Idősorok és többdimenziós statisztika

5. előadás

Zempléni András

2019.10.08

Állítás

Az autokovariancia-függvény pozitív szemidefinit, azaz
$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j R(|t_i-t_j|)\geq 0$ *minden* $t_1,\dots,t_n$ *időpontra, és* $\alpha_1,\dots,\alpha_n$ *valós számokra.*

Definíció

Legyenek X, Y, Z valószínűségi változók, amelyekre X és Y Z-szerinti **parciális kovarianciája** $cov(X,Y|Z)=E((X-E(X|Z))(Y-E(Y|Z)))$. X és Y Z-szerinti **parciális korrelációja** pedig

$$\varrho(X,Y|Z)=\frac{cov(X,Y|Z)}{(cov(X,X|Z)\cdot cov(Y,Y|Z))^{1/2}}.$$

Definíció

Egy idősor **parciális autokorreláció-függvénye** $\varrho_k=\varrho(X_{t+k},X_t|X_{t+1},\dots,X_{t+k-1})$

Zajfolyamatok

Definíció

Az $\varepsilon(t)$ folyamat **független értékű zaj**, ha $\varepsilon(t)$-k független, azonos eloszlású valószínűségi változók, 0 várható értékkel.

A 0 várható értéket nem mindig követeljük meg.

Definíció

Az $\varepsilon(t)$ folyamat **fehér zaj**, ha $\varepsilon(t)$-k azonos eloszlásúak minden t-re és korrelálatlanok (de nem feltétlen függetlenek), és $E\varepsilon(t)=0$.

A fehér zaj autokovariancia-függvénye $R(0)=\sigma^2$, $R(\tau)=0 \quad (\tau\geq 1)$.

A független értékű zaj erősen, a fehér zaj gyengén (másodrendben) stacionárius.

Autoregressziós folyamatok

Definíció

Elsőrendű autoregresszió, AR(1):
Az elsőrendű autoregresszió egy olyan idősor, amely kielégíti a

$$X(t)=\alpha X(t-1)+\sigma_\varepsilon\cdot\varepsilon(t)$$

$$X(0)=X_0$$

(dinamikai) egyenletet, melyben $\varepsilon(t)$ független értékű zaj, 1 szórással, $\alpha\in\mathbb{R},\sigma_\varepsilon>0$.

Definíció

Egy idősor **oksági megoldása** a dinamikai egyenletének, ha a folyamat jelene független a generáló zaj jövőjétől, azaz

$$X(t),\ \text{és}\ \{\varepsilon(t+1),\varepsilon(t+2)\dots\}$$

teljesen függetlenek.

Általában $EX(t) = 0$, de nem mindig, mivel a zaj 0 várható értékét nem mindig követeljük meg.

Amennyiben oksági a megoldás, és így $X(t-1)$ és $\varepsilon(t)$ függetlenek, valamint véges a szórás, akkor

$$D^2X(t) = \alpha^2 D^2X(t-1) + \sigma_\varepsilon^2 D^2\varepsilon(t)$$

(a függetlenség miatt a szórásnégyzetek összeadódnak).

Ha létezik stacionárius megoldás, akkor $\sigma_X^2 = D^2X(t) = D^2X(t-1)$-ből következik, hogy

$$\sigma_X^2 = \alpha^2\sigma_X^2 + \sigma_\varepsilon^2,$$

azaz

$$\sigma_X^2 = \frac{\sigma_\varepsilon^2}{1-\alpha^2} > 0.$$

Így $|\alpha| \geq 1$ esetén nincs stacionárius megoldás (nem lenne véges vagy nem lenne pozitív a szórásnégyzet).

