat{\theta}) - h(\theta) = (\hat{\theta} - \theta)\nabla h(\hat{\theta}) + \text{reste}$$

d'où

$$\begin{aligned} Var\left(h(\hat{\theta}) - h(\theta)\right) &= E\left((h(\hat{\theta}) - h(\theta)))'(h(\hat{\theta}) - h(\theta)))\right) + \text{reste} \\ &= E(\nabla h(\hat{\theta})'(\hat{\theta} - \theta)'(\hat{\theta} - \theta)\nabla h(\hat{\theta}) + \text{reste} \\ &\approx \nabla h(\hat{\theta})' Var(\hat{\theta} - \theta)\nabla h(\hat{\theta}) \end{aligned}$$

Pour les valeurs de retour, on a

$$\nabla z_p' = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] = [1, -\xi^{-1}(1 - y_p^{-\xi}), \sigma\xi^{-2}(1 - y_p^{-\xi}) - \sigma\xi^{-1} y_p^{-\xi} log(y_p)]$$

avec $y^p = -log(1-p)$.

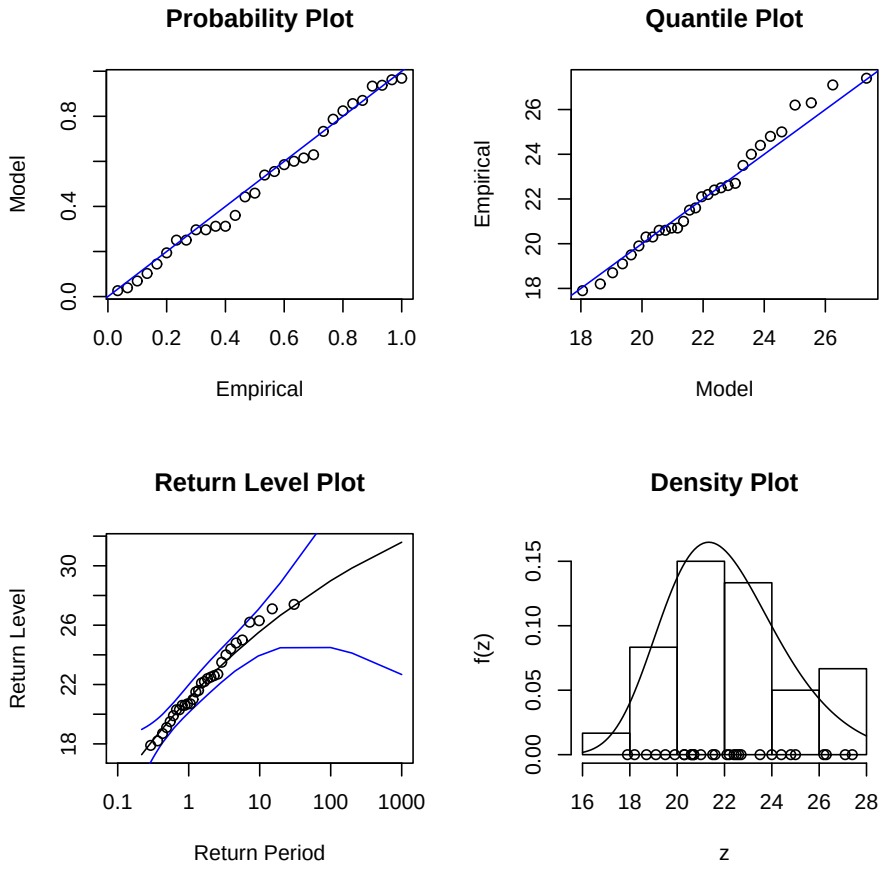

FIG. 4.4 – Valeur de retour du maximum annuel du vent à Brest - Modèle GEV

Chapitre 5

Modèles de franchissement

5.1 Introduction

Dans les méthodes basées sur les extremas par blocs (par exemple, les extremas annuels) on utilise une très faible partie des données disponibles. En particulier, si la série de données présente deux valeurs très élevées une même année, on ne va retenir que la plus forte alors que l'autre sera éventuellement plus forte que les extrema d'autres années. Quand on dispose de données horaires, journalières, mensuelles ... on peut proposer d'autres approches.

