Estimation des probabilités et quantiles extrêmes

Stéphane Girard

INRIA Rhône-Alpes, projet Mistis
http ://mistis.inrialpes.fr/~girard/dissemination.html

janvier 2009

Estimation des probabilités et quantiles extrêmes

1. En utilisant le théorème des valeurs extrêmes
2. En utilisant le théorème de Pickands
3. Approche semi-paramétrique

Estimation des probabilités et quantiles extrêmes

1. En utilisant le théorème des valeurs extrêmes
2. En utilisant le théorème de Pickands
3. Approche semi-paramétrique

Rappel : théorème des valeurs extrêmes

[Gnedenko, 43] Sous des conditions générales sur F, il existe trois paramètres a_n, b_n et γ tels que :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - a_n}{b_n} \leq x\right) = H_\gamma(x),$$

avec, si $\gamma \neq 0$,

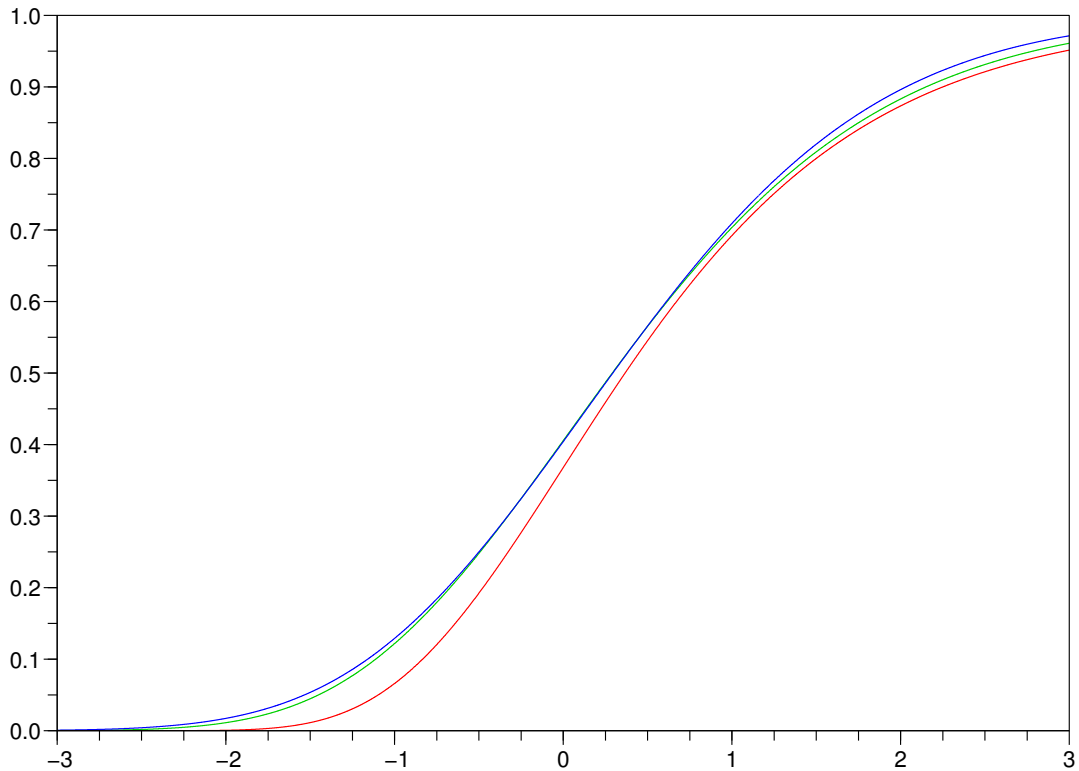
$$H_\gamma(x) = \exp\left(-(1+\gamma x)_+^{-1/\gamma}\right)$$

où $y_+ = \max(0, y)$ et $H_0(x) = \exp\left(-\mathrm{e}^{-x}\right)$.

Vocabulaire :

- H_γ est la **loi des valeurs extrêmes (EVD)**,
- γ est l'**indice des valeurs extrêmes**.
- a_n et b_n sont des paramètres de normalisation.

Illustration sur une loi normale



Comparaison entre $H_\gamma(x)$, $\mathbb{P}\left(\frac{X_{n,n}-a_n}{b_n} \leq x\right)$ avec $n = 10$ et $\mathbb{P}\left(\frac{X_{n,n}-a_n}{b_n} \leq x\right)$ avec $n = 100$

Rappel : Loi des valeurs extrêmes

En pratique,

$$\mathbb{P}(X_{n,n} \leq x) \simeq H_\gamma \left(\frac{x - a_n}{b_n} \right),$$