Az AR(1) autokovariancia függvénye

- Az elsőrendű autoregressziós folyamat autokovariancia függvénye:

$$R(\tau) = cov(X(t), X(t+\tau)) =$$
$$= cov(X(t), \alpha \cdot X(t+\tau-1) + \sigma_\varepsilon \cdot \varepsilon(t+\tau)) =$$
$$= \alpha \cdot cov(X(t), X(t+\tau-1)) = \alpha \cdot R(\tau-1),$$

ami kielégíti a zaj nélküli rekurziót. Mivel $R(0) = D^2X(t) = \sigma_X^2$, így

$$R(\tau) = \sigma_X^2 \cdot \alpha^\tau = \frac{\alpha^\tau}{1-\alpha^2} \cdot \sigma_\varepsilon^2, \quad \text{és} \quad r(\tau) = \frac{R(\tau)}{R(0)} = \alpha^\tau\ .$$

- Az AR(1) parciális autokovariancia függvénye
A parciális autokorreláció-függvény (PACF):

$$\varrho(\tau) = \begin{cases} \alpha & \tau = 1 \\ 0 & \tau \geq 2\ . \end{cases}$$

Ehhez: feltételes várható érték

- Adott X, Y valószínűségi változók, amelyek között van összefüggés
- Szeretnénk Y-ból kinyerni minden X-re vonatkozó információt
- Az Y-ban lévő információt Y függvényei jelentik $\Leftrightarrow$ azt a $g(Y)$ valószínűségi változót szeretnénk meghatározni, amely leginkább hasonlít X-re, legközelebb van X-hez.
- A feladat általános megoldása egy alkalmas térben: nézzük a $g(Y)$-ok által kifeszített alteret és megkeressük az X merőleges vetületét erre az altérre $\Rightarrow$ ez lesz az X-hez legközelebbi $g(Y)$ alakú valószínűségi változó.
- Ez a vetület X-nek Y szerinti feltételes várható értéke: $g(Y) = E(X|Y)$

Feltételes várható érték

Definíció: Ha X diszkrét valószínűségi változó és $P(B) > 0$, akkor X feltételes várható értéke a B feltétel mellett:
$X \sim \begin{matrix} x_1, x_2, \ldots \\ p_1, p_2, \ldots \end{matrix}$

$$E(X|B) = \sum_{i=1}^{\infty} x_i \cdot P(X = x_i|B)$$

Valószínűségi változó szerinti feltételes várható érték
X, Y diszkrét valószínűségi változó $X = \begin{matrix} x_1, x_2, \ldots \\ p_1, p_2, \ldots \end{matrix}$,

$Y = \begin{matrix} y_1, y_2, \ldots \\ q_1, q_2, \ldots \end{matrix}$, $EX < \infty$
$Y = y_i$ esemény és $P(Y = y_i) > 0$, így $E(X|Y = y_i)$ definiált a fentiek szerint.
Definíció: $E(X|Y = y) = g(y)$ egy függvény az Y értékkészletén az $y_i \to E(X|Y = y_i)$ hozzárendelés szerint.

Valószínűségi változó szerinti feltételes várható érték

Definíció: Ha $g(y) = E(X|Y = y)$ az X feltételes várható értéke az $Y = y$ feltétel mellett, akkor a $g(Y) = E(X|Y)$ az X feltételes várható értéke az Y feltétel (vagy adott Y) mellett. $E(X|Y)$ valószínűségi változó.
Megjegyzés: Ha $Y(\omega) = y_i$, akkor $E(X|Y)(\omega) = g(y_i) = g(Y(\omega))$
Tulajdonságok:

- Lineáris : $E(c \cdot X_1 + X_2|Y) = c \cdot E(X_1|Y) + E(X_2|Y)$
- Ha $X \geq Z \Rightarrow E(X|Y) \geq E(Z|Y)$
- Ha X, Y függetlenek: $E(X|Y) = EX$
- Ha $X = h(Y)$, akkor $E(X|Y = y_j) = h(y_j)$ ezért $E(X|Y = y) = h(y) \Rightarrow E(X|Y) = h(Y) = X$
- $E(E(X|Y)|Y) = E(g(Y)|Y) = g(Y) = E(X|Y)$

A feltételes várható érték kiszámítása

Az abszolút folytonos eset: legyen X, Y együttes sfv.-e $f(x, y)$.
X, ill. Y sfv.-e: $\int_{-\infty}^{+\infty} f(x, y)\mathrm{d}y = f_X(x)$, $\int_{-\infty}^{+\infty} f(x, y)\mathrm{d}x = f_Y(y)$
Definíció: A feltételes sűrűségfüggvény:

$$f(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f(x, y)}{\int_{-\infty}^{+\infty} f(x, y)\mathrm{d}x}$$