Soit $\{X_1, X_2, \cdots, X_n\}$ une suite de v.a.i.i.d. admettant une distribution marginale F. Il est naturel de considérer comme étant extrêmes les évènements X_i qui sont supérieurs à un seuil fixé u. Dans ce cas, on peut caractériser le comportement extrême d'un évenement X par la probabilité conditionnelle

$$P(X > u + y | X > u) = \frac{P(X > u + y)}{P(X > u)} = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0$$

Si on connait F on en déduit la distribution des dépassements de niveau. En pratique, on ne connait pas F pour des niveaux élevés et on l'approche par une GEV par exemple.

5.2 Modèle de Pareto Généralisé - GPD

Le modèle de Pareto généralisé permet de caractériser la distribution des dépassements de niveaux u élevés. On le justifie de la façon suivante.

Soit X une v.a. de fonction de répartition F. D'après le théorème sur la GEV, pour n assez grand, on a

$$F^n(z) \approx \exp\left\{ - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}$$

avec $\mu, \sigma > 0$ On peut encore écrire

$$n \log F(z) \approx - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi}$$

Et si z est grand, le développement de Taylor du log donne,

$$\log F(z) \approx -(1 - F(z))$$

et

$$1 - F(u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}$$

pour u grand. Suivant les mêmes arguments,

$$1 - F(u+y) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u+y-\mu}{\sigma}\right)\right]^{-1/\xi}$$

Ainsi,

$$P(X > u + y | X > u) \approx \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi}$$

avec $\tilde{\sigma} = \sigma + \xi(u - \mu)$. La famille de distribution caractérisée par la fonction $H(y) = \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi}$ est appelée famille de Pareto généralisée.

On a donc le théorème suivant.

Théorème 3 *Soit $X_1, \cdots, X_n$ une suite de variables aléatoires indépendantes de même fonction de répartition F, et soit $M_n = \max\{X_1, \cdots, X_n\}$. En notant X un terme arbitraire X_i de la suite et en supposant que F vérifie le théorème 2 alors, pour u grand, la distribution de $(X - u)$ conditionnellement à $X > u$ est approximativement*

$$H(y) = \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi}$$

définie sur $\{y : y > 0$ et $(1 + \xi y/\tilde{\sigma}) > 0\}$ avec $\tilde{\sigma} = \sigma + \xi(u - \mu)$.

Le théorème 3 implique que si les maxima par blocs ont une distribution GEV alors les dépassements d'un niveau u élevé ont approximativement une distribution de Pareto généralisée avec des paramètres que l'on déduit des paramètres de la GEV. En particulier les paramètres de forme sont égaux.

On remarque que si le paramètre de forme est nul, la fonction de répartition $H(y) = 1 - \exp(-y/\tilde{\sigma})$ correspond à une loi exponentielle.

Exercice 8 *Déterminer la distribution limite des dépassements d'un niveau u élevé quand $F(x) = 1 - \exp(x)$ - modèle exponentiel.*

5.3 Choix du niveau

Les données considérées sont une suite de réalisations indépendantes $\{x_1, \cdots, x_n\}$ de variables de mêmes lois $X_1, \cdots, X_n$. Les évènements extrêmes sont caractérisés par $\{x_i : x_i > u\}$ pour un un seuil u fixé. Notons $x_{(1)}, \cdots, x_{(k)}$ les réalisations associées aux évènements extrêmes et $y_j = x_{(j)} - u$ les dépassements pour $j = 1, \cdots, k$. D'après le théorème 3, les y_j peuvent être vues comme des réalisations indépendantes de variables aléatoires de même loi appartenant à la famille de Pareto généralisée.

Une des difficultés consiste à bien choisir le seuil u. En effet, si u est trop petit, on ne peut pas utiliser les résultats asymptotiques (ie l'approximation par une distribution de Pareto généralisée (GPD) ne sera pas bonne) et si u est trop grand, on conserve très peu d'observations. Voir exemple matlab.