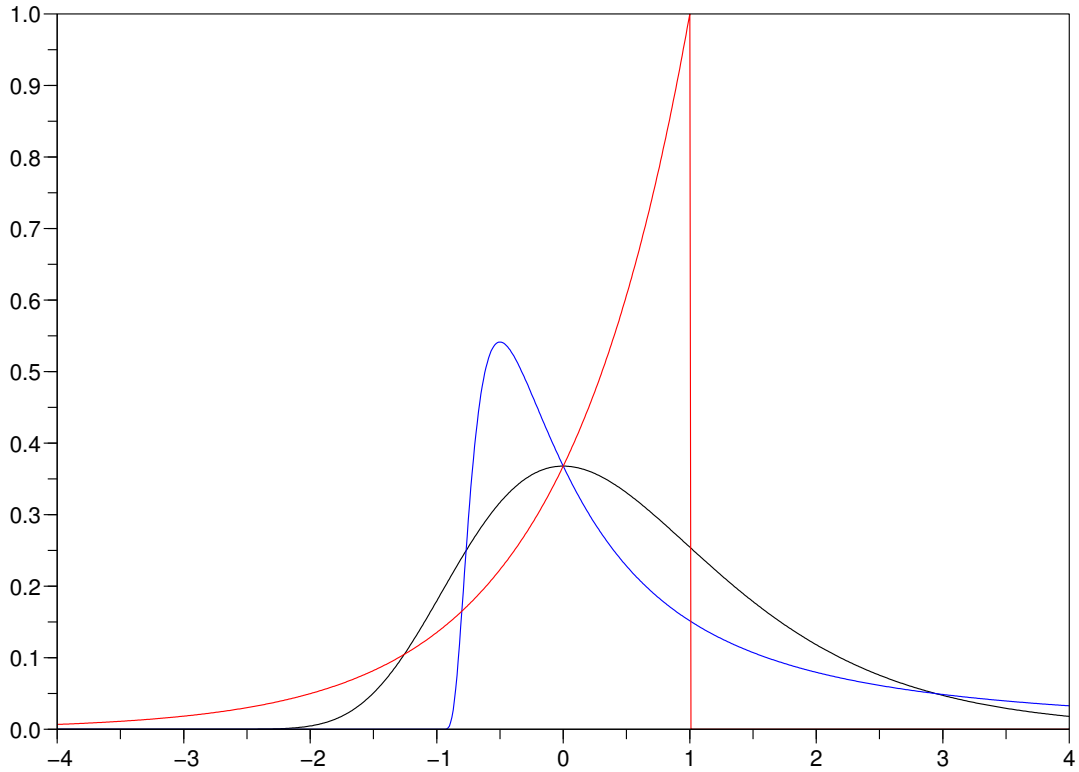
on a une loi à trois paramètres :

- a_n est un paramètre de position, jouant le rôle de $\mathbb{E}(X)$ dans le TCL,
- b_n est un paramètre d'échelle, jouant le rôle de $\sigma(X)/\sqrt{n}$ dans le TCL,
- γ un paramètre de forme, il n'a pas d'équivalent dans le TCL.

On distingue 3 cas (donc 3 types de lois) :

- Si $\gamma > 0$, on dit que F appartient au domaine d'attraction de Fréchet,
- si $\gamma = 0$, on dit que F appartient au domaine d'attraction de Gumbel,
- si $\gamma < 0$, on dit que F appartient au domaine d'attraction de Weibull.

Rappel : Loi des valeurs extrêmes



Exemples de densités associées à la loi des valeurs extrêmes ($\gamma = 0$, $\gamma = 1$ et $\gamma = -1$).

Application à l'extrapolation

Comme $\mathbb{P}(X_{n,n} \leq x) = F^n(x)$, on déduit du théorème des valeurs extrêmes une approximation de $F(x)$ pour les grandes valeurs de x,

$$F(x) = 1 - \bar{F}(x) \simeq H_\gamma^{1/n}\left(\frac{x - a_n}{b_n}\right),$$

et en passant au logarithme

$$\log(1 - \bar{F}(x)) \simeq \frac{1}{n} \log H_\gamma\left(\frac{x - a_n}{b_n}\right).$$

Comme x est grand, $\bar{F}(x)$ est petit, un développement limité au 1er ordre de $\log(1 + u)$ donne donc

$$\bar{F}(x) \simeq -\frac{1}{n} \log H_\gamma\left(\frac{x - a_n}{b_n}\right).$$

