3. előadás, 2022. február 24

Zempléni András

Valószínűségelméleti és Statisztika Tanszék
Természettudományi Kar
Eötvös Loránd Tudományegyetem

Áringadozások előadás

Az ML becslés tulajdonságai, alternatívák

- Maximum likelihood:
  - Nincs explicit megoldása
  - A szokásos aszimptotikus tulajdonságokkal (optimalitás, normalitás) rendelkezik, ha $\gamma > -0,5$.
  - $\gamma < -1$ esetén nincs lokális maximuma a sűrűségfüggvénynek, a maximális mintaelem a globális maximum - ez konzisztens.
- Alternatív módszerek: probability-weighted-moments
- Rendezett mintán alapuló eljárások (később visszatérünk rá)

Konfidencia tartomány

- A (számunkra érdekes) reguláris esetekben aszimptotikusan jó a normális határeloszlás alkalmazása
- De: a konvergencia általában nem túl gyors, különösen a VaR esetén általában nem pontosak a kapott eredmények
- Ezért célszerű alternatív módszerek alkalmazása

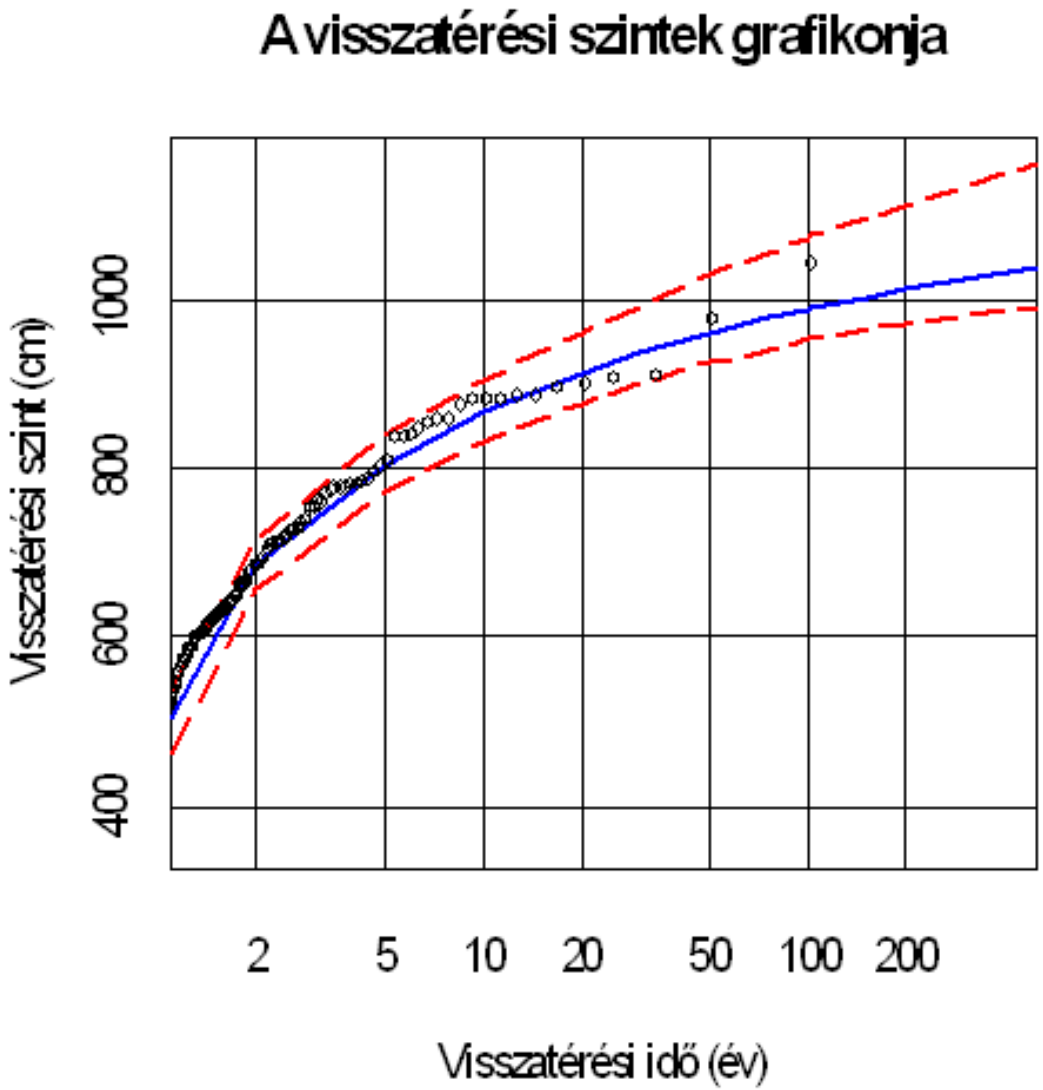
Profil likelihood

- Kis mintákra sokkal jobb tulajdonságú lehet a klasszikus, normalitáson alapuló módszernél
- A háttér itt is aszimptotikus eredmény: a reguláris esetben

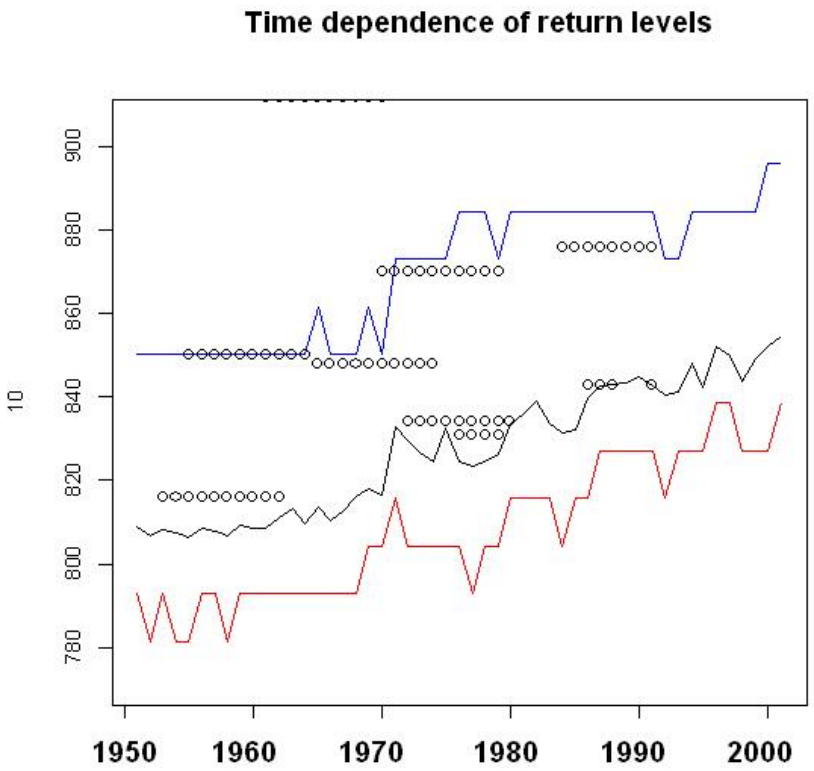
