**EÖTVÖS LORÁND TUDOMÁNYEGYETEM**
TERMÉSZETTUDOMÁNYI KAR

# Idősorok osztályozása

Írta:

Budai Fruzsina Mária
Alkalmazott matematikus MSc

Témavezető:

Prőhle Tamás
Valószínűségelméleti és Statisztika Tanszék

2013. 05. 31.

# Köszönetnyilvánítás

Köszönetet szeretnék mondani mindenkinek, aki a segítségemre volt az elmúlt két évben. Első sorban a szüleimnek, akik nemcsak anyagilag támogattak, és nyújtottak biztos hátteret, hanem szellemiségükkel, kritikájukkal és persze bíztató szavaikkal mind- mind erőt adtak. Aztán természetesen a témavezetőmnek Prőhle Tamásnak, akinek áldozatos munkája és önzetlen segítsége nélkül nem készülhetett volna el a dolgozatom. Továbbá szeretném megköszönni Stark Andársnak a sok-sok segítséget, valamint szaktársamnak, Józsa Mónikának ezt a másfél évet, amit együtt tölthettünk. Illetve az összes Tanárnak, akit megismerhettem, és tudásukkal tágították a látókörömet. Végül, de nem utolsó sorban a testvéremnek, Budai Kingának, hogy mindig mellettem állt.

# Tartalomjegyzék

**Köszönetnyilvánítás** **1**

**Ábrák jegyzéke** **4**

**1. Bevezetés** **5**
1.1. Bevezetés, a dolgozat célja . . . . . 5
1.2. Bevezetés, idősorokról általában . . . . . 6

**2. ARMA, ARIMA és FARIMA folyamatok** **9**
2.1. Bevezetés . . . . . 9
2.1.1. Idősorelemzés alapfogalmai . . . . . 9
2.1.2. ARMA folyamatok . . . . . 11
2.1.3. Autoregresszív (AR) folyamatok . . . . . 13
2.1.4. Mozgóátlag (MA) folyamatok . . . . . 15
2.1.5. ARMA folyamatok . . . . . 16
2.2. ARIMA folyamatok . . . . . 18
2.3. FARIMA folyamatok . . . . . 19
2.3.1. FARIMA modellek bevezetése . . . . . 22

**3. Spektrum és periodogram** **25**
3.1. Bevezetés . . . . . 25
3.1.1. Bevezető fogalmak . . . . . 25
3.2. Spektrum . . . . . 30
3.3. Spektrum becslése, a tapasztalati autokovarianciából . . . . . 31
3.4. Periodogram . . . . . 34
3.4.1. Spektrum becslése periodogrammal . . . . . 35
3.4.2. A periodogram aszimptotikus viselkedése . . . . . 35
3.5. Fehérzaj, AR(1) és MA(1) folyamatok spektruma . . . . . 38
3.5.1. A fehérzaj spektruma . . . . . 38
3.5.2. AR(1) spektruma . . . . . 38
3.5.3. AR(p) spektruma . . . . . 40
3.5.4. MA(q) spektruma . . . . . 41
3.5.5. ARMA(p,q) spektruma . . . . . 41

**4. Zajos ARIMA, FARIMA** **42**

4.1. Zajos ARIMA és FARIMA folyamatok . . . . . 42
4.1.1. Bevezetés . . . . . 42
4.1.2. Független sorok aggregációja . . . . . 46

**5. Távolságok** **52**
5.1. Idősorok klaszterezése . . . . . 52
5.1.1. Bevezetés . . . . . 52
5.1.2. Technika . . . . . 53
5.1.3. Alaptechnikák . . . . . 54
5.2. Távolságok többváltozós idősorok esetén . . . . . 56
5.2.1. Bevezetés . . . . . 56
5.2.2. Távolságok . . . . . 56

**6. Gyakorlat és eredmények** **60**
6.1. Adatok . . . . . 60
6.2. Matematikai módszerek . . . . . 64
6.3. Eredmények . . . . . 65
6.3.1. Összefoglalás . . . . . 71
6.4. Klasszikus idősor modellek klaszterezése . . . . . 72
6.4.1. Kanadai hiúz adatsor (lynx) . . . . . 72
6.4.2. Közel szinguláris lineáris modellek . . . . . 79

**Irodalomjegyzék** **83**

# Ábrák jegyzéke

1.1. Példa egy idősorra. . . . . 6

2.1. Példa a Gauss fehérzajra. . . . . 12

3.1. Prizma. . . . . 25
3.2. Hullámhossz, amplitúdó. . . . . 27
3.3. Fázis, fázisszög. . . . . 28
3.4. AR(1) spektruma . . . . . 39
3.5. AR(2) spektruma . . . . . 40
3.6. ARMA spektruma . . . . . 41

6.1. EU-27 skálázva . . . . . 66
6.2. Klaszter . . . . . 67
6.3. Mennyire stabil az osztályozás . . . . . 68
6.4. 10%-90%-es kvantilisek alapján . . . . . 70
6.5. 10%-75%-ös kvantilisek alapján . . . . . 70
6.6. 25%-90%-es kvantilisek alapján . . . . . 71
6.7. Lynx adatsora . . . . . 73
6.8. Lynx adatsorra illesztett 12 modell spektrumának skálázott képe . . 79
6.9. 12 modell spektrális távolságának skálázott képe . . . . . 80

