Idősorok és többdimenziós statisztika

10. előadás

Zempléni András

2019.11.19

Az $ARCH(1)$ folyamat

A most következő folyamatok és általánosításaik a pénzügyi modellezésben nagyon népszerűek. Az $ARCH(1)$ folyamatot Robert F. Engle vezette be 1982-ben, később közgazdasági Nobel-díjat kapott érte. Az ARCH elnevezés az Autoregressive Conditional Heteroscedasticity rövidítése. [1]

[1] Ez azt takarja, hogy a jelenlegi hiba varianciája függ a múltbeli értékektől (általában úgy, hogy a folyamat stacionárius maradjon). a Heteroscedasticity szó alapja a görög szkedásztikosz $\sigma\kappa\varepsilon\delta\alpha\sigma\tau\iota\kappa\omega\varsigma$ szó, melynek jelentése kb. (szét)szóródni képes.

Az $ARCH(1)$ folyamat definíciója

Legyen $\varepsilon(t)$ független értékű zaj. Az $X(t)$ folyamatot az

$$X(t) = \sigma(t)\varepsilon(t)$$

egyenlettel adjuk meg, azaz a független értékű zaj egy időfüggő valószínűségi változószorosaként. A valószínűségi változóra időtől, és véletlentől függő szórásként gondolhatunk. Azonban a véletlentől való függés csak a folyamat megelőző értékén (értékein) keresztül valósul meg. Eszerint $\sigma(t)$-t feltételes szórásként értelmezhetjük, feltéve, hogy a folyamat múltját ismerjük. E szórást a

$$\sigma^2(t) = \alpha_0 + \alpha_1 X^2(t-1)$$

egyenlet határozza meg. Az egyenletben α_0, α_1 nemnegatív valós konstansok.

Kvadratikus kifejezések

A fentiek alapján a feltételes szórásnégyzet

$$D^2\left(X(t)|X(t-1)=x\right) = \alpha_0 + \alpha_1 x^2$$

az előző érték *kvadratikus* függvénye. A négyzet helyett tetszőleges más hatványfüggvényt is választhatunk, ami az előző modell általánosítása – ezt a modellt Power ARCH-nak szokás hívni.
A fentebbi két egyenletből kapjuk, hogy

$$X^2(t) = \left(\alpha_0 + \alpha_1 X^2(t-1)\right)\varepsilon^2(t),$$

de ez nem ekvivalens velük, mert pl. Gauss zajjal történő generálás mellett az egyesített egyenletnek akár nemnegatív $X(t)$ megoldása is lehet, míg az eredeti két egyenlet megoldása biztos, hogy negatív értékeket is felvesz.

Stacionaritás

Keressük a stacionárius megoldást. Hasonlóan járunk el mint az $AR(1)$-esetén. Ehhez tegyük fel, hogy létezik ilyen, és iteráljuk az egyenletet:

$$X^2(t) = \alpha_0 \cdot \varepsilon^2(t) + \alpha_1 \alpha_0 \cdot \varepsilon^2(t) \cdot \varepsilon^2(t-1) + \alpha_1^2 \cdot X^2(t-2) \cdot \varepsilon^2(t) \cdot \varepsilon^2(t-1)$$

$$\vdots$$

$$X^2(t) = \alpha_0 \cdot \sum_{j=0}^{\infty} \alpha_1^j \cdot \varepsilon^2(t) \cdot \ldots \cdot \varepsilon^2(t-j).$$

Ez utóbbi akkor írható fel így, ha $\alpha_1 < 1$, mert a maradéktagokban α_1 egyre nagyobb hatványai jelennek meg, amik így nullához tartanak, miközben $X(t)$ stacionaritása és $\varepsilon(t)$ függetlensége, 1 szórása miatt a valváltozók szorzata korlátos a maradéktagokban (pl. $\mathcal{L}_2$ norma szerint).

