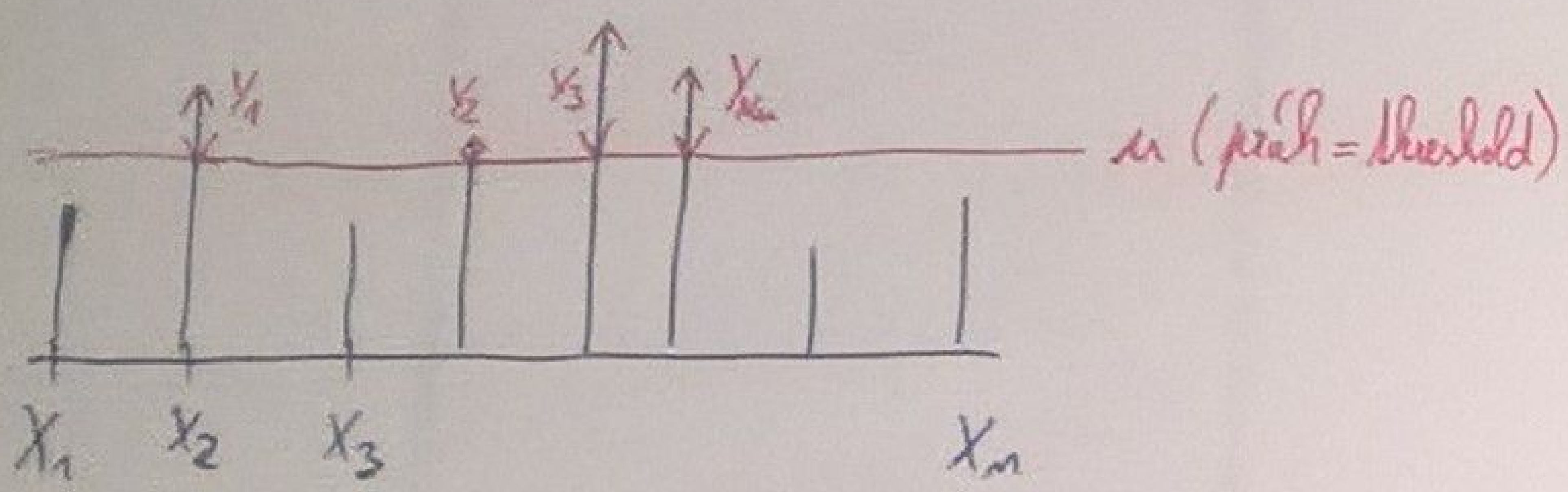
Metoda peaks-over-threshold (POT) – překročení hranice

$X_1,\dots,X_n$ je náh. výběr s distribuční funkcí F

$Y_i = X_i - u$, pro $X_i > u$ ... výše překročení meze (prahu) u (exces)