# 1. fejezet

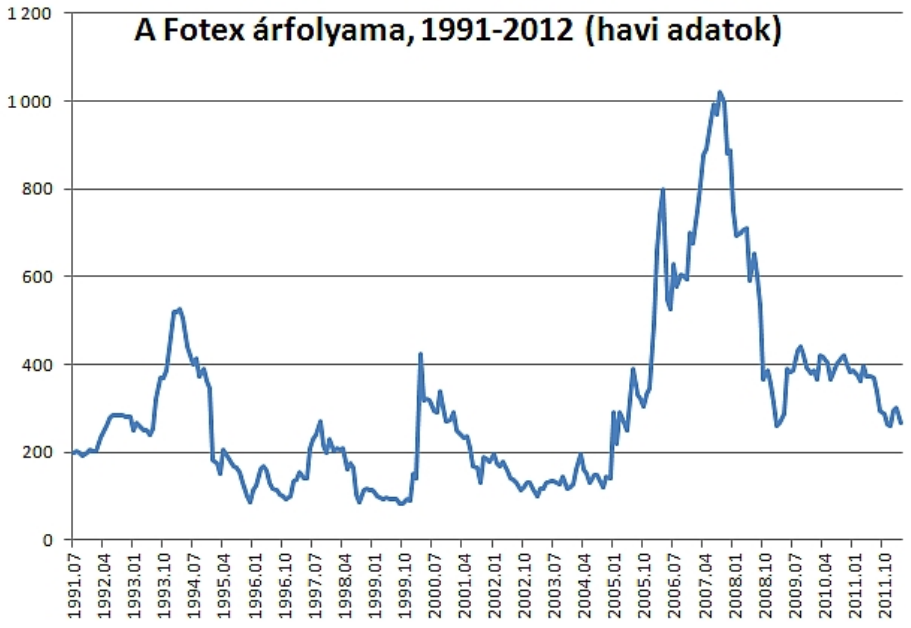
# Bevezetés

## 1.1. Bevezetés, a dolgozat célja

A dolgozat témája az idősorok osztályozása. A bevezetés során áttekintem ARMA, ARIMA, FARIMA folyamatokat, majd a zajjal megterhelt FARIMA folyamatokat. Ez után a spektrum, a spektrum-becslés, periodogram témaköre következik, aztán a távolságok, majd a 6. fejezetben a spektrum általánosításának lehetőségei, a kvantilis spektrum. Végül ezeket alkalmazva elemzem és osztályozom az ESI 1985.01-től 2013.04-ig tartó havi adatsorát. Befejezésképpen az eredményeket bemutatom.

## 1.2. Bevezetés, idősorokról általában

A matematikai statisztika egyik meghatározó részét képzi az idősorok elmélete, modellezése. Az idősorok gyakorlati modellezése az 1980-as évektől indult rohamos fejlődésnek. A kapott eredmények nagy hatással voltak az eloszlás-elméletre és az ökonometriai ismeretekre is. Egyre inkább elterjedt az időben nem állandó vagyis nem stacionárius (például időben növekvő) változókkal történő modellezés, és mindez kihatott nemcsak az idősorelemzésre, hanem az egész statisztika fejlődésére.




1.1. ábra. Példa egy idősorra, Fotex árfolyama, 1991-2012, havi adatok

Olyan esetekben, amikor az egyes megfigyelések időben nem feltétlenül függetlenek, valamint az egyes megfigyelési elemekre az azonos eloszlás hipotézise sem érvényes, célszerű lehet az elemek időbeni összefüggését feltételezni is vizsgálni. Tehát az egyik alapgondolat a megfigyelések közti időbeni kapcsolat, összefüggés. Az ilyen összefüggések feltérképezésére születtek a különböző idősormodellek. Az

ilyen modelleknek mind az az alapgondolata, hogy a megfigyelések adatsora tulajdonképpen egy véletlen folyamat, és amely folyamat változásait és alakulása következő tényezők határozzák meg:

- Trend: a változások hosszabb távú alapiránya, tendenciája (Jele: T)
- Ismétlődő komponensek: a trend körüli szabályszerű ingadozást írják le:
  - Rövid távon: Szezonalitás (Jele: S)
  - Hosszabb távon: Ciklikus hatás (Jele: C)
- Zaj: véletlen tényezőkkel összefüggő szabálytalan ingadozás.

A fent említett modellösszetevők az úgynevezett 'determinisztikus' idősorelemzés elméletéhez tartozik, ami azon alapszik, hogy az idősorok alakulása megközelítőleg megfejthető és leírható. Ez a gondolatkör vezet a dekompozíciós modellek felé, amik a fenti felsorolásban bemutatott komponensekre próbálják bontani az idősort, úgy mint trend, ciklus, szezonalitás és a zaj. A véletlen tényező a ciklikus komponens és a zaj.

A másik 'irányzat' a sztochasztikus idősorelemzés, ami pedig attól tér el az előzőtől, hogy itt a modellre ható véletlen, és az így kialakuló értékekre a véletlennek a későbbiekben is hatása van, és így az idősor időbeni alakulását is befolyásolja. Ezért úgy is szokták mondani, hogy az idősor időbeni alakulásában 'öngeneráló' folyamatok is érvényesülnek, tehát az adott időpontbeli véletlen befolyással bír a későbbi értékekre is, így a véletlen komponensnek folyamatalakító hatása van. Ez az irányzat sikeresen érvényesül a különböző közgazdasági, pénzügyi idősorok modellezésében, különösen rövid távon. Ez utóbbi a sztochasztikus modellezés része, illetve folytatása a "determinisztikusnak" egyrészt a ciklus, másrészt a zaj vonatkozásában.