Stacionaritás/2

Ha az összegzés és a várható érték felcserélhető, akkor

$$EX^2(t) = \alpha_0 \cdot \sum_{j=0}^{\infty} \alpha_1^j \cdot E\varepsilon^2(t) \cdot \ldots \cdot E\varepsilon^2(t-j) = \frac{\alpha_0}{1-\alpha_1},$$

ugyanis az $\varepsilon(t)$-k várható értéke 0, így második momentumuk a szórásnégyzetükkel egyenlő, ami 1, tehát egy egyszerű mértani sort kellett összegeznünk. Ebből látjuk, hogy $\alpha_0 = 0$ esetén $X(t)$ az azonosan 0 folyamat, ami nem túl érdekes.

Stacionaritás/3

Ha az

$$X(t) = \varepsilon(t) \cdot \sqrt{\alpha_0 \left(1 + \sum_{k=0}^{\infty} \alpha_1^{k+1} \cdot \varepsilon^2(t-1) \cdot \ldots \cdot \varepsilon^2(t-k-1)\right)} \quad (*)$$

felírásban a szumma konvergál, akkor stacionárius folyamatot állít elő, hiszen az $\eta(t) = \varepsilon(t+h)$ zaj véges dimenziós eloszlásai megegyeznek, és $(X(t_1+h), \ldots, X(t_m+h))$-t ugyanúgy állíthatjuk elő η-ból, mint $(X(t_1), \ldots, X(t_m))$-et ε-ból, tehát az eloszlásaik megegyeznek.
Bizonyítható az alábbi tétel.

Tétel

Ha $0 < \alpha_1 < 1$, akkor $(*)$ $\mathcal{L}_2$-ben konvergál, és az $ARCH(1)$ egyenlet egyértelmű, véges szórású, oksági, stacionárius megoldását adja.

Nem bizonyítjuk.

Erős stacionaritás

Megjegyzés

A most kapott megoldás tehát a gyengén stacionárius megoldás lesz.

Megjegyzés

A fenti szumma azonban 1 valószínűséggel is konvergálhat, és ekkor, mivel nincs szükség a második momentumra, a konvergencia feltétele más - adott esetben bővebb - is lehet, a zaj eloszlásától függően. Természetesen az így előállított folyamatnak $1 \leq \alpha_1$ mellett már nem lesz véges a szórásnégyzete, de megmutatható, hogy megoldás marad. Így előáll az a helyzet, hogy az erős stacionaritás létezési feltétele gyengébb, mint a gyenge stacionaritásé.

Várható érték és autokovariancia

Következmény

Az $\varepsilon(t)$ és $\sqrt{\cdot}$ tagok függetlensége miatt

$$EX(t) = E\varepsilon(t) \cdot E\sqrt{\cdot} = 0,$$

továbbá

$$D^2X(t) = \frac{\alpha_0}{1-\alpha_1}.$$

Az autokovariancia pedig

$$E\big(X(t+h)X(t)\big) = E\varepsilon(t+h) \cdot E\underbrace{\left(\sqrt{\cdot} \cdot \varepsilon(t) \cdot \sqrt{\cdot}\right)}_{\text{t+h múltja mind}} = 0,$$

azaz a stacionárius $ARCH(1)$ korrelálatlan, azonos eloszlású, 0 várható értékű, tehát **fehér zaj**.

Tulajdonságok

Az ARCH(1) azonban nem független értékű:

$$E\left(X^2(t)|X(t-1)\right) = \left[\alpha_0 + \alpha_1 X^2(t-1)\right] \cdot E\left(\varepsilon^2(t)|X(t-1)\right),$$

ahol $\varepsilon^2(t)$ és $X(t-1)$ függetlenek és $E\varepsilon^2(t) = 1$, tehát

$$E\left(X^2(t)|X(t-1)\right) = \alpha_0 + \alpha_1 X^2(t-1).$$

Ez pedig valódi valószínűségi változó, nem pusztán egy valós szám, mint ahogy az függetlenség esetén lenne. Ebből az is következik, hogy az $ARCH(1)$ nem is Gauss-eloszlású, hiszen akkor a korrelálatlanságból már a függetlenség is következne.