5.3.1 Méthode basée sur la moyenne de la GPD

On peut montrer (voir Reiss & Thomas) que si Y suit une GPD de paramètres σ et ξ alors

$$E(Y) = \frac{\sigma}{1-\xi} \text{ si } \xi < 1$$

Si ξ est plus grand que 1 la moyenne est infinie. Ainsi, on a

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1-\xi}$$

Si le modèle est valide pour u_0 alors il est aussi valide pour tout u supérieur à u_0, ie $u > u_0$

$$E(X - u | X > u) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0+\xi u}}{1-\xi}$$

On remarque que pour $u > u_0$, $E(X - u|X > u)$ est une fonction linéaire de u.

Ainsi, on propose de tracer le nuage de points

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{\max} \right\}$$

Le graphique obtenu est appelé *mean residual life plot*. Et on cherche le seuil u à partir duquel le nuage de points évolue linéairement. On peut ajouter les intervalles de confiances en supposant que la loi de l'estimateur empirique de la moyenne est bien approchée par une loi de Gauss.

L'interprétation du graphique n'est pas toujours évidente. En effet, plus u est grand et plus la variabilité (variance !) est importante, on ne voit alors pas clairement l'évolution linéaire...

5.3.2 Méthode basée sur la stabilisation des paramètres

Une autre façon de procéder repose sur l'idée que si le modèle est valide pour u_0 alors il doit être valide aussi pour $u > u_0$, et les paramètres de la GPD doivent donc rester constants quand on fait croitre u.

On ajuste une GPD sur les dépassements pour plusieurs valeurs de u, et on cherche le seuil à partir duquel la valeur des paramètres ne varie plus.

5.4 Inférence

Remarque

5.4.1 Estimation par la méthode des moments

Estimateur de Hill - Méthode de référence. Non paramétric, basé sur les statistiques d'ordre (-> dépasse le cadre des GPD) Facile à mettre en oeuvre et asymptotiquement sans biais. Pb quand on travaille avec de petits échantillons... Pb pour le choix du seuil.

5.4.2 Estimation par maximum de vraisemblance

Optimisation numérique. Voir GEV.

Les estimateurs du maximum de vraisemblance de σ et ξ sont asymptotiquement de loi de Gauss de moyenne (σ, ξ) et de variance Σ/k où k est la taille de l'échantillon (nombre de valeurs dépassant u) et

$$\Sigma = (1+\xi)\begin{pmatrix} 1+\xi & \sigma \\ \sigma & 2\sigma^2 \end{pmatrix}$$

5.4.3 Estimation par la méthode de Pickands

L'estimateur de Pickands est basé sur un ajustement de la fonction de répartition théorique sur la fonction de répartition empirique. Cette idée est assez naturelle si on se rappelle que les valeurs de retour se déduisent des quantiles. Pour estimer les paramètres σ et k, on ajuste 2 quantiles.

$$F(x_{(n/2)}; \xi, \sigma) = \frac{n/2}{n+1}$$
$$F(x_{(3n/2)}; k, \sigma) = \frac{3n/2}{n+1}$$

où $F(x) = 1-(1-kx/\sigma)^{-1/k}$ représente la fonction de répartition de $X-u|X>u$. On en déduit facilement

$$\hat{\sigma} = \frac{x^2_{(n/2)}}{2x_{(n/2)} - x_{(3n/2)}},\ \hat{k} = \frac{-1}{\log(2)} \log\left(\frac{x^2_{(n/2)}}{2x_{(n/2)} - x_{(3n/2)}}\right)$$

Le principal avantage de l'estimateur de Pickands est qu'il est défini quelque soit la valeur de k. On montre d'autre part que l'estimateur $(\hat{\sigma}, \hat{k})$ converge en loi vers une variable aléatoire de loi de Gauss centrée en (σ, k) et dont on peut calculer la variance.

5.5 Valeur de retour

Comme nous l'avons montré plus haut, il est plus facile d'interpréter un modèle d'extrême à partir des valeurs de retour. Supposons que le modèle GPD soit bien adapté pour modéliser les dépassements de u par une variable X. C'est à dire que si $X > u$,

$$P(X > x|X > u) = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$$

On a

$$P(X > x) = \zeta_u \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}$$

avec $\zeta_u = P(X > u)$ (Bayes) et si x_m est le niveau qui est dépassé en moyenne une fois toutes les m observations

$$\zeta_u \left[1 + \xi\left(\frac{x_m-u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}$$

Finalement

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right], \xi \neq 0$$

si m est assez grand pour que $x_m > u$.