Az idősorelemzés feladatai közé tartozik: a folyamatot leíró modell becslése, előrejelzése. Az esetleges zavaró tényezők kiszűrése és a véletlen meghatározása. A zaj nélküli folyamat, illetve a zaj meghatározása.

Az idősorok egyik alapmodellje az ARMA (autoregresszív mozgóátlag modell), ennek kiterjesztettje az ARIMA (integrált autoregresszív mozgóátlag modell) és a FARIMA (frakcionálisan integrált autoregresszív mozgóátlag modell).

# 2. fejezet

# ARMA, ARIMA és FARIMA folyamatok

## 2.1. Bevezetés

### 2.1.1. Idősorelemzés alapfogalmai

Ezt a fejezetet Dr. Márkus László egyetemi jegyzetére támaszkodva tárgyalom. Bevezetésképpen idősorelemzés alapfogalmait ismertetem, illetve az ARMA modellt, valamint végezetül rátérek a zajos ARIMA és FARIMA folyamatokra.

Akkor következzenek az alapfogalmak.

**2.1.1. Definíció.** $X_1, X_2, X_3, \ldots$ *valószínűségi változók sorozatát idősornak hívjuk, ha az indexparaméter a méréseket sorrendbe rendezi, például idősorrendbe.*

Az idősor egymás utáni állapotai, értekei összefüggethetnek. Az idősor valamilyen folyamat időbeni fejlődését írja le.

**2.1.2. Definíció.** *Az autokorreláció függvény (ACF):*

$$\rho(X(t), X(s)) = \frac{cov(X(t), X(s))}{\sigma_{X(t)}\sigma_{X(s)}} = \frac{E(X(t) - E(X(t)))E(X(s) - E(X(s)))}{\sigma_{X(t)}\sigma_{X(s)}}$$

**2.1.3. Definíció.** *$X_t$ folyamat, $t = 1, 2, 3, ...$, gyengén stacionárius, ha $E(X) =$ konstans, az autokovarianciafüggvény pedig: $cov(X_t, X_s) = R(t - s)$ csak a $(t - s)$ időpontok különbségétől függ.*

**2.1. Megjegyzés.** *A gyenge stacionaritást másodrendben stacionaritásnak is nevezhetjük.*

**2.1.4. Definíció.** *$X_t$ folyamatot erősen stacionáriusnak hívjuk, ha a véges dimenziós eloszlásai időben eltolás invariánsak, vagyis minden $l$-re, $n$-re és $t_1, t_2, t_3, ..., t_n$-re az $(X_{t_1}, X_{t_2}, X_{t_3}, ..., X_{t_n}) \sim (X_{t_1+l}, X_{t_2+l}, X_{t_3+l}, ..., X_{t_n+l})$.*

**2.1.5. Definíció.** *$X_t$ folyamat $k$-adrendben stacionárius, ha legfeljebb $k$-adrendű vegyes momentumai $k$-adrendben eltolásinvariánsak.*

**2.2. Megjegyzés.** *Ha erősen stacionárius egy idősor, akkor minden $k$-ra $k$-adrendben stacionárius, és $X_t$ folyamatot eloszlása állandó.*

**2.3. Megjegyzés.** *A stacionaritás tulajdonsága úgy is értelmezhető, hogy az idősor, mint egy véletlen folyamat által meghajtott dinamika, az idő előrehaladtával, hosszú távon stabil egyensúlyi helyzetbe kerül. Tehát a stabilitás miatt a stacinárius idősorok rövid távon megbízható módon előrejelezhetőek.*

**2.1.6. Definíció.** *Az idősoroknál az egyes időpontbeli értékek közti kovarianciát autokovariancia függvénynek hívjuk, jelölése: $cov(X(t), X(s)) = R(t, s)$. De szokták még így is jelölni $C(t, s), \gamma(t, s)$ vagy $B(t, s)$. Stacionárius idősornál $R(t, s) = R(t - s)$, és $R_X(0) = \sigma_X^2 = D^2X(t) = D^2X(t + h)$.*

### 2.1.2. ARMA folyamatok

Az idősorelemzés elméletében és gyakorlatában az autoregresszív és mozgóátlag (ARMA) folyamatok az 1970-es évektől nagy jelentőséget kaptak, mivel az ARMA folyamatok matematikailag jól kezelhetőek, és a gyakorlatban előforduló és stacionárius viselkedést követő véletlen folyamatok nagy része jól közelíthető és leírható velük. Ilyen folyamatok lehetnek akár gazdasági, biológiai, ipari, kémiai és egyéb folyamatok is. Így az ARMA folyamatok felhasználása valóban széleskörű.