McLeod – Li próba a fehér zaj tulajdonságra

McLeod és Li tesztje (1983) csak Gauss folyamatra alkalmazható.

- Ha igaz a nullhipotézis, akkor a folyamat Gauss fehér zaj, ami független értékű is, de akkor a négyzete is az, így az is korrelálatlan értékű, azaz fehér zaj, és akkor erre a Ljung-Box teszt használható.
- Az adatok négyzetét veszi: $Y(t) = X(t)^2$ és ennek autokorrelációfüggvény becslését, $\hat{r}_Y(\tau)$-t használja:

$$\tilde{Q} = T \cdot (T+2) \sum_{\tau=1}^{h} \hat{r}_Y^2(\tau)/(T-\tau)$$

- Ez a tesztstatisztika is χ^2_h eloszlású, ARCH ellenhipotézisre érdemes alkalmazni

Többdimenziós statisztika: az adatok

Mérni vagy megfigyelni tudunk n-szer valamilyen $X_1, \ldots, X_k$ mennyiségeket. Ezeket a mennyiségeket valváltozónak tekintjük. A mért értékek ezek realizációi: az $x_{i,j} = X_i(\omega_j)$ valós számok.

$$\begin{array}{c} \\ esetek(cases) \end{array} \overbrace{\begin{array}{cccc} X_1, & X_2, & \ldots & X_k \\ \left\{\begin{array}{cccc} x_{1,1} & x_{1,2} & \ldots & x_{1,k} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,k} \end{array}\right. \end{array}}^{változók\ (variables)} \tag{0.1}$$

- A változók között általában van összefüggés, gyakran ez az összefüggés térbeli struktúrából származik, de ezt az egyszerű statisztikai elemzés esetén nem vesszük figyelembe.
- Az egyes esetek (= sorok) gyakran függetlenek, ez sokkal egyszerűbben elemezhető.

Többdimenziós adatok

- Ha például adott helyeken az év n egymásutáni napján mérünk hőmérsékletet és ebből származik az n eset, akkor ezek már nem független adatok és az összefüggési struktúrát, (korrelációt és azon túl) figyelembe kell venni.
- Az is lehet, hogy nem csak egy jelenséget kívánunk vizsgálni, pl. hőmérséklet mellett páratartalmat, napsütést, szélsebességet, csapadékot. Ilyenkor ezeket az azonos helyen mért értékeket egy vektorban fogjuk össze tehát az adatbázisunk kap egy harmadik dimenziót is, pl az első sor elemei maguk is vektorok:

$$\underline{x}_{1,1}, \quad \underline{x}_{1,2}, \quad \cdots \quad \underline{x}_{1,k}.$$

- Ebben az esetben már többdimenziós statisztikai módszereket kell használni.

Egyszerű jellemzők becslése

- **Várható érték** EX_j becslése

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}, \quad \text{vagy} \quad \bar{\underline{X}}_j = \frac{1}{n}\sum_{i=1}^{n} \underline{x}_{i,j}$$

- **Szórásnégyzet** $\mathcal{D}^2 X_j = \sigma_j^2$ becslése

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,j} - \bar{X}_j\right)^2 = \frac{1}{n-1}SS(X_j) = \frac{1}{n-1}SS_j$$

SS=Sum of Squares

- **Kovariancia** $\operatorname{cov}(X_\ell, X_m)$ becslése

$$\widehat{\operatorname{cov}}(X_\ell, X_m) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,\ell} - \bar{X}_\ell\right)\left(x_{i,m} - \bar{X}_m\right) = \frac{1}{n-1}SP(X_\ell, X_m) = \frac{}{n}$$

SP=Sum of Procucts
Ez torzítatlan becslés.

Egyszerű jellemzők becslése/2

- **Korrelációs együttható** $\operatorname{corr}(X_\ell, X_m)$ becslése

$$\widehat{\operatorname{corr}}(X_\ell, X_m) = \frac{SP_{\ell,m}}{\sqrt{SS_\ell \cdot SS_m}}$$

Ez már nem torzítatlan becslés, de aszimptotikusan igen.