Application à l'extrapolation

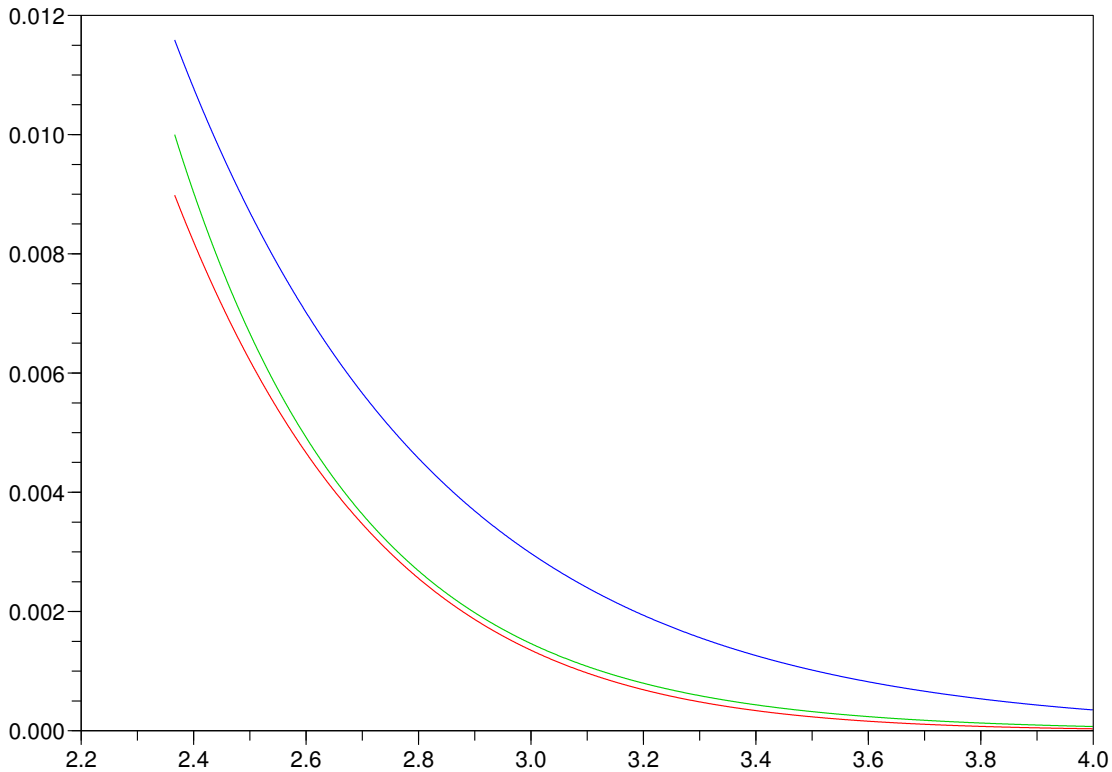
On a donc une approximation de la fonction de survie en queue :

$$\begin{aligned} \bar{F}(x) &\simeq \frac{1}{n}\left[1+\gamma\left(\frac{x-a_n}{b_n}\right)\right]^{-1/\gamma} \quad \text{si} \quad \gamma \neq 0 \\ &\simeq \frac{1}{n}\exp\left(-\frac{x-a_n}{b_n}\right) \quad \text{si} \quad \gamma = 0 \end{aligned}$$

et de son inverse :

$$\begin{aligned} \bar{F}^{-1}(p) &\simeq a_n + \frac{b_n}{\gamma}\left[(np)^{-\gamma}-1\right] \quad \text{si} \quad \gamma \neq 0 \\ &\simeq a_n - b_n \log(np) \quad \text{si} \quad \gamma = 0. \end{aligned}$$

Illustration sur une loi normale



Comparaison entre $\bar{F}(x)$, $\frac{1}{n}\exp\left(-\frac{x-a_n}{b_n}\right)$ avec $n = 10$ et $\frac{1}{n}\exp\left(-\frac{x-a_n}{b_n}\right)$ avec $n = 100$

Illustration sur une loi normale

Ici on a utilisé les valeurs théoriques de a_n, b_n et γ connues pour la loi normale centrée-réduite.

Problème : Les paramètres a_n, b_n et γ sont inconnus dans la pratique puisqu'on ne connait pas F, il faut les estimer.

Estimation des paramètres de la loi des valeurs extrêmes

On souhaite estimer les paramètres de la loi des valeurs extrêmes de fdr

$$H_{\gamma,a,b}(x) \overset{def}{=} H_\gamma\left(\frac{x-a}{b}\right) = \exp\left\{-\left[1+\gamma\left(\frac{x-a}{b}\right)\right]_+^{-1/\gamma}\right\}$$

Deux difficultés :

- Il faut un échantillon de maxima (parfois difficiles à extraire des données initiales, petit nombre d'observations utilisées).
- Les estimateurs du maximum de vraisemblance ne sont pas explicites.

Estimateurs du maximum de vraisemblance

Soit $\{Y_1, \ldots, Y_k\}$ un échantillon de k maxima indépendants tous de fdr $H_{\gamma,a,b}$.

- **Cas particulier :** γ est connu et $\gamma = 0$. Système de 2 équations à 2 inconnues :

$$\begin{aligned} \sum_{i=1}^{k} \exp\left(-\frac{Y_i - a}{b}\right) &= k, \\ \sum_{i=1}^{k} \frac{Y_i - a}{b}\left(1 - \exp\left(-\frac{Y_i - a}{b}\right)\right) &= k, \end{aligned}$$

  - Pas de solution explicite, méthodes numériques.
  - Propriétés asymptotiques connues.
- **Cas général :** Le support de la loi dépend des paramètres, propriétés asymptotiques [Smith, 1985].