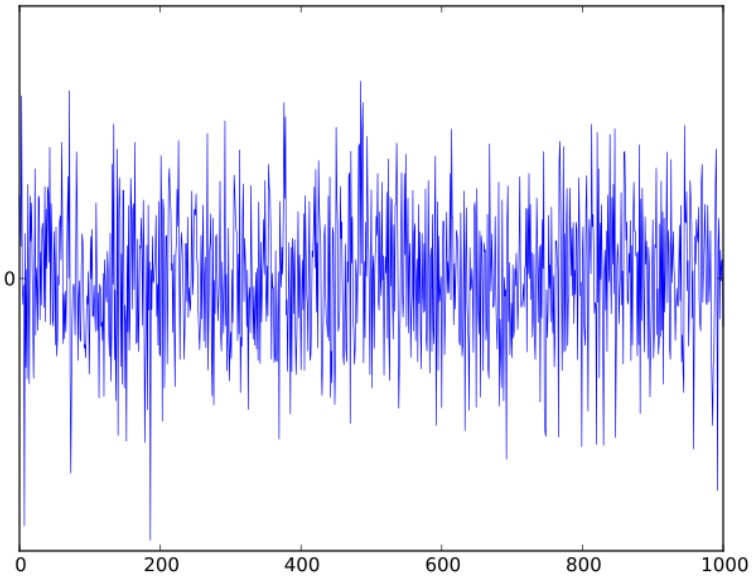
Az ARMA folyamatok a lineáris folyamatok családjába tartoznak, ugyanis ilyen alakban állnak elő:
$\sum_{i=-\infty}^{\infty} \alpha_i e_{t-i}$,ahol $t \in \mathbb{Z}$ és $e_t$ fehérzaj folyamat.

**2.1.7. Definíció.** *Az $\varepsilon(t)$ folyamat független értékű zaj, ha a folyamat várható értéke 0 és $\varepsilon(t)$ független, azonos eloszlású valószínűségi változók (röviden jelölve: i.i.d.).*

**2.1.8. Definíció.** *Az $\varepsilon(t)$ folyamat fehérzaj, ha a folyamat várható értéke 0, azonos eloszlású minden t-re és korrelálatlanok (nem feltétlenül függetlenek). Ha $\varepsilon(t)$ folyamat normális eloszlású, akkor a korrelálatlanság megegyezik a függetlenséggel. A fehérzaj folyamat autokovariancia függvénye: $R(0) = \sigma^2$, $R(\tau) = 0$ minden $\tau \geq 1$.*

**2.4. Megjegyzés.** *A független értékű zaj erősen, a fehérzaj pedig gyengén stacionárius.*

**2.1.9. Definíció.** *Az $\varepsilon(t)$ sztochasztikus folyamat Gauss fehérzaj, ha minden t-re független és normális eloszlású nulla várható értékkel valamint véges $\sigma^2$ varianciával.*




2.1. ábra. Gauss fehérzaj

### 2.1.3. Autoregresszív (AR) folyamatok

Az autoregresszivitás arra utal, hogy a folyamat részben a saját múltjától is függ, így felírható a saját múltjának lineáris regressziójaként is.

Jelölés: bármilyen folyamat indexelésénél az alsó index és a zárójeles index ugyanazt jelenti, például:
$x(t) = x_t$.

**2.1.10. Definíció.** *Elsőrendű autoregresszív folyamat* `AR(1)`*:* $X(t) = \alpha X(t-1) + \sigma_\varepsilon \cdot \varepsilon(t)$. *Az* $\varepsilon(t)$ *-k független, azonos eloszlású valószínűségi változók. Leggyakrabban az* $\varepsilon(t)$*-k* $N(0,1)$*-ek. Az* $X(t)$ *folyamat várható értéke 0, és* $X(t-1)$ *és* $\varepsilon(t)$ *függetlenek, vagyis* $\varepsilon(t)$ *és az* $X(t)$ *múltja függetlenek.*
*Így mivel a folyamat múltja független az* $\varepsilon(t)$*-tól, ezért az* $X(t)$ *szórásnégyzete az* $X(t-1)$ *és az* $\varepsilon(t)$ *szórásnégyzetének az összege lesz:*
$D^2X(t) = \alpha^2 D^2X(t-1) + \sigma_\varepsilon^2 D_\varepsilon^2 \varepsilon(t)$.
*Ha van stacionárius megoldás, akkor a stacionaritásból kifolyólag* $D^2X(t) = D^2X(t-1)$. *Ebből következik, hogy:*
$\sigma_X^2 = \alpha^2\sigma_X^2 + \sigma_\varepsilon^2$, *ahol* $\sigma_X^2 = D^2X(t)$.
*A fenti egyenlőséget átrendezve azt kapjuk, hogy:*
$\sigma_X^2 = \frac{\sigma_\varepsilon^2}{1-\alpha^2} > 0$.
*Ebből adódik,hogy akkor és csak akkor létezhet stacionárius megoldás, ha* $|\alpha| < 1$.

#### Magasabb rendű autoregresszív folyamatok

Másodrendű autoregresszív `AR(2)` folyamat:
$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \sigma_\varepsilon \varepsilon(t)$.

A p-edrendű autoregresszív `AR(p)` folyamat:

$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \cdots + \alpha_p X(t-p) + \sigma_\varepsilon \varepsilon(t)$.

Az egyenlőség átrendezve:

$X(t) - \sum_{k=1}^{p} \alpha_k X(t-k) = \sum_{k=0}^{p} \widetilde{\alpha_k} X(t-k) = \sigma_\varepsilon \cdot \varepsilon(t)$,

ahol $\widetilde{\alpha_0} = 1$ és $\widetilde{\alpha_k} = -\alpha_k$.