Definíció: A feltételes eloszlásfüggvény: $F(x|y) = \int_{-\infty}^{x} f(t|y)\mathrm{d}t$
Definíció: A feltételes valószínűség y függvénye:

$$P(X \in A|Y = y) = \int_A f(t|y)\mathrm{d}t$$

Teljes valószínűség tétele

Állítás:

$$P(X \in A, Y \in B) == \int_A \int_B f(x|y)f_Y(y)\mathrm{d}y\mathrm{d}x =$$

$$\overset{\text{integrál csere}}{=} \int_B P(X \in A|Y = y)f_Y(y)\mathrm{d}y$$

Legyen $g(y) = \int_{-\infty}^{+\infty} x \cdot f(x|y)\mathrm{d}x$, akkor $g(Y) = E(X|Y)$
Következmény: Beírva $f(x|y)$ definícióját is kapjuk, hogy

$$g(y) = \int_{-\infty}^{+\infty} x \cdot \frac{f(x, y)}{f_Y(y)}\mathrm{d}x = \int_{-\infty}^{+\infty} \frac{x \cdot f(x, y)}{\int_{-\infty}^{+\infty} f(t, y)\mathrm{d}t}\mathrm{d}x$$

Az AR(1) egyenlet megoldása

Állítás

Ha $|\alpha| < 1$, $\varepsilon(t)$ független értékű zaj, továbbá $E(\varepsilon(t)^2) < \infty$, akkor az $AR(1)$ egyenletnek létezik stacionárius megoldása.

Ehhez tegyük fel, hogy $|\alpha| < 1$, és iteráljuk az AR(1) elsőrendű autoregressziós egyenletet. Nem jelöljük külön az egyenletben a zaj szórását, azt tesszük fel, hogy $\varepsilon(t)$ szórása σ_ε.

$$X(t) = \alpha X(t-1) + \varepsilon(t)$$

$$X(t) = \alpha(\alpha X(t-2) + \varepsilon(t-1)) + \varepsilon(t)$$

$$\ldots$$

$$X(t) = \alpha^{s+1} X(t-s-1) + [\alpha^s \varepsilon(t-s) + \ldots + \alpha\varepsilon(t-1) + \varepsilon(t)].$$

Átrendezve:

$$X(t) - \sum_{u=0}^{s} \alpha^u \varepsilon(t-u) = \alpha^{s+1} X(t-s-1)$$

A két oldal négyzetének várható értéke:

$$E\left(X(t)-\sum_{u=0}^{s}\alpha^u\varepsilon(t-u)\right)^2=\alpha^{2s+2}E\left(X(t-s-1)^2\right)\longrightarrow 0,$$

Ha itt $X(t)$ stacionárius folyamat, akkor $EX^2(t)=\sigma_X$ minden t-re, és így a jobboldal 0-hoz tart. Ez azt jelenti, hogy

$$\sum_{u=0}^{s}\alpha^u\varepsilon(t-u)\xrightarrow{\mathcal{L}_2}X(t),\ s\to\infty.$$

Azaz

$$X(t)=\sum_{u=0}^{\infty}\alpha^u\varepsilon(t-u)$$

$\mathcal{L}_2$ értelemben, tehát az $X(t)$ stacionárius folyamat kifejezhető a zajjal $\mathcal{L}_2$-beli határértékként.

Előrejelzés

$AR(1)$ folyamat esetén, ha a megoldás oksági, azaz a zaj jövőjétől független és 0 várható értékű, akkor

$$E(X(t)|X(t-1))=E(\alpha X(t-1)+\varepsilon(t)|X(t-1))=\alpha X(t-1).$$

Továbbá

$$E(X(t)|X(t-1))=E(X(t)|X(t-1),X(t-2),\dots)$$

Ez abból látszik, hogy a feltételhez hozzávett további múltbéli tagok nem változtatnak azon, hogy ε független a feltételtől, $\alpha X(t-1)$ pedig a feltétel függvénye.

Tehát az $AR(1)$ folyamat Markov-folyamat, legjobb előrejelzése pedig a zaj nélküli egyenlettel adódik. (Miért a feltételes várható érték adja a legjobb előrejelzést? Milyen értelemben legjobb?)

Ebből azt is megkapjuk, hogy az $AR(1)$ esetén a legjobb lineáris előrejelzés adja a legjobb előrejelzést.