En général, on préfère parler de valeur de retour à N ans. Si on a conservé en moyenne $n_a(u)$ observations par an, ça correspond à la valeur de retour x_m avec $m = N * n_a(u)$.

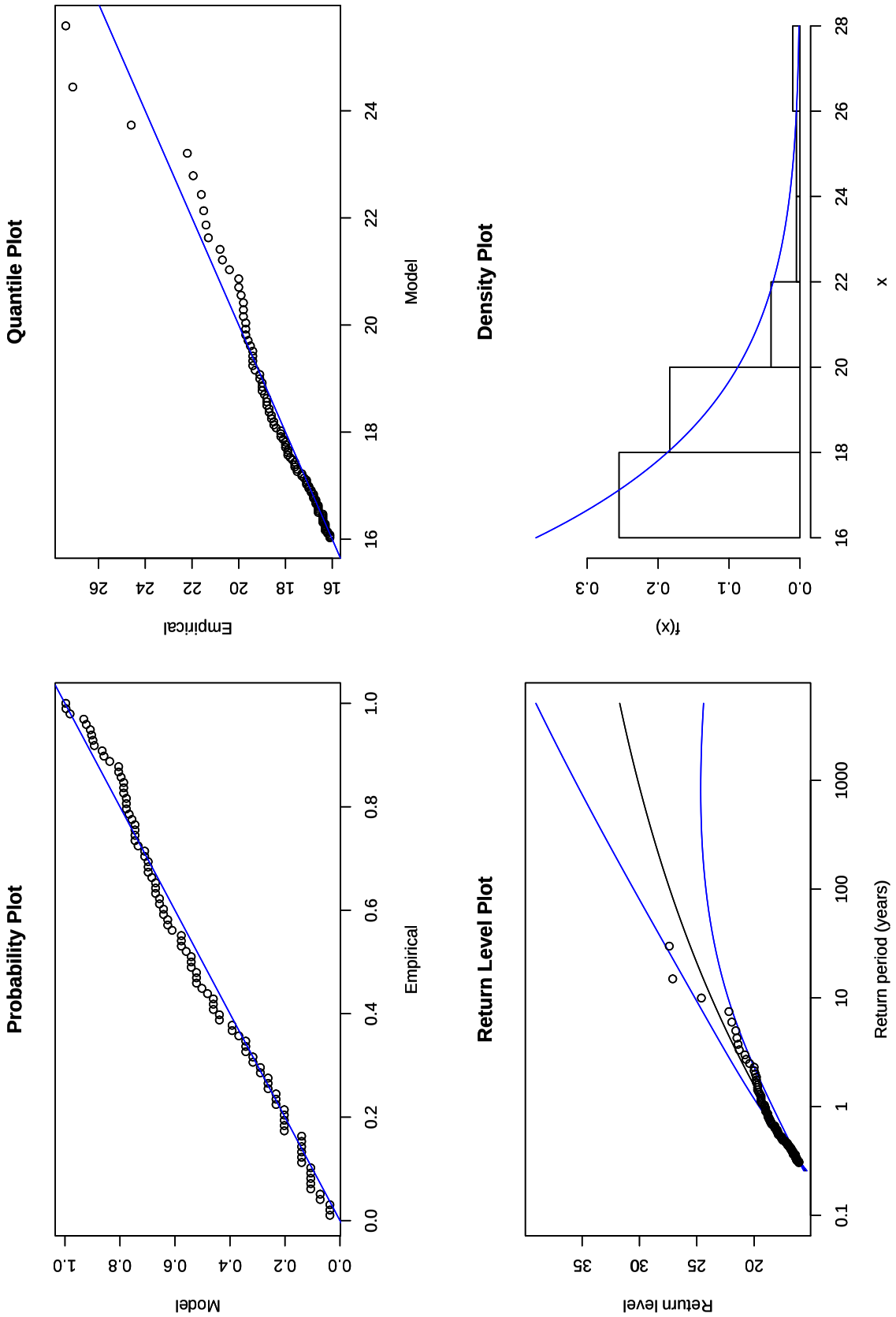

FIG. 5.1 – Valeur de retour du maximum annuel du vent à Brest ua mois de janvier - Modèle GPD