**2.5. Megjegyzés.** *Az AR folyamat karakterisztikus polinomját $P(x)$-szel szokás jelölni. Az:*

$X(t) - \sum_{k=1}^{p} \alpha_k X(t-k) = \sum_{k=0}^{p} \widetilde{\alpha_k} X(t-k) = \sigma_\varepsilon \cdot \varepsilon(t)$

*egyenlettel felírt p-ed rendű AR folyamat karakterisztikus polinomja:*

$P(x) = \sum_{k=0}^{p} \widetilde{\alpha_k} x^k$.

**2.6. Megjegyzés.** *Az AR(p) egyenletnek akkor és csak akkor van egyértelmű stacionárius megoldása (indítása), ha $P(z)$ karakterisztikus polinom gyökei az egységkörön kívül vannak. Ha gyengén stacionárius folyamat Gauss folyamat, akkor erősen stacionárius is.*

**2.1.11. Definíció.** *Jelöljük $B$-vel vagy $L$-lel az eltolás vagy késleltető vagy hátralépetető operátort, ami a folyamat értékeit időben lépteti vissza, a következők szerint:*

$BX(t) = X(t-1)$, $B^2X(t) = X(t-2)$ *és így tovább.*

**2.7. Megjegyzés.** $(\sum_{k=0}^{p} \widetilde{\alpha_k} B^k) X(t) = \varepsilon(t)$, *átrendezve:*

$X(t) = (\sum_{k=0}^{p} \widetilde{\alpha_k} B^k)^{-1} \cdot \varepsilon(t)$.

*Természetesen a fenti sor csak akkor írható fel, ha a $\sum_{k=0}^{p} \widetilde{\alpha_k} B^k$-nek létezik inverze.*

### 2.1.4. Mozgóátlag (MA) folyamatok

Az ARMA folyamat másik 'összetevője' a mozgóátlag folyamat.

A mozgó átlag arra utal, hogy a lineáris regresszió hibatagja az $\varepsilon(t)$ mozgó átlaga, vagyis a jelen és a múlt lineáris kombinációja. A gyakorlatban az $\varepsilon(t)$ együtthatói nem szükségképpen pozitívak, és az összegük sem feltétlenül 1, így az elnevezés nem teljesen pontos, hiszen nem valódi átlagról van szó.

**2.1.12. Definíció.** *q-adrendű mozgóátlag* `MA(q)` *folyamat:*

$X(t) = \beta_0 \cdot \varepsilon(t) + \beta_1 \cdot \varepsilon(t-1) + \cdots + \beta_q \cdot \varepsilon(t-q)$

*ahol $\varepsilon(t)$ fehérzaj vagy speciálisan Gauss fehérzaj.*

**2.8. Megjegyzés.** *A lineáris folyamatok felírhatóak $\infty$-rendű mozgóátlag folyamatként.*

**2.9. Megjegyzés.** *$X(t)$ MA folyamat autokovariancia-függvénye a következőképpen alakul, ha $\varepsilon(t)$ Gauss fehérzaj nulla várható értékű, 1 szórásnégyzettel:*

$R(\tau) = E(X(t) \cdot E(X(t+\tau))) - E(X(t)) \cdot E(X(t+\tau)) = E(X(t) \cdot E(X(t+\tau)))$

*mivel ha $E(\varepsilon_t) = 0$, akkor $E(X(t)) = 0$ is igaz.*

*$R(\tau)$ kovariancia előállítható így:*

$R(\tau) = cov(X(t), X(t+\tau)) = E(X(t) \cdot E(X(t+\tau))) =$

$= cov(\beta_0 \cdot \varepsilon(t) + \beta_1 \cdot \varepsilon(t-1) + \cdots + \beta_q \cdot \varepsilon(t-q), \beta_0 \cdot \varepsilon(t+\tau) + \beta_1 \cdot \varepsilon(t+\tau-1) +$

$\cdots + \beta_q \cdot \varepsilon(t+\tau-q) =$

$= \beta_0 \cdot \beta_\tau + \beta_1 \cdot \beta_{\tau-1} + \beta_2 \cdot \beta_{\tau-2} + \cdots + \beta_{\tau-q} \cdot \beta_q$

**2.10. Megjegyzés.** *$R(\tau)$ tényleg nem függ az időtől, t-től, így $X(t)$ stacionárius.*

**2.1. Tétel.** *Ha a $R(\tau)$ autokovariancia-függvényre teljesül az előző előállítás, akkor van olyan MA(q) folyamat, hogy az autokovariancia-függvény együtthatói pont a fenti felbontás $\beta$ paraméterei.*

*Ha $X(t)$ stacionárius Gauss-folyamat és a várható értéke 0, valamint ha $\tau > q$-ra $R(\tau) = 0$, akkor $X(t)$ egy MA(q).*

Egy $MA(q)$ $X(t)$ folyamat karakterisztikus polinomja $Q(x)$:
$Q(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_q x^q$.