- **Ferdeség** (Skewness) standardizált 3. momentum: $\frac{E(X_j - EX_j)^3}{\sigma_j^3}$, becslése

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,j} - \bar{X}_j\right)^3}{S_j^3}$$

- **Lapultság** (Kurtosis) 4.kumuláns/(2.kumuláns négyzete): $\frac{E(X_j - EX_j)^4}{\sigma_j^4} - 3$, becslése

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,j} - \bar{X}_j\right)^4}{S_j^4}$$

Kovariancia mátrix

Az $\underline{X} = X_1, X_2, \ldots, X_k$ valószínűségi vektorváltozó jellemzője a kovariancia mátrix, ami az egydimenziós szórásnégyzet megfelelője erre az esetre.

$$\Sigma = E\left(\underline{X} - E\underline{X}\right)\left(\underline{X} - E\underline{X}\right)^T$$

Σ diagonálisában a vektor komponenseinek szórásnégyzetei, a felső háromszögben kovarianciái állnak, és a mátrix szimmetrikus.

- **Kovariancia mátrix** becslése $\widehat{\Sigma}$
  A fentebbi formulákkal becsült szórásnégyzeteket és kovarianciákat beírjuk a $\widehat{\Sigma}$ mátrix megfelelő helyére.
- Gyakran fontos a kovariancia mátrix **sajátértékeinek** becslése, amit a $\widehat{\Sigma}$ mátrix sajátértékeivel becslünk.

Hipotézisvizsgálat valószínűségi vektorváltozóra

- A hipotézisvizsgálat valószínűségi vektorváltozóra bonyolultabb, mert a paraméterek száma jóval nagyobb, pl. egy több- (d-)dimenziós normális eloszlást várható érték vektora és kovariancia mátrixa jellemez, ez összesen $2d + \binom{d}{2}$ paraméter
  Lehetne egyesével tesztelni a paramétereket, de nézzük meg mi történne a hibákkal.
- Legyen pl. d=10. Ha a várható értékről minden egyes tesztben $\alpha = 0.05$ szinten döntünk, akkor az együttes döntésünk elsőfajú hibája így alakul:

$$\begin{aligned} P_0(\text{Legalább egy elut.}) &= 1 - P_0(\text{mindet elfogadom}) \sim \\ &\sim 1 - (1 - 0.05)^{10} = 1 - (0.95)^{10} = 0.4 \end{aligned}$$

  ami elfogadhatatlanul nagy. (P_0 a 0-hipotézis igaz volta mellett számított valószínűség.)

Egyváltozós próbák problémái

- Az egyváltozós próbák teljesen figyelmen kívül hagyják a valószínűségi vektorváltozó komponensei közötti összefüggéseket, korrelációkat.
- A többváltozós próbák azért is erősebbek, mert az egyes változókon, komponenseken fellépő kis hatások, melyek önmagukban elhanyagolhatók lennének, együttesen már lényeges, szignifikáns eltéréssé állnak össze.
- Az egyváltozós próbákkal az elfogadási tartomány csak egy tégla lehet (téglalap, téglatest, hipertégla), míg többváltozós próbák esetén ez tetszőleges alakú tartomány is lehet.

Próbák többdimenziós normális eloszlás várható érték vektorára: ismert kovariancia mátrix

- Legyen $\underline{X}_1,\ \underline{X}_2,\ \ldots\ \underline{X}_n$ $N(\mu, \Sigma)$ eloszlású független d-dimenziós minta.
- A H_0: $\{\mu = \mu_0\}$ nullhipotézist teszteljük teljes (kétoldali) alternatíva mellett.
- Az U-próba megfelelője az az eset, ha a kovariancia mátrix, Σ, ismert. A próbastatisztika az egydimenziós esetbeli $\sqrt{n} \cdot \frac{(\bar{X} - \mu_0)}{\sigma}$ analógiájára készülhetne, azonban az abban szereplő σ *szórás* megfelelőjét nem tudjuk egyértelműen megtalálni
- Szerencsére a próbastatisztika négyzetének többdimenziós analogonjával nincs ilyen probléma, és mivel maga a próbastatisztika $N(0, 1)$-eloszlású ezért négyzete χ^2_1 eloszlású tehát ennek alapján a próbastatisztika négyzetével is készíthetnénk az egydimenziós U-próbát.
- Ezt a gondolatot lehet átvinni a többdimenziós esetre.