Estimateurs des moments pondérés

[Hosking, Wallis, Wood, 1985]. On peut définir le moment pondéré d'ordre r par

$$\mu_r = \mathbb{E}\left[Y H^r_{\gamma,a,b}(Y)\right].$$

Cette quantité existe pour $\gamma < 1$ et est donnée par

$$\mu_r = \frac{1}{r+1}\left[a - \frac{b}{\gamma}\left\{1 - (r+1)^\gamma \Gamma(1-\gamma)\right\}\right],$$

où Γ est la fonction définie par

$$\Gamma(t) = \int_0^{+\infty} x^{t-1} \exp(-x) dx.$$

Estimateurs des moments pondérés

Pour calculer a, b et γ, trois moments pondérés suffisent :

$$\begin{aligned} \mu_0 &= a - \frac{b}{\gamma}\{1 - \Gamma(1-\gamma)\} \\ 2\mu_1 - \mu_0 &= -\frac{b}{\gamma}(1 - 2^{\gamma})\Gamma(1-\gamma) \\ \frac{3\mu_2 - \mu_0}{2\mu_1 - \mu_0} &= \frac{1 - 3^{\gamma}}{1 - 2^{\gamma}}. \end{aligned}$$

En inversant ces formules, on obtient (a, b, γ) en fonction de (μ_0, μ_1, μ_2). Il reste à estimer ces trois moments.

Estimateurs des moments pondérés

On remplace l'espérance par une moyenne empirique

$$\mu_r \simeq \frac{1}{k}\sum_{i=1}^{k} Y_i H^r_{\gamma,a,b}(Y_i) = \frac{1}{k}\sum_{i=1}^{k} Y_{i,k} H^r_{\gamma,a,b}(Y_{i,k})$$

en ordonnant les observations. On remplace $H_{\gamma,a,b}$ par la fdr empirique :

$$\mu_r \simeq \frac{1}{k}\sum_{i=1}^{k} Y_{i,k} \hat{F}^r_k(Y_{i,k}) = \frac{1}{k}\sum_{i=1}^{k} Y_{i,k} \left(\frac{i-1}{k}\right)^r .$$

On obtient alors un estimateur sous forme d'une combinaison linéaire :

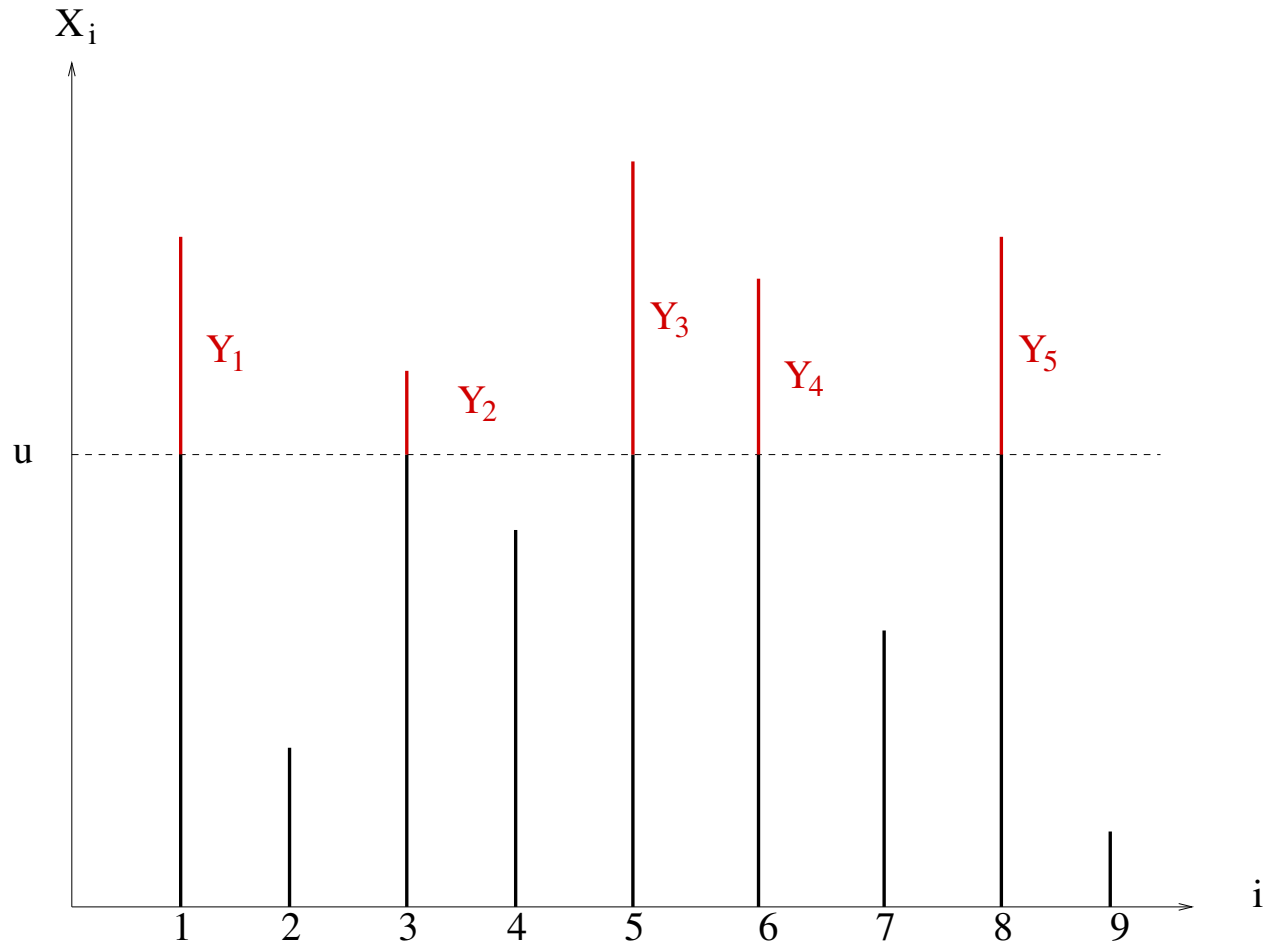
$$\hat{\mu}_r = \frac{1}{k}\sum_{i=1}^{k} Y_{i,k} \left(\frac{i-1}{k}\right)^r .$$