pro $i = 1,2,\dots,N_u$ $\quad N_u$ = počet X_i, které překročí hranici u

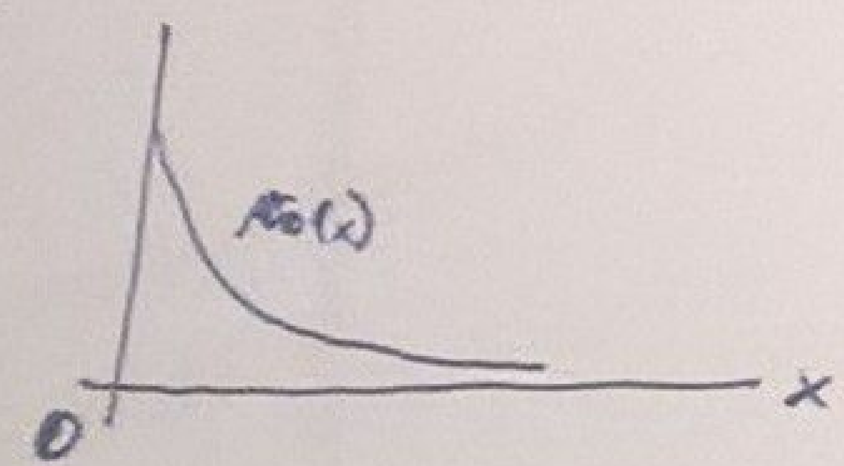


Hledejme distribuční funkci excesů Y_i. Označme ji $F_u(x) = P(Y_i \le x) = P(X_i - u \le x \mid X_i > u) = P(X_i \le u + x \mid X_i > u) = \dfrac{P(u < X_i \le u+x)}{P(X_i > u)} = \dfrac{F(u+x) - F(u)}{1 - F(u)}$, $x \ge 0$.

Rozdělení pro modelování excesů

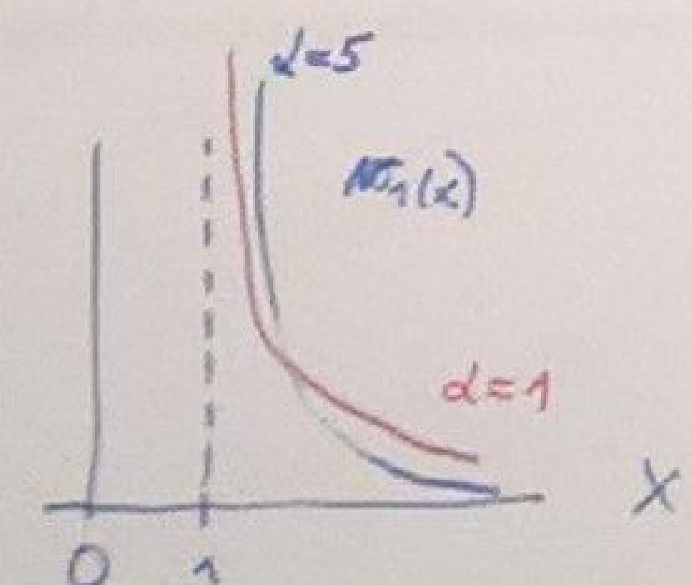
1) Exponenciální rozdělení

$$W_0(x) = 1 - e^{-x}, \quad x \ge 0$$



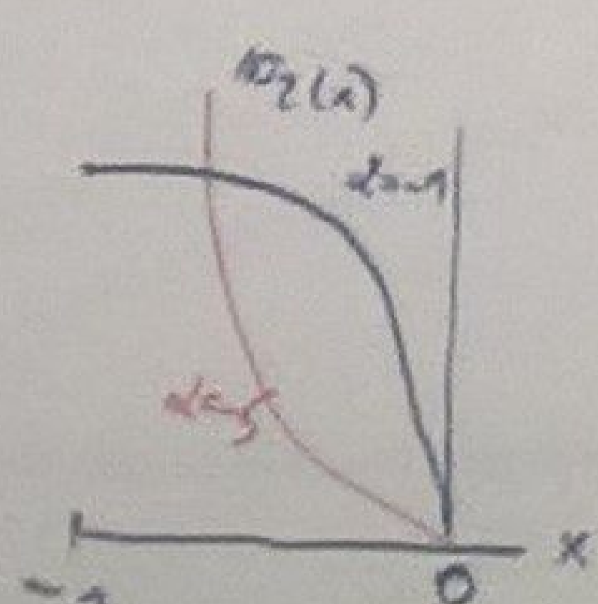
2) Paretovo rozdělení s parametrem tvaru $\alpha > 0$

$$W_1(x) = 1 - x^{-\alpha}, \quad x \ge 1$$



3) Beta rozdělení s parametrem tvaru $\alpha < 0$

$$W_2(x) = 1 - (-x)^{-\alpha}, \quad -1 \le x \le 0$$



Opět můžeme přidat parametr polohy μ a parametr měřítka σ:

$$W_{i,\mu,\sigma}(x) = W_i\left(\frac{x-\mu}{\sigma}\right) \quad , \text{pro } i = 0,1,2.$$

Předchozí 3 distribuční funkce můžeme zapsat pomocí jednoho vzorečku:

Zobecněné Paretovo rozdělení (GPD)

$$W_\gamma(x) = 1 - (1+\gamma x)^{-\frac{1}{\gamma}} \quad , \gamma \in \mathbb{R} \text{ parametr}$$

Poznámka

pro $\gamma = 0$ chápeme $W_\gamma(x)$ jako $\lim_{\gamma \to 0} W_\gamma(x) = 1 - e^{-x}$, $x > 0$ ... exponenciální rozdělení

$\gamma > 0$ ... Paretovo rozdělení , $x \geq 0$.