Chapitre 6

Modélisation des extrêmes pour les séries temporelles

Dans les méthodes que nous avons étudiées jusqu'à présent nous faisons l'hypothèse que les observations sont des réalisations de variables indépendantes de même loi. Dans la théorie classique des valeurs extrèmes (GEV), on considère en général des extrémas annuels et l'hypothèse est vérifiée la plupart du temps même si les observations sont des réalisations d'une suite de variables aléatoires dépendantes telle qu'une série météorologique ou plus généralement des données environnementales. Dans le cas où l'on étudie les extrêmes d'une série temporelle, les observations sont dépendantes et on ne peut pas utiliser le modèle GPD.

Dans la suite, nous considérons des séries temporelles qui sont des réalisations de processus stationnaires.

6.1 Indépendance asymptotique

Pour construire des modèles d'extrême pour des processus, nous avons besoin d'introduire une notion d'indépendance asymptotique.

Définition 2 *Soit une série $X_1, X_2, \cdots$ de variables aléatoires identiquement distribuées. On dit que cette série vérifie la condition $D(u_n)$ si pour tout $i_1 < i_2 < \cdots < i_p < j_1 < \cdots < j_q$ avec $j_1 - i_p > l$*

$$\begin{aligned}&\big|P(X_{i_1} \leq u_n, \cdots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \cdots, X_{j_q} \leq u_n)\\ &P(X_{i_1} \leq u_n, \cdots, X_{i_p} \leq u_n)P(X_{j_1} \leq u_n, \cdots, X_{j_q} \leq u_n)\big| \leq \alpha(n,l)\end{aligned}$$

où $\alpha(n,l) \to 0$ pour une suite l_n telle que l_n/n tend vers 0 et n tend vers l'infini.

Cette condition implique que si deux valeurs sont assez éloignées dans le temps alors elle sont indépendantes. Elle est vérifiée en particulier par les processus autorégressifs. De plus, elle est presque toujours réaliste pour les processus environnementaux.

Théorème 4 *Soit $\{X_i\}_{i\in I}$ un processus stationnaire et soit $M_n = \max\{X_1, \cdots, X_n\}$. Alors si $\{\alpha_n > 0\}$ et $\{b_n\}$ sont des suites de constantes telles que*

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G(z)$$

où G est une fonction de distribution non dégénérée et si $D(u_n)$ est vérifiée pour $u_n = a_n z + b_n$ alors G appartient à la famille GEV.

Ce théorème implique que la distribution des extrêmes d'une série dépendante tend vers la même famille de distribution que celle des extrêmes de v.a. indépendantes. Cependant, on va montrer que les paramètres de la loi sont affectés par la dépendance.

6.2 Extremal index

Considérons tout d'abord un exemple : soit $Y_0, Y_1, Y_2, \cdots$ une suite de v.a.i.i.d. de distribution exponentielle ayant pour fonction de répartition

$$F_Y(y) = \exp\left\{\frac{-1}{(a+1)y}\right\} \text{ pour } y > 0$$

où $a \in [0,1]$ est un paramètre. On définit le processus $\{X\}$ par

$$\begin{cases} X_0 = Y_0 \\ X_i = \max\{aY_{i-1}, Y_i\},\ i = 1, \cdots, n \end{cases}$$

pour tout $i \in 1, \cdots, n$ et $x > 0$ on a

$$\begin{aligned} P(X_i \leq x) &= P(aY_{i-1} \leq x, Y_i \leq x) \\ &= P\left(Y_{i-1} \leq \frac{x}{a}\right) P(Y_i \leq x) \\ &= \exp\left\{\frac{-a}{(a+1)x} - \frac{1}{(a+1)x}\right\} \\ &= \exp\left\{-\frac{1}{x}\right\} \end{aligned}$$

On en déduit donc que le processus stationnaire X a une distribution marginale de Fréchet.