A fenti definíció szerint $X(t) = \beta_0 \cdot \varepsilon(t) + \beta_1 \cdot \varepsilon(t-1) + \cdots + \beta_q \cdot \varepsilon(t-q)$.
Vagyis $X(t) = Q(B)\varepsilon(t) = (\beta_0 + \beta_1 B + \beta_2 B^2 + \cdots + \beta_q B^q)\varepsilon(t)$, ahol
$B\varepsilon(t) = \varepsilon(t-1)$.

Így ha $(Q(x))^{-1} = \sum_{j=0}^{\infty} \alpha_j x^j$ sor konvergens, akkor $(Q(B))^{-1} = \sum_{j=0}^{\infty} \alpha_j B^j$ is létezik, és így $\varepsilon(t)$ megadható $\sum_{j=0}^{\infty} \alpha_j X(t-j)$ alakban, vagyis $X(t)$ MA(q)-nak létezik $AR(\infty)$ előállítása.

**2.2. Tétel.** $\sum_{j=0}^{\infty} \alpha_j X(t-j)$ *akkor és csak akkor konvergens, ha a $Q(x)$ karakterisztikus polinom gyökei az egységkörön kívül vannak.*

Az előzőekből is látható az $AR(\infty)$ és a $MA(\infty)$ 'dualitása'. MA(q)-nak létezik $AR(\infty)$ előállítása, akkor $X(t)$ MA(q) folyamat invertálható, és $X(t)$ akkor és csak akkor invertálható, ha a karakterisztikus polinom gyökei az egységkörön kívül vannak.

### 2.1.5. ARMA folyamatok

**2.1.13. Definíció.** $\sum_{k=0}^{p} \widetilde{\alpha_k} X(t-k) = \sum_{m=0}^{q} \beta_m \varepsilon(t-m)$

A két karakterisztikus polinom $P(x)$ és a $Q(x)$.

**2.3. Tétel.** *Ha az AR folyamat karakterisztikus polinomjának, $P(x)$-nek, a gyökei az egységkörön belül helyezkednek el, akkor létezik stacionárius ARMA folyamat, és ennek előállítása $MA(\infty)$.*

*Ha pedig a MA folyamat karakterisztikus polinomjának, $Q(x)$-nek, a gyökei az egységkörön kívül helyezkednek el, akkor a folyamatunknak $AR(\infty)$ előállítása létezik.*

**2.11. Megjegyzés.** *Az $MA(\infty)$ előállításához a $\frac{Q(x)}{P(x)}$ kell, míg az $AR(\infty)$ meghatározásához pedig ennek reciproka a $\frac{P(x)}{Q(x)}$ hatványsor előállítása szükséges.*

## 2.2. ARIMA folyamatok

Az ARIMA folyamatok egyfajta általánosítását vagy kiterjesztését jelentik az ARMA folyamatoknak.

Az ARIMA folyamat 'jelentése': autoregresszív integrált mozgóátlag folyamat. Paraméterezése: ARIMA(p,d,q), ahol a d jelenti az integráltságnak a fokát. Az ARIMA folyamatok bevezetésére azért van szükség, mert nem mindig stacionárius a folyamatunk. Ugyanakkor lehetséges, hogy ha a folyamat egy ARIMA(p,d,q)-t d-szer differenciálunk, akkor már stacionárius folyamatot kapunk. Így a modellből eltüntethető a nem stacionárius rész.

**2.2.1. Definíció.** *$X(t)$ ARIMA(p,1,q) folyamat, ha az egyszeres különbségképzéssel kapott új folyamat: $Y(t) = X(t) - X(t-1) = (1-B)X(t)$ ARMA folyamat.*

**2.12. Megjegyzés.** *Az ARIMA(p,1,q)-nál a differenciálással a nem stacionaritást okozó (lokális) lineáris trend tüntethető el.*

**2.2.2. Definíció.** *$X(t)$ ARIMA(p,d,q) folyamat, ha a d-szeres differenciáltja, az új $Y(t) = (1-B)^d X(t)$ folyamat ARMA.*

**2.13. Megjegyzés.** *Az ARIMA(p,d,q)-nál d-szeres differenciálással a d-edfokú polinomiális trend tüntethető el.*

## 2.3. FARIMA folyamatok

A FARIMA (frakcionálisan integrált autoregresszív-mozgóátlag) folyamat vagy szokták még ARFIMA-nak is hívni, az ARIMA(p,d,q) folyamat kiterjesztése vagy általánosítása úgy, hogy a $d$-vel jelölt integrálási paraméternek nem kell egész számnak lennie, vagyis frakcionálisan (tört) integrált lesz a folyamat. A FARIMA modelleknek nagy a jelentőségük a hosszú-memóriájú idősorok leírásánál. Mi a hosszú memória? Egy kis kitérő következik.

### Hosszú emlékezetű folyamatok

A hosszú emlékezetű folyamatokat az autokovarianciájuk viselkedés különbözteti meg a rövid emlékezetű folyamatoktól. A hosszú memóriájú folyamatoknál az autokovariancia-függvény nem exponenciális lecsengésű, hanem annál lassabb (hiperbolikusan cseng le), így "mutatja a hosszútávú összefüggést". A Hurst-együttható mutatja a hosszú emlékezetűség mértékét. A Hurst-együttható 0 és 1 között mozog, ha 1/2-del egyenlő, akkor nem hosszú emlékezetű a folyamat, és minél közelebb van az 1-hez, annál inkább hosszú emlékezetű az adott folyamat.