$\gamma < 0$ ... Beta rozdělení , $0 \leq x \leq -\frac{1}{\gamma}$.

Úplnou třídu dostaneme přidáním parametru měřítka σ: $W_{\gamma,\sigma}(x) = 1 - \left(1 + \frac{\gamma}{\sigma} x\right)^{-\frac{1}{\gamma}}$.

Poznámka

Protože modelujeme excesy, které jsou nezáporné, parametr polohy $\mu = 0$.

Věta (Balkema, de Haan, Pickands)

Nechť $X_1, \dots, X_n$ je náh. výběr s distribuční funkcí F. Nechť existují posloupnosti konstant $a_n > 0$ a $b_n \in \mathbb{R}$ takové, že $F_{u_n}(a_n x + b_n) \xrightarrow[u_n \nearrow x_F]{} H(x)$ pro nějakou spojitou distribuční funkci H a $x_F = \sup\{x : F(x) < 1\}$ je pravý koncový bod nosiče F, pak H je distribuční funkce zobecněného Paretova rozdělení.

Poznámka

- podle předchozí věty je jediné možné limitní rozdělení excesů GPD rozdělení
- excesy $Y_i = X_i - u$ pro $X_i > u$ budeme modelovat pomocí GPD rozdělení s distribuční funkcí $W_{\gamma,\sigma}(x)$. Jeho parametry γ a σ odhadneme z dat.

Věta

Distribuční funkce $F \in MDA(G_\gamma) \iff$ existuje kladná funkce $\beta(u)$ taková, že $\lim_{u \nearrow x_F} \sup_{0 < x < x_F - u} |F_u(x) - W_{\gamma,\beta(u)}(x)| = 0.$

Poznámka

Rozdělení, pro která normovaná maxima konvergují k GEV rozdělení jsou stejná jako rozdělení, pro která rozdělení excesů konvergují k GPD rozdělení $\left(W_\gamma(x) = 1 + \log G_\gamma(x)\right)$.

Navíc, parametr γ je v obou případech stejný (a nezávisí na u).

Stabilita excesů GPD rozdělení

Nechť $X_1, \dots, X_n$ je náh. výběr ze zobecněného Paretova rozdělení s parametry γ a σ.

$$W_{\gamma,\sigma}(x) = 1 - \left(1 + \frac{\gamma}{\sigma} x\right)^{-\frac{1}{\gamma}}.$$

distribuční funkce excesů Y_i je $F_u(y) = \dfrac{F(u+y)-F(u)}{1-F(u)} = \dfrac{\left(1+\frac{\gamma}{\sigma}u\right)^{-\frac{1}{\gamma}} - \left(1+\frac{\gamma}{\sigma}(u+y)\right)^{-\frac{1}{\gamma}}}{\left(1+\frac{\gamma}{\sigma}u\right)^{-\frac{1}{\gamma}}} = 1-\left[\dfrac{1+\frac{\gamma}{\sigma}(u+y)}{1+\frac{\gamma}{\sigma}u}\right]^{-\frac{1}{\gamma}} = 1-\left[\dfrac{\sigma+\gamma(u+y)}{\sigma+\gamma u}\right]^{-\frac{1}{\gamma}} =$

$= 1-\left(1+\dfrac{\gamma y}{\sigma+\gamma u}\right)^{-\frac{1}{\gamma}} = W_{\gamma,\sigma+\gamma u}(y)$ ... distribuční funkce zob. Paretova rozdělení s parametry γ a $\sigma+\gamma u$.