Considérons maintenant $X_1^*, X_2^*, \cdots$ une suite de va indépendantes telle que pour tout i

$$P(X_i^* \leq x) = \exp\left\{-\frac{1}{x}\right\}$$

On définit

$$M_n^* = \max\{X_1^*, \cdots, X_n^*\}$$

et on a

$$P(M_n^* \leq nz) = \left(\exp\left\{-\frac{1}{nz}\right\}\right)^n = \exp\left(-\frac{1}{x}\right)$$

Maintenant, si $M_n = \max\{X_1, \cdots, X_n\}$ on obtient

$$\begin{aligned} P(M_n \leq nz) &= P(X_1 \leq nz, \cdots, X_n \leq nz) \\ &= P(Y_1 \leq nz, aY_1 \leq nz, \cdots, Y_{n-1} \leq nz, aY_{n-1} \leq nz, Y_n \leq nz) \\ &= P(Y_1 \leq nz, Y_2 \leq nz, \cdots, Y_n \leq nz) \text{ car } a < 1 \\ &= \left(\exp\left(\frac{-1}{(a+1)nz}\right)\right)^n \\ &= \left(\exp\left(\frac{-1}{z}\right)\right)^{1/(a+1)} \\ &= P(M_n^* \leq nz)^{1/(a+1)} \end{aligned}$$

Exercice 9 *A faire en TD : réaliser des simulation qui mettent en évidence que la dépendance induit une modification du conportement des extrèmes.*

Le théorème suivant permet de généraliser ce qui est mis en évidence dans l'exemple précédent.

Théorème 5 *Soit $\{X\}$ un processus stationnaire et $X_1^*, X_2^*, \cdots$ une suite de va indépendantes de même loi marginale que X_i. On note $M_n = \max\{X_1, \cdots, X_n\}$ et $M_n^* = \max\{X_1^*, \cdots, X_n^*\}$. Sous les conditions de régularité adéquates,*

$$P\left(\frac{M_n^* - b_n}{a_n} \leq z\right) \to G_1(z) \text{ quand } n \to +\infty$$

pour des suites $\{a_n > 0\}$ et $\{b_n\}$. G_1 est une fonction de répartition non dégénérée si et seulement si

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \to G_2(z) \text{ quand } n \to +\infty$$

où $g_2(z) = G_1(z)^\theta$ pour une constante θ telle que $0 < \theta \leq 1$.

Ce théoreme implique que si le maximum d'une suite stationnaure converge, la distribution limite est reliée à la distribution du maimum d'une suite de variable aléatoires indépendantes de même loi marginale. L'effet de la dépendance est de remplacer la loi limite par une puissance θ inférieure à un de cette loi limite. Le paramètre θ est appelé **index extrêmal**[1].

On peut interpreter l'index extrêmal d'une suite stationnaire d'après sa propension à former des clusters dans les extrêmes, et on a

$$\theta = (\text{limiting mean cluster size})^{-1}$$

Dans le cas de l'indépendance, on a $\theta = 1$.

Estimation de θ

L'approche paramétrique consiste à déduire un estimateur de la relation suivante

$$P(M_{\text{an}} < z) = P(M_{\text{bloc}} < z)^{n\theta}$$

avec n le nombre d'observations par an.

[1] en anglais : extremal index

Certains auteurs proposent des estimateurs non paramétriques. On a par exemple un estimateur basé sur une équation des moments. En posant pour un niveau u choisi

$$\hat{\theta}(u) = \begin{cases} \min(1, \hat{\theta}_n(u) \text{ si } \max\{T_i : 1 \leq i \leq n-1\} \leq 2 \\ \min(1, \hat{\theta}^*_n(u) \text{ si } \max\{T_i : 1 \leq i \leq n-1\} > 2 \end{cases}$$

avec T_i le temps écoulé entre 2 franchissements du niveau u et

$$\hat{\theta}_n(u) = \frac{2\left(\sum_{i=1}^{n-1} T_i\right)^2}{(n-1)\sum_{i=1}^{n-1} T_i^2}$$

$$\hat{\theta}^*_n(u) = \frac{2\left(\sum_{i=1}^{n-1} (T_i - 1)\right)^2}{(n-1)\sum_{i=1}^{n-1} (T_i - 1)(T_i - 2)}$$

6.3 Modèle pour les maxima par blocs

Pour modéliser les extrêmes de séries de variables dépendantes, on raisonne classiquement de la façon suivante : on détermine tout d'abord un niveau u (ou une statistique d'ordre) au dessus duquel on considére les observations (les observation plus faibles sont négligées) et on sélectionne le maximum de chaque cluster de valeurs supérieures à u. Si les blocs sont de taille suffisante, on peut faire l'hypothèse que les extremas par bloc sont indépendants et on se retrouve donc dans la situation des chapitres précédents. On ajuste alors un modèle d'extrême (GEV) sur les maxima par blocs.