$$\{\vartheta : 2(l(\hat{\vartheta}) - l_p(\vartheta)) \leq h_{1-\alpha,k}\}$$

aszimptotikusan $1-\alpha$ megbízhatóságú konfidencia intervallum. $h_{1-\alpha,k}$ a k szabadságfokú chi-négyzet eloszlás $1-\alpha$ kvantilise, k pedig a vizsgált paraméterek száma (tipikusan $k=1$ és l_p a loglik. fv. maximuma, ha a többi paraméterben maximalizálunk).

Alkalmazás: árvíz-idősorok



Az időfüggés vizsgálata mozgó ablakokkal



ábra: 10 éves visszatérési szintek, 50 éves ablakok alapján

- Minden évben csak a megelőző 50 évet használtuk
- Fekete: becsült 90%-os kvantilis (10 éves visszatérési szint)
- Kék/piros: 95% felső/alsó profil likelihood konfidencia intervallum
- Fekete körök: azok az árvizek 10 éven belül, amik nagyobbak a becsült kvantilisnél (több van, mint a várt, felfelé mutató trend látszik)

Illeszkedésvizsgálat

Klasszikus tesztek:

- Chi-négyzet
- Kolmogorov-Szmirnov

nem túl erősek

- Cramér-von Mises típusú tesztek: a tapasztalati és az elméleti eloszlás eltérésének (esetleg súlyozott) integrálját használják

$$C_n = \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dF(x)$$

- De a becsléses eset miatt a K-S, a CvM (AD) teszteknél szimulált kritikus értékeket kell használni!