Estimation des probabilités et quantiles extrêmes

1. En utilisant le théorème des valeurs extrêmes
2. En utilisant le théorème de Pickands
3. Approche semi-paramétrique

Rappel : définition d'un excès

Plutôt que de se focaliser sur le maximum, on étudie les valeurs dépassant un seuil donné. **L'excès Y de la variable X au dessus du seuil** u est défini par $X - u$ quand $X \geq u$.



Fonction de survie d'un excès

La fonction de survie $\bar{F}_u$ d'un excès au dessus de u est donnée pour $y > 0$ par

$$\begin{aligned}
\bar{F}_u(y) &= \mathbb{P}(Y \geq y) \\
&= \mathbb{P}(X - u \geq y | X \geq u) \\
&= \frac{\mathbb{P}(X \geq u + y, X \geq u)}{\mathbb{P}(X \geq u)} \\
&= \frac{\bar{F}(u + y)}{\bar{F}(u)}
\end{aligned}$$

Lorsque le seuil est grand, on peut approcher cette quantité par la fonction de survie d'une **loi de Pareto Généralisée (GPD)**.

Rappel : Loi de Pareto Généralisée

Sa fonction de survie est donnée par

$$\begin{aligned} \bar{G}_{\gamma,\sigma}(y) &= \left(1+\gamma\frac{y}{\sigma}\right)^{-1/\gamma} \quad \text{si} \quad \gamma \neq 0, \\ &= \exp\left(-\frac{y}{\sigma}\right) \quad \text{sinon.} \end{aligned}$$

Son ensemble de définition est $\mathbb{R}^+$ si $\gamma \geq 0$ ou $[0, -\sigma/\gamma[$ si $\gamma < 0$.
Elle dépend de deux paramètres :

- $\sigma > 0$ est un paramètre d'échelle,
- $\gamma \in \mathbb{R}$ est un paramètre de forme.

Deux cas particuliers :

- $\gamma = 0$, loi exponentielle d'espérance σ,
- $\gamma = -1$, loi uniforme sur $[0, \sigma]$.

Rappel : Théorème de Pickands

[Pickands, 1975] Il y a équivalence entre la convergence en loi du maximum vers une EVD et la convergence en loi d'un excès vers une GPD :

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - a_n}{b_n} \leq x\right) = H_\gamma(x),$$

si et seulement si

$$\lim_{u\to x_F} \sup_{y\in[0,x_F-u]} |\bar{F}_u(y) - \bar{G}_{\gamma,\sigma(u)}(y)| = 0.$$

On remarque que le paramètre de forme γ est le même pour l'EVD et la GPD.