Néammoins, il faut faire attention à la validité des extrapolation obtenues par cette méthode. L'approche basique consiste à dire que la distribution estimée est une apporximation valide si n est très grand. POur les séries stationnaire, nous avons vu que que M_n a les mêmes propriétés statistiques que $M^*_{n\theta}$. Cependant, il faut prendre en compte le fait qu'en réduisant le nombre d'observations de n à $n\theta$, on diminue la qualité des estimateurs !

6.4 Modèles à seuil

De même que le modèle GEV est valide pour les extrema par blocs de séries stationnaires, les mêmes arguments suggèrent que la distribution GPD reste appropriée pour les dépassements de niveaux. Cependant, dans les séries stationnaires les extrêmes ont tendance a former des clusters. Ceci nécessite d'adapter l'usage de ce modèle. En particulier, l'approche développée au chapitre précédent permet de modéliser un dépassement du niveau u considéré mais ne donne aucune information sur la dépendance avec les observations voisines (dans le temps). Par exemple si on considère des données de température extrême, on observe en général qu'un jour de très forte chaleur est suivi et/ou précédé de fortes températures : dépendance ! ! ! On ne peut alors plus utiliser les estimateurs du maximum de vraisemblance pour estimer les paramètres de la GPD.

Une alternative consiste à ne retenir que les maxima par bloc pour "dégrouper" les observations... Mais ceci engendre une perte d'information....

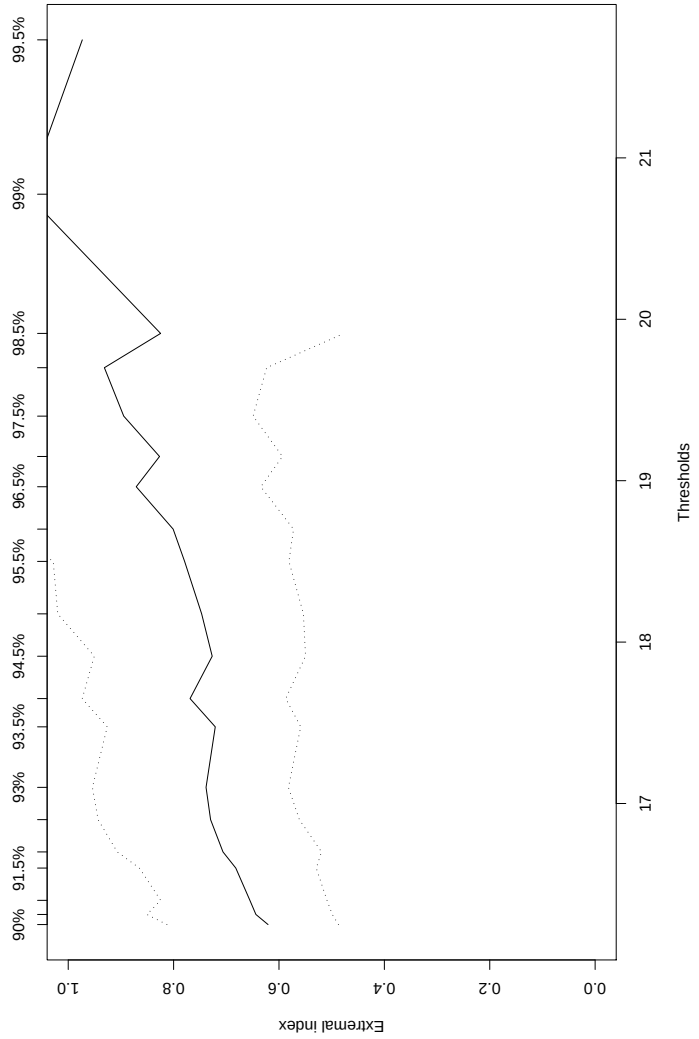

FIG. 6.1 – Variation de l'extremal index pour la série journalière de vent à Brest au mois de janvier. Les intervalles de confiance sont estimés par bootstrap.