3. STATISZTIKA GYAKORLAT: Elégséges statisztikák és ML becslések

1. Legyen $X_1, \dots, X_n$ fae. minta az $f_\theta(x) = 2\theta x(1-x^2)^{\theta-1}$ (ha $0 < x < 1$, különben 0) sfv.-el definiált eloszlásból.

   (a) Keressünk elégséges statisztikát az ismeretlen $\theta > 0$ paraméterre!

   (b) Keressünk ML-becslést az ismeretlen $\theta > 0$ paraméterre!

   (c) Exponenciális eloszláscsaládban vagyunk-e? Ha igen, alkalmazzuk az ott tanultakat!

2. Legyen $X_1, \dots, X_n$ fae. minta az $f_\theta(x) = \frac{1.5x^2}{\theta^3}$ (ha $-\theta \le x \le \theta$, különben 0) sfv.-el definiált eloszlásból.

   (a) Keressünk elégséges statisztikát az ismeretlen $\theta > 0$ pareméterre!

   (b) Keressünk ML-becslést az ismeretlen $\theta > 0$ pareméterre!

   (c) Exponenciális eloszláscsaládban vagyunk-e? Ha igen, alkalmazzuk az ott tanultakat!

3. A következő két kísérletet végezték el annak érdekében, hogy megállapítsák egy adott párt népszerűségét:

   (a) addig kérdezték véletlenszerűen kiválasztva az embereket, amíg 10 olyat nem találtak, aki az adott pártra szavazna. Azt tapasztalták, hogy ehhez 1000 embert kellett megkérdezni;

   (b) Véletlenszerűen megkérdeztek 1000 embert és azt találták, hogy közülük 10 választaná az adott pártot.

   Mutassuk meg, hogy mind a két kísérlet ugyanahhoz a maximum likelihood becsléshez vezet!

4. Legyen $X_1, \dots, X_n$ n megfigyelés az $f(x) = 0.5e^{-|x-\vartheta|}$ sfv-ű eloszlásból. Adjunk maximum likelihood becslést ϑ-ra!

5. Tégla alakú „kockával" dobunk, melynek élhosszai 1,1,θ. Egy adott oldalra esés val.sége arányos az oldal területével. Legyenek a négyzet alapú hasáb lapjai úgy megszámozva, hogy az egységnyi területű lapokon van '1' és '6', míg a '2,3,4,5' számok a θ területű lapokon vannak. n dobás után azt tapasztaljuk, hogy az 'i' kimenetel gyakorisága x_i $(i = 1, \dots, 6)$. Ennek alapján adjunk ML becslást θ-ra és keressünk elégséges statisztikát is!

6. Az ún. α modell (l. pl. Csiszár,V., Hussami,P., Komlós,J., Móri,T.F., Rejtő,L. and Tusnády,G. (2011), When the degree sequence is a sufficient statistic, Acta Math. Hung. 134, 45-53) a következ?. Adott egy véletlen gráf n csúccsal, melynek szomszédsági mátrixa $\mathbf{A} = (A_{ij})$. $\mathbf{A}$ diagonálisa zéró, a diagonalis feletti A_{ij}-k pedig függetlenek, és A_{ij} Bernoulli eloszlású $p_{ij} = \mathbb{P}(A_{ij} = 1)$ paraméterrel, különben $\mathbf{A}$ szimmetrikus. A modell szerint a $\frac{p_{ij}}{1-p_{ij}}$ ún. esélyhányadosokra

$$\frac{p_{ij}}{1-p_{ij}} = \alpha_i \alpha_j \quad (1 \le i < j \le n)$$

   teljesül, ahol $\alpha_1, \dots, \alpha_n$ pozitiv valós paraméterek. Adjunk meg elégséges statisztikát ezekre a paraméterekre!