Többféle megközelítése létezik a hosszú emlékezetűség definiálására, így többféle definíció létezik rá (ezek többé-kevésbé ekvivalensek egymással). Én kétféleképpen fogom definiálni.

**2.3.1. Definíció.** *$X(t)$ véges szórású, stacionárius folyamat, az autokorreláció függvénye, $r(k)$. Ekkor $X(t)$ hosszú emlékezetű folyamat, ha*
$\lim_{k\to\infty} \sum_{k=-n}^{n} |r(k)| = \infty$.

**2.3.2. Definíció.** *$X(t)$ véges szórású, stacionárius folyamat, az autokovariancia függvénye, $r(k)$, hiperbolikusan lecsengő, és létezik spektrálsűrűség-függvénye, jele: $\varphi(\lambda)$, valamint létezik $0 < \beta < 1$ és $d > 0$ konstansok, hogy:*
$\lim_{\lambda\to 0} \frac{\varphi(\lambda)}{d|\lambda|^{-\beta}} = 1$,

*tehát a spektrálsűrűség-függvénynek létezik $\beta$-rendű pólusa 0-ban ($\varphi(0) = \infty$). Ekkor $X(t)$ hosszú emlékezetű.*

**2.14. Megjegyzés.** *$H = \frac{\beta+1}{2}$ és $1/2 < H < 1$, ahol $H$ a Hurst-együttható.*

**2.1. Állítás.** *$H$ a Hurst-együttható, $0 < H < 1$.*

**2.3.3. Definíció.** *$X(t)$ d-ed rendben integrált folyamat vagy $I(d)$-beli, ha $-1/2 < d < 1/2$ és $Y(t) = (1-B)^d X(t)$, ahol $Y(t)$ folyamat stacionárius, pozitív, korlátos spektruma van minden frekvencián.*

**2.15. Megjegyzés.** *Ha $-1/2 < d < 0$-ra $\sum_{k=-\infty}^{k=\infty} |r(k)| < \infty$, akkor $X(t)$ folyamat rövid memóriájú.*

**Frakcionális Brown mozgás**

Legyen $B(t)$ Brown mozgás, azaz $B(0) = 0$, független növekményű, $0 \leq t_0 < t_1 < t_2 < \cdots < t_n$-re $B(t_1) - B(t_0), B(t_2) - B(t_1), \ldots B(t_n) - B(t_{n-1})$ függetlenek és $B(t) - B(s) \sim N(0, t-s)$, $s < t$, $E(B(t)) = 0$ minden $t$-re, folytonos trajektóriájú. Ennek (másik) ekvivalens definíciója: $B(t)$ Brown mozgás, ha 0-ból induló Gauss folyamat, $E(B(t)) = 0$ minden $t$-re és $cov(B(t), B(s)) = \min(t, s) = t \wedge s$.

**2.3.4. Definíció.** *$W(t)$ Gauss folyamat, ha minden $n$-re és minden $t_1, \ldots, t_n$-re és $c_1, \ldots, c_n \in \mathbb{R}$-re $\sum_{i=1}^{n} c_i W(t_i)$ normális eloszlású.*

Legyen $B_H(t)$ frakcionális Brown mozgás $H$ paraméterrel. $B_H(t)$ stacionárius növekményű Gauss folyamat, $EB_H(t) = 0$, $D^2 B_H(t) = t^{2H}$,
$cov(B_H(t), B_H(s)) = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H})$.
A $B_H(t)$ frakcionális Brown mozgás növekményei: $B_H(s) - B_H(u)$, $B_H(t)$ és $B_H(s) - B_H(u)$ kovarianciája:
$cov(B_H(t), B_H(s) - B_H(u)) = cov(B_H(t), B_H(s)) - cov(B_H(t), B_H(u)) =$
$= \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H} - t^{2H} - u^{2H} + |t-u|^{2H}) =$

$= \frac{1}{2}(s^{2H} - |t-s|^{2H} - u^{2H} + |t-u|^{2H}) = \frac{1}{2}(s^{2H} - u^{2H} - [(t-s)^{2H} - (t-u)^{2H}])$,

ha $s < u < t$.

Ha $H = 1/2$, akkor kovariancia 0, ha $H > 1/2$, akkor a kovariancia nagyobb, mint 0.

A frakcionális Brown mozgás nem független növekményű.

$B_H(t)$ frakcionális Brown mozgás felírható a közönséges $B(t)$ Brown mozgás $1/2-H$ rendű frakcionális deriváltjaként.

$B_H(t) = \frac{1}{\Gamma(H+1/2)} \int\limits_0^t (t-x)^{H-1/2} dB(x)$.

**2.16. Megjegyzés.** *$H = 1/2$, akkor Brown mozgást kapunk.*

**Frakcionális fehérzaj**

**2.3.5. Definíció.** *$\xi(t) = B_H(t) - B_H(t+h)$, ekkor $X(t)$ folyamat frakcionális fehérzaj, és $B_H(t)$ frakcionális Brown mozgás növekményeivel egyenlő.*

A frakcionális Brown mozgás növekményei hosszú memóriájú folyamatok. A Brown mozgásból a diszkrét fehérzaj adódik, a frakcionális Brown mozgásból pedig a már fent definiált frakcionális fehérzaj. Ha $1/2 < H < 1$, akkor hosszú emlékezetű stacionárius folyamat lesz.