Próbák többdimenziós normális eloszlás várható érték vektorára: ismert kovariancia mátrix

- Legyen a próbastatisztika:

$$U_d^2 = n \cdot \left(\underline{\bar{X}} - \mu_0\right)^T \Sigma^{-1} \left(\underline{\bar{X}} - \mu_0\right)$$

- Ha $\Sigma = AA^T$ és $Z = \sqrt{n} \cdot A^{-1} \left(\underline{\bar{X}} - \mu_0\right)$, akkor egyfelől Z egy d-dimenziós standard normális vektor (független $N(0, 1)$-es komponensekből áll), másfelől $U_d^2 = Z^T Z$, tehát U_d^2 χ^2_d-eloszlású.
- Próbastatisztikánkat tehát a χ^2_d eloszlás kritikus értéke ellenében kell tesztelni.
- Az alternatív hipotézis igaz volta mellett is meg lehet adni a próbastatisztika eloszlását, ez nem centrális $\chi^2_{d,\nu}$ eloszlás lesz, ahol

$$\nu = \nu_\mu = n \cdot (\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0)$$

  az u.n. nemcentralitási paraméter. Ebből az erőfüggvény számolható.

Ismeretlen kovariancia mátrix

- Ugyanúgy mint az előbb legyen $\underline{X}_1,\ \underline{X}_2,\ \ldots\ \underline{X}_n \quad N(\mu, \Sigma)$ eloszlású független d-dimenziós minta. A H_0: $\{\mu = \mu_0\}$ nullhipotézist teszteljük teljes alternatíva mellett.
- Az *ismeretlen* Σ kovariancia mátrix esete a t-próba megfelelője. A próbastatisztika az előzőnek megfelelő, csupán az előző U-próbában szereplő Σ helyére annak becslését írjuk be.

$$T^2 = n \cdot \left(\bar{\underline{X}} - \mu_0\right)^T \hat{\Sigma}^{-1} \left(\bar{\underline{X}} - \mu_0\right)$$

  ahol $\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{\underline{X}}\right) \left(X_i - \bar{\underline{X}}\right)^T$. Megmutatható, hogy $\hat{\Sigma}$ pozitív definit, p változós Wishart eloszlást követ $n-1$ szabadsági fokkal.
- A próbát Hotelling féle T^2-próbának hívják.
- A likelihood hányados próba normális eloszlás esetén a T^2-próbához vezet, ez **legerősebb** próba lesz, csakúgy mint a $t-$ ill. $F-$próbák egydimenzióban.

Ismeretlen kovariancia mátrix/2

- A t-statisztika jelentése: hány szórásnyira van egymástól a mintaátlag és a hipotetikus várható érték. A T^2-nek nincs ilyen szemléletes jelentése.
- Nagy mintaszámra van F-közelítése a próbastatisztikának

$$\frac{n-d}{d(n-1)} T^2 \sim F_{d,n-d},$$

  ami azért fontos, mert az F eloszlás sok szempontból jobban ismert mint a T^2.
- A $T^2_{d,n-1}$ ferde eloszlás, nem úgy, mint pl. a t.
- Határeloszlása is van a $T^2_{d,n-1}$-nek: $T^2_{d,n-1} \underset{n\to\infty}{\longrightarrow} \chi^2_d$ úgy, ahogy az várható, mivel a t eloszlás normális eloszláshoz tart ezért "négyzete" χ^2-hez.
- De ez a konvergencia jóval lassúbb mint a t eloszlás normális közelítése, pl. $d = 5$ esetén hasonló precizitáshoz már 100 körüli mintaelemszám kell, tehát a becsült kovariancia mátrix jóval tovább "rontja" a tesztet.