Megoldások

1. (a) A likelihood fv.:

$$\begin{aligned} L_\theta(\mathbf{x}) = L_\theta(x_1,\dots,x_n) &= \prod_{i=1}^n f_\theta(x_i) \\ &= [2^n\theta^n(\prod_{i=1}^n(1-x_i)^2)^{\theta-1})]\cdot[\prod_{i=1}^n x_i] \\ &= [2^n\theta^n(\prod_{i=1}^n(1-x_i)^2)^{\theta})]\cdot[\prod_{i=1}^n \frac{x_i}{1-x_i^2}], \end{aligned}$$

ahol az első kapcsos zárójelben álló fv. $g_\theta(T(\mathbf{x}))$, míg a másodikban álló nem függ a paramétertől, ez $h(\mathbf{x})$. (Megjegyezzük, hogy $h(\mathbf{x})$-be az $I(0 \le x_1^* \le \dots \le x_n^* \le 1)$ indikátorfv-t is bevehettük volna, ez sem függ a paramétertől.) Így a $\prod_{i=1}^n(1-X_i^2)$ statisztika elégséges lesz, de $\sum_{i=1}^n \ln(1-X_i^2)$ is az lenne.

(b) A log-likelihood fv. θ szerint deriválható, a derivált gyökhelye lehet az ML-becslés:

$$\frac{\partial}{\partial\theta}\ln L_\theta(\mathbf{x}) = \frac{\partial}{\partial\theta}\ln g_\theta(T(\mathbf{x})) = \frac{n}{\theta} + \sum_{i=1}^n \ln(1-X_i^2) = 0,$$

ahonnan

$$\hat\theta = -\frac{n}{\sum_{i=1}^n \ln(1-X_i^2)}.$$

Könnyen látható, hogy $\hat\theta$ 1 val.séggel pozitív, és ez az egyetlen lokális, sőt globális maximum; továbbá $\hat\theta$ egy elégséges statisztika fv-e.

(c) Mivel

$$f_\theta(x) = 2\theta x(1-x^2)^{\theta-1} = 2\theta e^{\theta\ln(1-x^2)}\cdot\frac{x}{1-x^2}I_{(0,1)},$$

ezért exp. eo. családban vagyunk és $\sum_{i=1}^n \ln(1-X_i^2)$ a kanonikus elégséges stat. Mivel a paramétertér tartalmaz 1-dim. téglát, ez teljes, és min. elégséges is.

2. (a) A likelihood fv.:

$$L_\theta(\mathbf{x}) = \prod_{i=1}^n \frac{1.5x_i^2}{\theta^3} I(-\theta \le x_i \le \theta) = \left[\frac{1}{\theta^{3n}} I(\max_i |x_i| \le \theta)\right]\cdot\left[\prod_{i=1}^n 1.5x_i^2\right],$$

ahol az első kapcsos zárójelben álló fv. $g_\theta(\mathbf{x})$, míg a másodikban álló nem függ a paramétertől, ez $h(\mathbf{x})$. Így $\max_i |X_i|$ elégséges.

(b) A likelihood fv. most nem deriválható θ szerint (ez általában így van, ha a sfv. tartója függ a paramétertől), viszont látható, hogy θ csökkenésével monoton nő mindaddig, amíg $\theta \ge \max_i |x_i|$. Így $\hat\theta = \max_i |X_i|$ ML-becslés.

(c) Mivel az eloszlás tartója függ a paramétertől, nem vagyunk exp. eo. családban.

3. (a) A minta alapján Y :=hány embert kell megkérdezni, amíg a 10. támogatót megtaláljuk? $Y \sim \mathcal{N}_{10}(\theta)$ negatív binomiális eloszlású. A likelihood fv. annak a szituációnak a val.sége, hogy $Y = 1000$:

$$\mathbb{P}(Y = 1000) = \binom{999}{9}\theta^{10}(1-\theta)^{990}.$$

(b) Most a minta alapján X :=hány ember támogatja 1000 közül a pártot? Itt $X \sim \mathcal{B}_{1000}(\theta)$ binomiális eloszlású. A likelihood fv. annak a szituációnak a val.sége, hogy $X = 10$:

$$\mathbb{P}(X = 10) = \binom{1000}{10}\theta^{10}(1-\theta)^{990}.$$

Szemmel láthatóan a két likelihood fv. csak egy konstansban (binomiális együttható) különbözik, így maximumhelyük ugyanaz. Keressük ezt meg logaritmálással. Mindkát esetben a maximalizálandó log-likelihood fv:

$$c + 10\ln(\theta) + 990\ln(1-\theta),$$

melynek θ szerinti deriváltja:

$$\frac{10}{\theta} - \frac{990}{1-\theta} = 0,$$

ami ekvivalens azzal, hogy

$$10(1-\theta) - 990\theta = 0,$$

ui. $0 < \theta < 1$. Innen $\hat{\theta} = \frac{10}{1000} = 0.01$ az ML becslés (a második derivált alapján tényleg maximum hely). Nem meglepő, hogy a relatív gyakoriságot kapjuk.

