Idősorok és többdimenziós statisztika

1. előadás

Zempléni András

2019. szeptember 10.

Tematika

- Markov láncok
- Speciális sztochasztikus folyamatok:
  - Poisson folyamat
  - Születési-halálozási folyamatok
  - Felújítási folyamatok
  - Brown mozgás
- Stacionárius folyamatok: autokorreláció, parciális autokorreláció (segédeszköz: feltételes várható érték), spektrálelőállítás
  - Lineáris folyamatok: $ARMA(p, q)$ modellek
  - Példa nemlineáris modellekre: ARCH(1) folyamat
- Többdimenziós normális eloszlás és paraméterbecslése
- A statisztikából tanultak áttekintése, kiegészítése ("big data")

Tematika/2

- Idősorok statisztikája
  - A várható érték becslése
  - Az autokovariancia becslése
  - Az ARMA folyamatok paraméterbecslése, rendszelekció, Akaike-féle információs kritérium
- Többdimenziós statisztika elemei
  - Főkomponens analízis
  - Faktoranalízis
  - Klaszteranalízis
  - További, modern módszerek (SVM, döntési fa)
- Sok példával, kevés bizonyítással
- Az eszköz: R, de python-t is próbálok mutatni
- Vizsga: írásbeli: elméleti kérdések, példák (R: 10-15%), kevés bizonyítás

Markov láncok

- Olyan sztochasztikus folyamat ($X_t : \Omega \to \mathbb{R}$ val.változó, $t \in T$), ahol az állapottér (az X_t értékkészlete) véges vagy megszámlálható és a jelen meghatározza a következő időpontban az eloszlást (nem kell a múlt ismerete).
- Formális def.:

$$\mathbb{P}(X_{n+1} = i \mid X_0 = i_0, X_1 = i_1 \ldots, X_n = i_n) = \mathbb{P}(X_{n+1} = i \mid X_n = i_n)$$

- Markov:
  - Véletlen bolyongás
  - i.i.d. sorozat
- nem Markov:
  - Visszatevés nélküli mintavétel
  - A napi középhőmérséklet
  - Bolyongásnál az adott időpontig a maximum értéke

Alaptulajdonságok, példák

- Feltesszük, hogy a Markov-lánc stacionárius, azaz hogy $\mathbb{P}(X_m = j \mid X_n = i)$ csak $m - n$-től függ (jel.: $p_{ij}^{(m-n)}$, $m - n$ lépéses átmenetvszg.)
- $p_{ij} := p_{ij}^{(1)}$ 1 lépéses átmenetvszg
- $\sum_{j \in I} p_{ij} = 1$
- Példák:
  - (szimmetrikus) bolyongás
  - Általánosabban: $S_n = X_1 + ... + X_n$, ahol X_n független, azonos (diszkrét) eloszlású val. változó-sorozat ($n = 1, 2, ..., S_0 = 0$)
  - Elnyelő falas bolyongás (addig játszunk, míg valamelyikünk el nem veszti az összes pénzét; S_n a pénzünk n játék után)
  - Visszaverő falas bolyongás: részecske mozgása egy szakasz mentén
  - Genetikai modell: N fix elemszámú populáció, kétfajta gén. Állapotok: az "a" típus gyakorisága. Átmenet: binomiális eloszlás szerint

Átmenetvalószínűségek

- A véges dimenziós eloszlások:

$$\begin{aligned}&\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) =\\ &= \mathbb{P}(X_0 = i_0)\mathbb{P}(X_1 = i_1 | X_0 = i_0) \cdots \mathbb{P}(X_n = i_n | X_0 = i_0, \ldots, X_{n-1}\\ &= i_{n-1}) = \mu_{i_0} p_{i_0 i_1} \cdots p_{i_{n-1} i_n}.\end{aligned}$$

- Legyenek $p_{ij}^{(n)} = \mathbb{P}(X_{n+m} = j | X_m = i)$ az n-lépéses átmenetvalószínűségek. Ezekre teljesül a Chapman-Kolmogorov egyenlőség:

$$p_{ij}^{(n+m)} = \sum_{k \in I} p_{ik}^{(n)} p_{kj}^{(m)}.$$

  Azaz $p_{ij}^{(n)} = (p^n)_{ij}$ (mátrixhatvány).
- Kérdés: hova tart X_n eloszlása?

Az állapotok osztályozása

- Az i állapotból *elérhető* j ($i \to j$), ha van olyan $n \geq 0$, hogy $p_{ij}^{(n)} > 0$. Ez reflexív ($p_{ii}^{(0)} = 1$) és tranzitív
- i és j *közlekednek*, ha $i \to j$ és $j \to i$. Ez ekvivalenciareláció, tehát osztályokra bontja az állapotteret
- A Markov-lánc *irreducibilis*, ha egyetlen osztályból áll.
- Az i állapot *lényeges*, ha $i \to j$ esetén $j \to i$ is teljesül.
- A lényegesség osztálytulajdonság
- Az $\{n > 0 : p_{ii}^{(n)} > 0\}$ halmaz legnagyobb közös osztója az i *periódusa*, jelölése $d(i)$. Ha a halmaz üres, akkor a periódust nem értelmezzük. Ha $d(i) = 1$, akkor az állapot *aperiodikus*
- Egy osztály minden állapotának ugyanannyi a periódusa
- Legtöbbször elég irreducibilis és aperiodikus Markov-láncokat vizsgálni

Visszatérőség

- Legyen

  $$f_{ij}^{(0)} = 0, \quad f_{ij}^{(n)} = \mathbb{P}(X_n = j, X_k \neq j : k = 1, 2, \ldots, n-1 \mid X_0 = i)$$

  $n \geq 1$ annak valószínűsége, hogy az i állapotból indulva a lánc először az n-edik lépésben ér el a j állapotba.
- $f_{ij}^* = \sum_{n=1}^{\infty} f_{ij}^{(n)}$ annak valószínűsége, hogy az i állapotból indulva, a lánc előbb-utóbb elér a j állapotba
- Az i állapot *visszatérő vagy rekurrens*, ha $f_{ii}^* = 1$, egyébként pedig *átmeneti vagy tranziens*. Ha i visszatérő, akkor az átlagos visszatérési idő $m_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$. Az i állapot *pozitív rekurrens*, ha $m_i < \infty$, ha pedig $m_i = \infty$, akkor *nulla rekurrens*.

Tulajdonságok, példák

- Legyen $i \neq j$. $i \rightarrow j$ akkor és csak akkor, ha $f_{ij}^* > 0$.
- A nem lényeges állapotok tranziensek (fordítva azonban nem feltétlenül igaz).
- Az egydimenziós bolyongás akkor és csak akkor visszatérő, ha szimmetrikus.
- Ha i rekurrens állapot d periódussal, akkor $\lim_{n\to\infty} p_{ii}^{(nd)} = \frac{d}{m_i}$.
- Legyen i, j két tetszőleges állapot, és jelölje j periódusát d. Ekkor $r = 1, 2, \ldots, d$ esetén

$$\lim_{n\to\infty} p_{ij}^{(nd+r)} = f_{ij}^*(r)\frac{d}{m_j},$$

ahol $f_{ij}^*(r)$ annak a valószínűsége, hogy d-vel osztva r maradékot adó lépésben jut el a lánc i-ből j-be. Így $\sum_{r=1}^{d} f_{ij}^*(r) = f_{ij}^*$. (Tranziens állapotra legyen $m_j = \infty$.)