Alternatívák

- Anderson-Darling teszt:

$$A^2 = \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x)$$

Kiszámítása:

$$A^2 = -n - \sum_{i=1}^{n} \frac{(2i-1)(\log(z_i) + \log(1 - z_{n+1-i}))}{n}$$

ahol $z_i = F(X_i)$. Az eloszlás mindkét szélén érzékeny.

- Módosítás:

$$B^2 = \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{(1 - F(x))} dF(x)$$

(a maximumra; az eloszlás felső szélére súlyoz). Ennek kiszámítása:

$$B^2 = \frac{n}{2} - \sum_{i=1}^{n} \frac{(2i-1)\log(1 - z_{n+1-i})}{n} - \overline{z}$$

.

További lehetőségek

- Egy másik teszt pedig a GEV eloszlások stabilitási tulajdonságát használja: minden $m \in \mathbb{N}$ re megadható a_m, b_m hogy $F(x) = F^m(a_m x + b_m)$ $(x \in \mathbb{R})$
- A tesztstatisztika:

$$h(a,b) = \sup_x \sqrt{n} \left| F(x) - F^2(ax+b) \right| .$$

- Becslési alternatívák:
  - Találjuk meg azt az az a, b párt, ami minimalizálja $h(a,b)$ értékét (számításigényes algoritmus kell hozzá).
  - Becsüljük meg a GEV paramétereit maximum likelihood módszerrel és ezeket helyettesítsük be a stabilitási tulajdonságba.

Kritikus értékek

- A határeloszlás eloszlásmentes az ismert paraméterek esetére. Például:

$$\sup_x \sqrt{n} \left| F(x) - F^2(a_2 x + b_2) \right| \to \sup_x \left| B(x) - \sqrt{x} B(\sqrt{x}) \right|$$

  ahol B a Brown híd a [0,1]-en.
- Mivel a határeloszlások a normális eloszlás függvényei, a maximum likelihood becslés hatását be lehet építeni a kovariancia struktúra transzformációjával.
- Gyakorlatban: szimulált kritikus értékeket célszerű használni (különösen előnyös kis mintáknál).

A teszt erejének vizsgálata

A helyes döntés valószínűsége (p=0,05):

n	100	200	400	100	200	200	400
teszt eloszlás		Neg.bin.		exp.		norm.	
K-S	0,27	0,49	0,88	0,36	0,61	0,19	0,23
B	0,02	0,27	0,49	0,17	0,58	0,05	0,08
A-D	0,31	0,62	0,96	0,72	0,97	0,21	0,34
h	0,67	0,87	0,99	0,75	0,91	0,10	0,14

A tipikus alternatívák esetére az A-D teszt tűnik a legerősebbnek. A *h* teszt ereje erősen függ az aktuális eloszlás alakjától.

Alkalmazások

- Speciális esetekben, ahol az eloszlás felső széle a legfontosabb, (pl. árvízi adatoknál), a B-teszt a legérzékenyebb.
- Ha a fenti teszteket árvízi adatokra alkalmaztuk (éves maximumok; 50 évnyi ablakok), akkor jónéhány esetben el kellett utasítani a GEV hipotézist a 95%-os szinten.
- Lehetséges okok:
  - Nemstacionaritás
  - A folyó medrének változása miatt (alak, vegetáció stb).
  - Klímaváltozás?
  - Periodikusság?

Időfüggés

- Mostanáig nem vettük figyelembe az egymás utáni megfigyelések lehetséges összefüggését (kivéve a POT módszernél a declusterezést)
- Vannak rutinszerűen alkalmazható idősoros modellek
- Ezek tanulmányozása nem tartozik a tárgyunk témái közé, mi csak az összefüggőség hatását vizsgáljuk

Extrémumok stacionárius sorozatokra

- Ha csak gyenge összefüggőség áll fenn, a maximumok határeloszlása továbbra is GEV
- Ehhez elég az alábbi feltétel ($D(u_n)$):

$$\left|P\Big(\max_{\{i\in A_1\cup A_2\}} X_i < u_n\Big) - P\Big(\max_{\{i\in A_1\}} X_i < u_n\Big) P\Big(\max_{\{i\in A_2\}} X_i < u_n\Big)\right| \leq \alpha(n,l)$$

- ahol $A_1 = \{i_1, \ldots, i_p\}$ és $A_2 = \{j_1, \ldots, j_q\}$, $1 \leq i_1 \leq \cdots < i_p < j_1 < \cdots < j_q$ és $j_1 - i_p > l$,$\alpha(n,l) \to 0$, ha $n \to \infty$ megfelelő $l = l_n = o(n)$ sorozatra.