4. A maximalizálandó likelihood függvény

$$L_\theta(\mathbf{x}) = 0.5^n \cdot e^{-\sum_{i=1}^n |\vartheta - x_i|},$$

a log-likelihood fv pedig

$$\ln L_\theta(\mathbf{x}) = n\ln(0.5) - \sum_{i=1}^n |\vartheta - x_i|.$$

Az első tag irreleváns, a második $-\sum_{i=1}^n |\vartheta - x_i|$, melynek maximumhelye ugyanaz, mint $\sum_{i=1}^n |\vartheta - x_i|$ minimumhelye, ami $\hat{\vartheta} = m_n$, ahol m_n jelöli az $x_1, \ldots, x_n$ számsokaság mediánját. (Ez $n = 2k$ esetén nem egyértelmű: ekkor az x_k^*, x_{k+1}^* intervallum tetszőleges pontja megoldása a szélsőérték feladatnak).

- Az eredmény például abból az okoskodásból is adódik, hogy log-likelihood függvényünk folytonos és szakaszonként deriválható; a derivált konstans minden (x_m^*, x_{m+1}^*) intervallumon és a konstansok fogyó sorozatot alkotnak. A 0 értéket éppen a fentiekben megadott helyeken veszi fel (illetve a páratlan esetben „ugorja át" a függvény).

- Másik megoldás, hogy ha ϑ egyik oldalán több x_i van, mint a másikon, akkor az $-\sum_{i=1}^{n}|\vartheta - x_i|$ távolságösszeg nőni fog, ha a medián(ok) felé mozdulunk el.

5. Az '1,6' oldalakra esés val.sége $\frac{1}{2+4\theta}$, a többi oldalara esés val.sége $\frac{\theta}{2+4\theta}$. Ezért a likelihood fv:

$$L_\theta(\mathbf{x}) = \left(\frac{1}{2+4\theta}\right)^{x_1}\cdot\left(\frac{1}{2+4\theta}\right)^{x_6}\cdot\prod_{i=2}^{5}\left(\frac{\theta}{2+4\theta}\right)^{x_i} = \left(\frac{1}{2+4\theta}\right)^{z}\cdot\left(\frac{\theta}{2+4\theta}\right)^{y},$$

ahol $y = \sum_{i=2}^{5} x_i$ és $z = x_1 + x_6$. Nyilván $y + z = n$ és bármelyikük elégséges stat. θ-ra. A log-likelihood fv:

$$\ln L_\theta(\mathbf{x}) = z\ln\frac{1}{2+4\theta} + y\ln\frac{\theta}{2+4\theta} = -z\ln(2+4\theta) + y\ln(\theta) - y\ln(2+4\theta) = -n\ln(2+4\theta) + y\ln(\theta).$$

Ezért a likelihood egyenlet:

$$\frac{\partial \ln L_\theta(\mathbf{x})}{\partial\theta} = \frac{-4n}{2+4\theta} + \frac{y}{\theta} = 0,$$

ahonnan

$$\hat{\theta} = \frac{2y}{4(n-y)} = \frac{y}{2(n-y)} = \frac{n-z}{2z}$$

az ML becslés. (Ha ez 1, azaz a tégla oldalai egyenlőek, akkor $n = 3z$, $y = 2z$, $n = y + z$ és x_i-k közel azonosak.)