Application à l'extrapolation

En utilisant le théorème de Pickands, on a, pour $y \geq 0$,

$$\bar{F}_u(y) = \frac{\bar{F}(u+y)}{\bar{F}(u)} \simeq \bar{G}_{\gamma,\sigma}(y).$$

Avec le changement de variable $x = u + y$ on obtient l'approximation (valable pour $x \geq u$) :

$$\bar{F}(x) \simeq \bar{F}(u)\bar{G}_{\gamma,\sigma}(x-u).$$

Finalement, on introduit la probabilité α que X dépasse u, $\alpha = \bar{F}(u)$, d'où

$$\bar{F}(x) \simeq \alpha\bar{G}_{\gamma,\sigma}(x - \bar{F}^{-1}(\alpha)).$$

Application à l'extrapolation

On a donc une approximation de la fonction de survie en queue :

$$\begin{aligned}\bar{F}(x) &\simeq \alpha\left[1+\gamma\left(\frac{x-\bar{F}^{-1}(\alpha)}{\sigma}\right)\right]^{-1/\gamma} \quad \text{si} \quad \gamma \neq 0 \\ &\simeq \alpha \exp\left(-\frac{x-\bar{F}^{-1}(\alpha)}{\sigma}\right) \quad \text{si} \quad \gamma = 0\end{aligned}$$

et de son inverse :

$$\begin{aligned}\bar{F}^{-1}(p) &\simeq \bar{F}^{-1}(\alpha)+\frac{\sigma}{\gamma}\left[\left(\frac{p}{\alpha}\right)^{-\gamma}-1\right] \quad \text{si} \quad \gamma \neq 0 \\ &\simeq \bar{F}^{-1}(\alpha)-\sigma\log\left(\frac{p}{\alpha}\right) \quad \text{si} \quad \gamma = 0.\end{aligned}$$

Comparaison avec l'approche EVD

Les expressions sont les mêmes, il y a trois paramètres inconnus :

- l'indice des valeurs extrêmes γ,
- σ qui joue le rôle de b_n dans l'approche EVD,
- $\bar{F}^{-1}(\alpha)$ qui joue le rôle de a_n dans l'approche EVD.

Avantages :

- Il est plus facile d'avoir un échantillon d'excès que de maxima,
- $\bar{F}^{-1}(\alpha)$ est un quantile classique, facile à estimer par inversion de la fonction de survie empirique.
  En pratique : on choisit $\alpha = k/n$, où k est le nombre d'excès,

Il reste à estimer γ et σ.

Estimateurs du maximum de vraisemblance

Soit $\{Y_1, \ldots, Y_k\}$ un échantillon de k excès indépendants tous de fdr $G_{\gamma,\sigma}$.

- **Cas particulier 1 :** γ est connu et $\gamma > 0$. Une équation en σ :

$$\sum_{i=1}^{k} \frac{Y_i - \sigma}{\gamma Y_i + \sigma} = 0.$$

  - Pas de solution explicite, méthodes numériques.
  - Propriétés asymptotiques connues :

$$\sqrt{k}\frac{\hat{\sigma}_k - \sigma}{\sigma\sqrt{2\gamma + 1}} \xrightarrow{L} N(0,1)$$

Estimateurs du maximum de vraisemblance

- **Cas particulier 2 :** γ est connu et $\gamma = -1$. Le support de la loi dépend des paramètres. Estimateur explicite :

$$\hat{\sigma}_k = \max_i Y_i$$

Propriétés asymptotiques connues :

$$k\frac{\hat{\sigma}_k - \sigma}{\sigma} \xrightarrow{L} EVD$$

- **Cas général :**
  - Le système de 2 équations à 2 inconnues se ramène à 1 équation à 1 inconnue par un changement de variable astucieux.
  - Propriétés asymptotiques : [Smith, 1985].

Estimateurs des moments pondérés

[Hosking, Wallis, 1987]. On peut définir un autre type de moment pondéré d'ordre s par

$$\nu_s = \mathbb{E}\left[Y \bar{G}^s_{\gamma,\sigma}(Y)\right].$$

Cette quantité existe pour $\gamma < 1$ et est donnée par

$$\nu_s = \frac{\sigma}{(s+1)(s+1-\gamma)}.$$

Pour obtenir γ et σ, deux moments suffisent

$$\gamma = \frac{4\nu_1 - \nu_0}{2\nu_1 - \nu_0} \text{ et } \sigma = \frac{2\nu_1\nu_0}{\nu_0 - 2\nu_1},$$

on estime ensuite ν_0 et ν_1 classiquement.

Estimation des probabilités et quantiles extrêmes

1. En utilisant le théorème des valeurs extrêmes
2. En utilisant le théorème de Pickands
3. Approche semi-paramétrique

Modèle semi-paramétrique

On se restreint au **domaine d'attraction de Fréchet** où l'on a la caractérisation

$$\bar{F}(x) = x^{-1/\gamma}\ell(x),$$

avec ℓ une fonction à variations lentes et $\gamma > 0$. Ce modèle de fonction de survie comporte :

- une partie paramétrique $x^{-1/\gamma}$ ne dépendant que d'un paramètre réel (γ).
- une partie non-paramétrique $\ell(x)$ sur laquelle on sait seulement que

$$\lim_{u\to\infty} \frac{\ell(tu)}{\ell(u)} = 1,$$

pour $t > 1$.