Tulajdonságok

- Független azonos eloszlású sorozatra minden u_n-re teljesül a $D(u_n)$ feltétel
- Ha normális eloszlású a sorozat, akkor elég az autokorrelációkra a $\rho_n \log(n) \to 0$ feltétel
- Ez gyengébb, mint az általában szokásos gyenge keverés
- Ha teljesül $u_n = a_n z + b_n$-re, akkor a normalizált maximumok határeloszlása szintén GEV (Leadbetter, 1974)
- De: a paraméterek eltérhetnek a független azonos eloszlású esetre adódótól

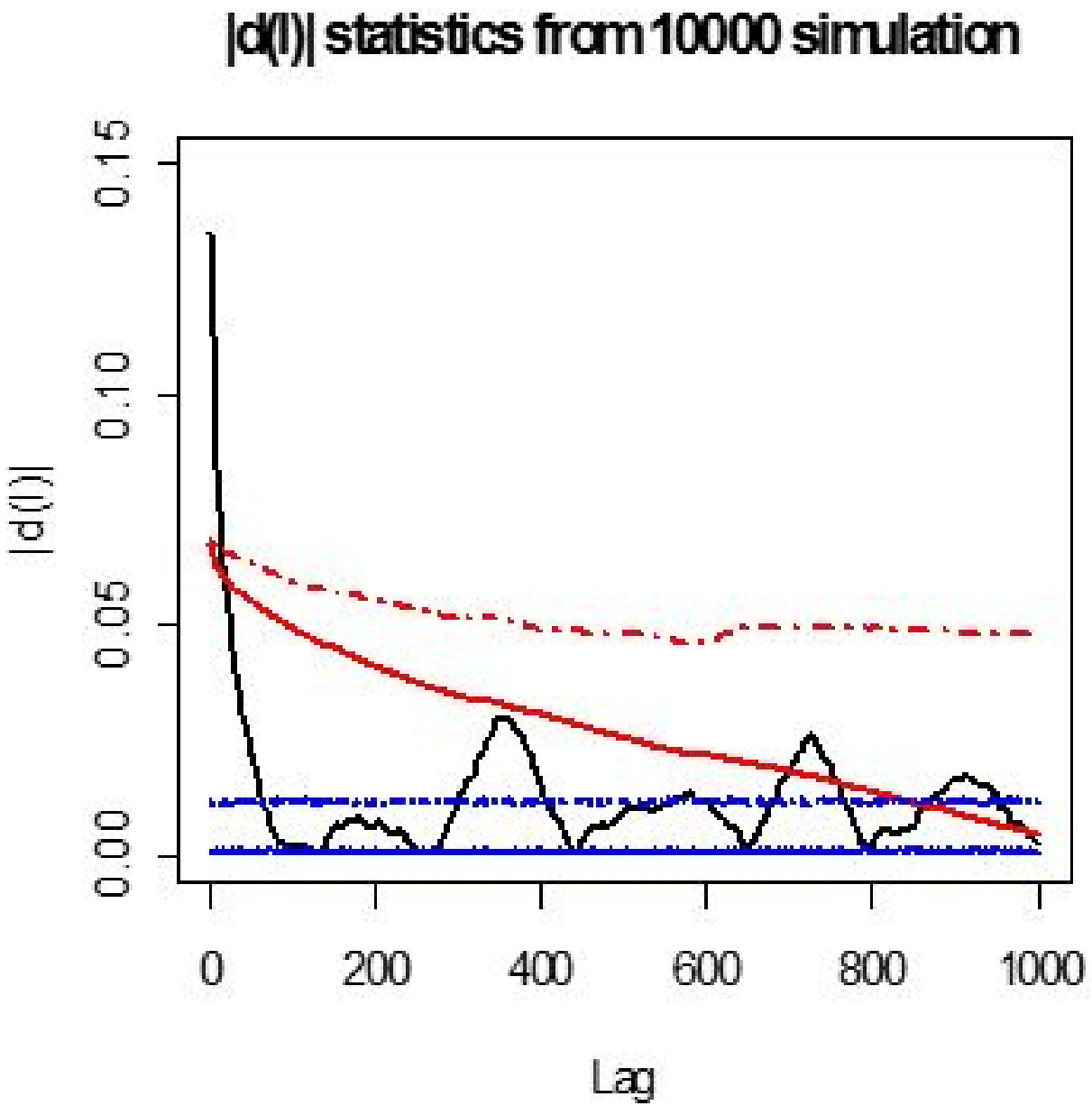
Hogyan ellenőrizzük a D feltételt?