6. Könnyen látható, hogy

$$p_{ij} = \frac{\alpha_i\alpha_j}{1+\alpha_i\alpha_j} \quad \text{és} \quad 1 - p_{ij} = \frac{1}{1+\alpha_i\alpha_j}.$$

Úgy tűnhet, hogy egyelemű mintánk van, azonban itt az élek a független mintaelemek (n rögzített). Jelölje $\underline{D} = (D_1, \ldots, D_n)$ a fokszámsorozatot, ahol $D_i = \sum_{j=1}^{n} A_{ij}$ $(i = 1, \ldots n)$. Belátjuk, hogy ez elégseges az $\underline{\alpha} = (\alpha_1, \ldots, \alpha_n)$ paramétervektorra. $\mathbf{A}$ szimmetriája és a $0^0 = 1$ konvenció miatt

$$\begin{aligned}
L_{\underline{\alpha}}(\mathbf{A}) &= \prod_{i=1}^{n-1}\prod_{j=i+1}^{n} p_{ij}^{A_{ij}}(1-p_{ij})^{1-A_{ij}} = \left\{\prod_{i=1}^{n}\prod_{j=1}^{n} p_{ij}^{A_{ij}}(1-p_{ij})^{1-A_{ij}}\right\}^{1/2} \\
&= \left\{\prod_{i=1}^{n}\prod_{j=1}^{n}\left(\frac{p_{ij}}{1-p_{ij}}\right)^{A_{ij}}\prod_{i=1}^{n}\prod_{j=1}^{n}(1-p_{ij})\right\}^{1/2} \\
&= \left\{\prod_{i=1}^{n}\alpha_i^{\sum_{j=1}^{n}A_{ij}}\prod_{j=1}^{n}\alpha_j^{\sum_{i=1}^{n}A_{ij}}\prod_{i\neq j}(1-p_{ij})\right\}^{1/2} \\
&= \left\{\prod_{i\neq j}\frac{1}{1+\alpha_i\alpha_j}\right\}^{1/2}\left\{\prod_{i=1}^{n}\alpha_i^{D_i}\prod_{j=1}^{n}\alpha_j^{D_j}\right\}^{1/2} \\
&= \left\{\prod_{i<j}\frac{1}{1+\alpha_i\alpha_j}\right\}\left\{\prod_{i=1}^{n}\alpha_i^{D_i}\right\} = C_{\underline{\alpha}}\times\prod_{i=1}^{n}\alpha_i^{D_i}.
\end{aligned}$$

Itt kihasználtuk, hogy $A_{ij} = A_{ji}$, $p_{ij} = p_{ji}$ $(i < j)$ és $A_{ii} = 0$, $p_{ii} = 0$ $(i = 1, \ldots, n)$. $C_{\underline{\alpha}} = \prod_{i<j} \frac{1}{1+\alpha_i \alpha_j}$ csak az $\underline{\alpha}$ paramétertől függ, és a likelihood fv. az A_{ij} mintaelemektől csak D_i-ken keresztül függ. Így $\underline{D}$ elégséges. A csak mintaelemektől függő tényező itt 1, amiből következik, hogy a mátrixelemek együttes eloszlasa, feltéve a fokszámsorozatot, egyenletes. Azaz adott fokszámsorozat esetén ezzel a modellel generálhatunk véletlen gráfokat.

Az ML-becsléshez egy olyan n egyenletből álló likelihood egyenlet rendszert kell megoldanunk, amire csak numerikus iteráció létezik. A fenti cikkben belátják, hogy ez konvergens és az ML becslés pontosan akkor létezik és egyértelmű, ha a mintából számolt fokszámsorozat az Erdős–Gallai feltételeknek eleget tevő politóp belsejébe esik (ha a határára esik, akkor nincs megoldás, ami nem 0 val.ségű, pl. az ún. threshold gráfok ilyenek).

Megjegyezzük, hogy az Erdős–Rényi véletlen gráf ennek az a specialis esete, melyben az összes α, és így p_{ij} megegyezik. Véletlen, téglalap alakú 0-1 mátrixokra is általánosítható a módszer, l. Rasch modell és Bolla, M., Elbanna, A. (2015), Estimating parameters of a probabilistic heterogeneous block model via the EM algorithm, Journal of Probability and Statistics. Article 657965.

Az elégséges statisztikáknál nézzük meg azt is, hogy exp. eloszláscsaládba tartozik-e az eloszlás, és ha igen, akkor keressük így is meg az elégséges statisztikát, és állapítsuk meg, hogy teljes-e, minimális elégséges-e. Győződjünk meg, hogy az ML-becslés a minimális elégséges stat. fv.-e.