Application à l'extrapolation

Pour $t > 1$,

$$\lim_{u\to\infty} \frac{\bar{F}(tu)}{\bar{F}(u)} = t^{-1/\gamma} \left(\lim_{u\to\infty} \frac{\ell(tu)}{\ell(u)} \right) = t^{-1/\gamma}.$$

On en déduit l'approximation

$$\bar{F}(tu) \simeq \bar{F}(u) t^{-1/\gamma}.$$

En posant $x = tu$ et $\alpha = \bar{F}(u)$, on a

$$\begin{aligned} \bar{F}(x) &\simeq \alpha \left(\frac{x}{\bar{F}^{-1}(\alpha)} \right)^{-1/\gamma} \\ \bar{F}^{-1}(p) &\simeq \bar{F}^{-1}(\alpha) \left(\frac{p}{\alpha} \right)^{-\gamma}, \end{aligned}$$

pour $x > u$ ou de façon équivalente $p \leq \alpha$.

Application à l'extrapolation

Remarques :

- Ces approximations sont des cas particuliers de l'approche GPD avec $\sigma = \gamma \bar{F}^{-1}(\alpha)$;
- $\bar{F}^{-1}(\alpha)$ s'estime comme nous l'avons déjà vu par une des observations ordonnées.
- Il reste uniquement à estimer γ, en se basant encore sur

$$\bar{F}^{-1}(p) \simeq \bar{F}^{-1}(\alpha) \left(\frac{p}{\alpha}\right)^{-\gamma},$$

que l'on peut réécrire

$$\log \bar{F}^{-1}(p) - \log \bar{F}^{-1}(\alpha) \simeq \gamma \log(\alpha/p).$$

Estimation semi-paramétrique de γ

On choisit comme précédemment, $\alpha = k/n$ et on considère plusieurs valeurs de $p = i/n$, $i = 1, \ldots, k-1$. (on doit avoir $p < \alpha$). On obtient :

$$\log \bar{F}^{-1}(i/n) - \log \bar{F}^{-1}(k/n) \simeq \gamma \log(k/i),$$

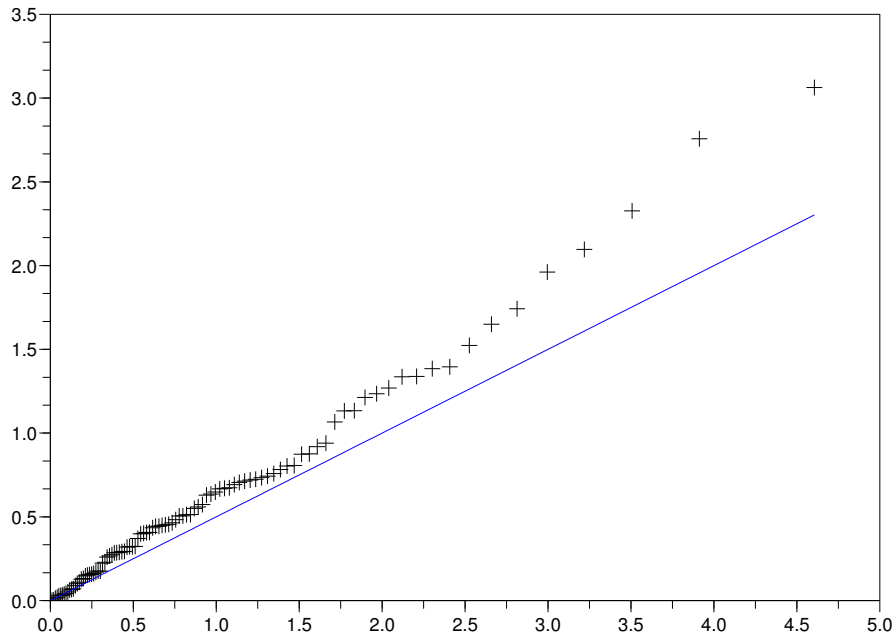
et en estimant les fonctions de survies par leurs équivalents empiriques,

$$\log X_{n-i+1,n} - \log X_{n-k+1,n} \simeq \gamma \log(k/i).$$

Il est possible de vérifier graphiquement cette approximation.

Estimation semi-paramétrique de γ

Simulation de $n = 500$ réalisations d'une loi de Student à 2 degrés de liberté ($\gamma = 1/2$). On a choisi $k = 100$.



En abscisse : $\log(k/i)$. En ordonnée : $y = x/2$ et $\log X_{n-i+1,n} - \log X_{n-k+1,n}$ pour $i = 1, \ldots, k-1$.