- Legyen $p = 1$ és $q = 1$ a $D(u)$ definíciójában és válasszunk egy magas u küszöböt
- Számoljuk ki a

$$d(l) = \frac{1}{n-l} \sum_{i=1}^{n-l} I\{\max(X_i, X_{i+l}) < u\} - \left(\frac{1}{n} \sum_{i=1}^{n} I\{X_i < u\}\right)^2$$

  értéket $l = 1, \ldots, 1000$ -re
- Ábrázoljuk $|d(l)|$ -et l függvényében és hasonlítsuk össze ismert sorozatokra adódó $|d(l)|$ értékekkel

Alkalmazás: vízállás-adatok



- Folytonos vonal: becslések
- Szaggatott vonal: 95%-os konf. int
- Kék: iid N(0;1)
- Piros: AR(1)
- Az adatok (fekete vonal) a konfidencia intervallumon belül vannak (gyenge a függésük a küszöbtől), tehát elfogadható a *D* feltétel

Az extremális index

- Ha az eredeti $X_1, X_2, \ldots, X_n$ sorozathoz képezzük az $X_1^*, X_2^*, \ldots, X_n^*$ független, azonos eloszlású sorozatot és feltesszük, hogy
- $[\max(X_1^*, X_2^*, \ldots, X_n^*) - a_n]/b_n$
- $[\max(X_1^*, X_2^*, \ldots, X_n^*) - a_n]/b_n \to G_1$ és
- $[\max(X_1, X_2, \ldots, X_n) - a_n]/b_n \to G_2$ akkor a $D(u_n)$ feltétel esetén $G_1^\theta = G_2$
- Tulajdonságok:
  - $0 < \theta \leq 1$
  - Az alakparaméter ugyanaz a két esetben
  - Független sorozatra $\theta = 1$, de a megfordítás nem igaz

Becslés

- θ becsülhető például abból a tulajdonságból, hogy θ az átlagos (küszöb feletti) klaszterméret reciproka
- Másik lehetőség: futam-módszer
- De nem könnyű a becslés: különböző küszöbökre és becslési módszerekre igencsak eltérő értékek adódhatnak

Blokkmódszer

- Legyen n megfigyelésünk
- Osszuk fel k_n csoportra, mindegyik r_n nagyságú
- N_n a küszöböt meghaladó megfigyelések száma
- Z_n azon blokkok száma, amelyekben van küszöb fölötti megfigyelés
- Becslések:

$$\overline{F}(u_n) \sim \frac{N_n}{n}, \ \overline{F}_{r_n}(u_n) \sim \frac{Z_n}{k_n}$$

- Innen:

$$\overline{F}_{r_n}(u) \sim \overline{F}^{\theta r_n}(u) \text{ és így } \hat{\theta} = \frac{\log\left(1 - \frac{Z_n}{k_n}\right)}{r_n \log\left(1 - \frac{N_n}{n}\right)}$$

Hivatkozások

- Rachev, S.T.(ed): Handbook of Heavy Tailed Distributions (2003)
- Royston, P.: Profile likelihood for estimation and confidence intervals (2007)
- Embrechts, P., Klüppelberg, K., Mikosch, T.: Modelling Extremal Events. for Insurance and Finance (2000)
- Smith. R.L.: Maximum likelihood in a class of nonregular cases. Biometrika, 1985.
- Profil likelihood: `http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture11.htm#marginal`
-
- Zempléni, A.: Goodness of fit for generalized extreme value distributions (1991).
- Coles: Intro. to statistical modelling of extreme values (2001)
- Leadbetter, M.R. and Rootzen, H.: Extremal Theory for Stochastic Processes, 1988.
- Smith, R.L. and Weissman, I.: Estimating the extremal index,