A dokázáno, zobecněné Paretovo rozdělení je jediné spojité rozdělení s touto vlastností.

<u>Metoda POT při praktickém zpracování dat</u>

$X_1,\dots,X_n$ náh. výběr s distribuční funkcí F

Budeme předpokládat, že pro F jsou splněny předpoklady Balk.-de Haanovy-Pickandsovy věty

zvolíme hranici u a definujeme excesy $Y_i = X_i - u$ pro $X_i > u$ pro $i = 1, 2, \dots, N_u$ (N_u je počet excesů)

$Y_1,\dots,Y_{N_u}$ pokládáme za náh. výběr z GPD rozdělení s parametry γ a σ. A ty budeme dále odhadovat.

metoda maximální věrohodnosti:

$$W_{\gamma,\sigma}(x) = 1-\left(1+\frac{\gamma}{\sigma}x\right)^{-\frac{1}{\gamma}}$$

hustota je

$$w_{\gamma,\sigma}(x) = \frac{1}{\gamma}\left(1+\frac{\gamma}{\sigma}x\right)^{-\frac{1}{\gamma}-1}\cdot\frac{\gamma}{\sigma} = \frac{1}{\sigma}\left(1+\frac{\gamma}{\sigma}x\right)^{-\left(1+\frac{1}{\gamma}\right)}$$

$$\log w_{\gamma,\sigma}(x) = -\log\sigma - \left(1+\frac{1}{\gamma}\right)\log\left(1+\frac{\gamma}{\sigma}x\right)$$

logaritmická věrohodnostní funkce: $\ell(\gamma,\sigma) = \sum_{i=1}^{N_u} \log w_{\gamma,\sigma}(Y_i) = -N_u \log\sigma - \left(1+\frac{1}{\gamma}\right) \sum_{i=1}^{N_u} \log\left(1+\frac{\gamma}{\sigma} Y_i\right)$.

<u>Poznámka</u>

$\ell(\gamma,\sigma)$ se maximalizuje přímo $\Rightarrow$ maximálně věrohodný odhad $(\hat{\gamma},\hat{\sigma})^T$.

pro $\gamma > -\frac{1}{2}$ je dokázána jejich asymptotická nestrannost, konzistence a as. normalita.

metoda pravděpodobnostně vážených momentů (PWM)

připomeňme, že má-li n.v. X distr. fci F definujeme $M_{p,r,s} = E\left[X^p[F(X)]^r[1-F(X)]^s\right]$ pro $p,r,s \in \mathbb{R}$ pravděpodobnostně vážené momenty.

speciální případ: $p=1$, $r=0$

$$\alpha_s = M_{1,0,s} = E\left[X \cdot [1-F(X)]^s\right] \quad \text{pro } s=0,1$$

pro GPD je $\alpha_s = \dfrac{\sigma}{(s-\gamma+1)(s+1)}$.

odhad α_s je $\hat{\alpha}_s = \dfrac{1}{N_u} \sum_{j=1}^{N_u} \left(\prod_{r=1}^{s} \dfrac{N_u-j-k+1}{N_u-k} \right) Y_{(j)}$ pro $s=0,1$.

Odhad $(\gamma,\sigma)^T$ získáme řešením soustavy $\alpha_s = \hat{\alpha}_s$ pro $s=0,1$.

Úkoly

1.) Odhadnout $P(X_i > x)$ pro x velké

pro $x \geq u$ platí $P(X_i > x \mid X_i > u) = \dfrac{P(X_i > x)}{P(X_i > u)} \Rightarrow$

$$P(X_i > x) = P(X_i > u) \cdot P(X_i > x \mid X_i > u) = P(X_i > u) \cdot \overbrace{P(Y_i > x-u \mid X_i > u)}^{1 - P(Y_i \leq x-u \mid X_i > u)} = P(X_i > u) \cdot (1 - F_u(x-u)) = P(X_i > u) \cdot (1 - W_{\gamma,\sigma}(x-u)) =$$

$$P(X_i > u) \cdot \left(1 + \frac{\gamma(x-u)}{\sigma}\right)^{-\frac{1}{\gamma}}$$

$$\widehat{P(X_i > x)} = \frac{N_u}{n} \cdot \left(1 + \frac{\hat{\gamma}(x-u)}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\gamma}}} \quad \text{pro } x \geq u$$