Estimation semi-paramétrique de γ

En sommant de part et d'autre sur $i = 1, \ldots, k-1$, on obtient

$$\gamma \simeq \frac{\displaystyle\sum_{i=1}^{k-1} \log X_{n-i+1,n} - \log X_{n-k+1,n}}{\displaystyle\sum_{i=1}^{k-1} \log(k/i)}$$

Le dénominateur se réécrit $\log[k^{k-1}/(k-1)!]$, en utilisant la formule de Stirling,

$$m! \sim \sqrt{2\pi} m^{m+1/2} \mathrm{e}^{-m}$$

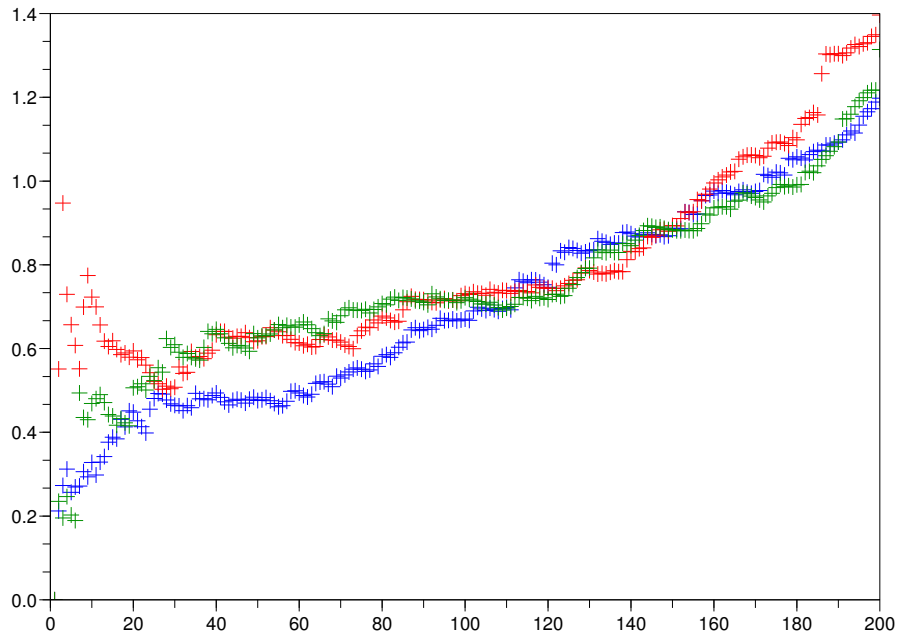
il est équivalent à k au voisinage de l'infini. On obtient l'**Estimateur de Hill**

$$\hat{\gamma}(k) = \frac{1}{k} \sum_{i=1}^{k-1} (\log X_{n-i+1,n} - \log X_{n-k+1,n}),$$

[Hill, 1975].

Comportement de l'estimateur de Hill

Trois simulations de $n = 500$ réalisations d'une loi de Student à 2 degrés de liberté ($\gamma = 1/2$).



En abscisse : k. En ordonnée : $\hat{\gamma}(k)$ pour $k = 1, \ldots, 200$.

En pratique ...

Le choix de k est difficile :

- Si k est petit, $\hat{\gamma}(k)$ utilise peu d'observations, il a alors une **grande variance**.
- Si k est grand, le seuil estimé $X_{n-k+1,n}$ est petit, on sort de la zône où la fonction de survie est approximativement une puissance, $\hat{\gamma}(k)$ a alors un **grand biais**.

Bibliographie

- Livres de référence :

P. Embrechts, P., C. Klüppelberg, and T. Mikosch (1997), *Modelling extremal events*, Springer.

M. Falk, J. Hüsler and R. Reiss (2004), *Laws of small numbers : Extremes and rare events*, 2nd edition, Birkhäuser.

R. Reiss and M. Thomas (2001), *Statistical analysis of extreme values*, Birkhäuser, Basel.

N. Bingham, C. Goldie and J. Teugels (1987), *Regular variation*, Encyclopedia of Mathematics and its Applications, **27**, Cambridge University Press.

Bibliographie

- Article fondateur :
  B. Gnedenko (1943), Sur la distribution limite du terme maximum d'une série aléatoire, *The annals of Mathematics*, 2nd Ser., **44**, 423–453.

- Moments pondérés :
  J.R.M. Hosking, J.R. Wallis and E.F. Wood (1985), Estimation of the Generalized Extreme-Value distribution by the method of probability-weighted moments, *Technometrics*, **27**, 251–261.

  J.R.M. Hosking and J.R. Wallis (1987), Parameter and quantile estimation for the Generalized Pareto Distribution, *Technometrics*, **29**, 1339–1349.

Bibliographie

- Premier estimateur de l'indice des valeurs extrêmes :
  B.M. Hill (1975), A simple general approach to inference about the tail of a distribution, *The Annals of Statistics*, **3**, 1163–1174.

- Comportement théorique du maximum de vraisemblance :
  R. Smith (1985), Maximum likelihood estimation in a class of non regular cases, *Biometrika*, **72**, 67–90.