2.) Úroveň návratu – extrémní jev, který nastane v průměru jednou za k pozorování

Hledáme $(1-\frac{1}{k})$-kvantil rozdělení n.v. X_i, označme jej $q = q_{1-\frac{1}{k}}$

$$P(X_i > q) = \frac{1}{k}$$

$$P(X_i > u) \cdot \left(1 + \frac{\gamma(q-u)}{\sigma}\right)^{-\frac{1}{\gamma}} = \frac{1}{k} \quad \ldots \quad \rightsquigarrow \quad q = u + \frac{\sigma}{\gamma}\left[\left[k \cdot P(X_i > u)\right]^{\gamma} - 1\right]$$

tedy odhad úrovně návratu je $\hat{q} = \hat{q}_{1-\frac{1}{k}} = u + \frac{\hat{\sigma}}{\hat{\gamma}}\left[\left(\frac{k \cdot N_u}{n}\right)^{\hat{\gamma}} - 1\right]$.

3) Doba návratu

známe hodnotu $q_{1-\frac{1}{k}}$ „q" a hledáme příslušné k

$$k = \frac{1}{P(X_i > u) \cdot \left(1 + \frac{\gamma(q-u)}{\sigma}\right)^{-\frac{1}{\gamma}}} = \frac{\left(1 + \frac{\gamma(q-u)}{\sigma}\right)^{\frac{1}{\gamma}}}{P(X_i > u)} \quad \overset{\text{odhad}}{\Longrightarrow} \quad \hat{k} = \frac{n\left(1 + \frac{\hat{\gamma}(q-u)}{\hat{\sigma}}\right)^{\frac{1}{\hat{\gamma}}}}{N_u}.$$

Jak volit velikost prahu u?

- hodnota prahu u musí být dostatečně vysoká, aby „fungovala" aproximace GPD rozdělením
- příliš vysoká hodnota prahu u však znamená nedostatečný počet excesů N_u pro odhady parametrů GPD rozdělení

mean excess plot

definujme střední hodnotu velikosti překročení prahu u (za podmínky, že k překročení došlo): mean excess: $e(u) = E(X_i - u \mid X_i > u)$.

Poznámka

Má-li n.v. Z zobecněné Paretovo rozdělení s parametry γ a σ, pak $EZ = \frac{\sigma}{1-\gamma}$.

Je-li $X_1, \dots, X_n$ je náh. výběr z GPD rozdělení s parametry γ, σ, pak $X_i - u \mid X_i > u$ má GPD rozdělení s parametry γ a $\sigma + \gamma \cdot u$ $\Rightarrow e(u) = \frac{\sigma + \gamma u}{1-\gamma}$.

$e(u)$ je v tomto případě lineární funkce v u. To je charakteristická vlastnost GPD rozdělení.

Pro naše data, tedy odhadneme funkci $e(u)$. Použijeme empirický odhad $\underbrace{\widehat{e(u)} = \frac{\sum_{i=1}^{N_u} Y_i}{N_u}}_{\text{průměr excesů}}$, kde $Y_i = X_i - u$ pro $X_i > u$ $i = 1, \dots, N_u$.

Na závěr vykreslíme do grafu body $\{ X_{(i)}, \widehat{e(X_i)}, i = 2, 3, \dots, m \}$.

Pokud pozorování pochází z GPD rozdělení, pak graf vykazuje lineární trend.

Práh u určíme tedy jako bod z grafu, kde závislost „začíná být lineární".

Poznámka

V praxi téměř nepoužitelné.

Práh u se proto často volí jako vyšší empirický kvantil našich dat, např. 96 %.