A frakcionális fehérzaj előállítható $d = H - 1/2$-del:

$(1-B)^d(\xi(t) - \mu) = \varepsilon(t)$,

ahol $\varepsilon(t)$ fehérzaj, $E\varepsilon(t) = 0$, $D^2(\varepsilon(t)) = \sigma_\varepsilon^2$ és $E[\varepsilon(t)\varepsilon(s)] = 0$,

és $\xi(t)$ gyengén stacionárius frakcionális fehérzaj.

Az egyenlet alapján $\xi(t)$ $I(d)$-beli.

**2.17. Megjegyzés.** $(1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k}(-B)^k$

$= \sum_{k=0}^{\infty} \frac{\prod_{a=0}^{k-1}(d-a)(-B)^k}{k!}$

$= 1 - dB + \frac{d(d-1)}{2!}B^2 - \dots$

**2.18. Megjegyzés.** *$\xi(t)$ frakcionális fehérzaj variancia-függvénye:*

$R(0) = D^2(\xi(t)) = \sigma_\xi^2 = \sigma_\varepsilon^2 \frac{\Gamma(1-2d)}{\Gamma^2(1-d)}$

*$\xi(t)$ frakcionális fehérzaj k-adrendű autokovariancia-függvénye:*

$R(k) = \sigma_\xi^2 r(k) = \sigma_\varepsilon^2 \frac{\Gamma(k+d)\Gamma(1-2d)}{\Gamma(k+1-d)\Gamma(1-d)\Gamma(d)}$

$= \frac{\sigma_\varepsilon^2}{2\pi} \sin(\pi \cdot d) \frac{\Gamma(k+d)\Gamma(1-2d)}{\Gamma(k+1-d)}$,

*ahol $r(k)$ $\xi(t)$ k-adrendű autokorreláció-függvénye.*

$\xi(t)$ frakcionális fehérzaj korrelált folyamat, elsőrendű autokorreláció-függvénye:

$r(1) = \frac{d}{1-d}$, másodrendű: $r(2) = \frac{d(d+1)}{(1-d)(2-d)}$

k-adrendű autokorreláció-függvénye: $r(k) = \frac{\Gamma(k+d)\Gamma(1-d)}{\Gamma(k-d+1)\Gamma(d)}$

**2.19. Megjegyzés.** *A Stirling- formulából ha $k \to \infty$, akkor $\frac{\Gamma(k+a)}{\Gamma(k+b)} \approx k^{a-b}$, ebből $r(k) \sim konstans \cdot k^{2d-1} = konstans \cdot k^{2(H-1)}$.*

**2.2. Állítás.** *$\xi(t)$ spektrálsűrűség-függvénye:*

$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} |1 - e^{-i\lambda}|^{-2d}$,

*kis frekvenciákon:*

$f(\lambda) \approx konstans \cdot \lambda^{-2d} \to \infty$, *ha $\lambda \to \infty$ és $d > 0$.*

### 2.3.1. FARIMA modellek bevezetése

Az ARIMA modellnél az integrált részt a $(1 - B)$ adja, ahol $B$ a visszaléptető operátor.

$(1 - B)^2 = 1 - 2B + B^2$, ahol $B^k X(t) = X(t - k)$

$(1 - B)^2 X(t) = X(t) - 2X(t - 1) + X(t - 2)$

Frakcionális modell esetén (2.18.Megjegyzés szerint):

$(1 - B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k$

$= \sum_{k=0}^{\infty} \frac{\prod_{a=0}^{k-1}(d-a)(-B)^k}{k!}$

$= 1 - dB + \frac{d(d-1)}{2!} B^2 - \dots$

**FARIMA(0,d,0)**

Ez a legegyszerűbb FARIMA modell. Jelölje $X(t)$-vel a FARIMA(0,1,0) folyamatot.

Ekkor $(1-B)^d X(t) = \varepsilon(t)$, ahol $\varepsilon(t)$ fehérzaj,

másképpen:

$X(t) - dX(t-1) + \frac{d(d-1)}{2!}X(t-2) - \cdots = \varepsilon(t)$

A FARIMA(0,d,0) frakcionális fehérzaj, ha $d = H - 1/2$. A FARIMA(0,d,0) spektrálsűrűség-függvénye:

$f(\lambda) = (1/2\pi)(2\sin(\lambda/2))^{-2d}$.

**FARIMA(p,d,q)**

A FARIMA(p,d,q) az ARIMA(p,d,q) "általánosítása", hogy $d \notin \mathbb{Z}$.

$\underline{\alpha}(B)(1-B)^d X(t) = \underline{\beta}(B)\varepsilon(t)$

$\underline{\alpha}(B)X(t) = (1-B)^{-d}\underline{\beta}(B)\varepsilon(t)$.

A frakcionális fehérzaj bekezdésénél, az előzőekben $\xi(t)$vel jelöltem a frakcionális fehérzajt és $\xi(t) = (1-B)^{-d}\varepsilon(t)$, $\xi(t)$. Ezt felhasználva a fenti egyenlet átírható:

$\underline{\alpha}(B)X(t) = \underline{\beta}(B)\xi(t)$.

A fenti egyenletekből látható, hogy egy frakcionális fehérzajból generált ARMA folyamat egy FARIMA folyamattal egyezik meg.

A FARIMA folyamatok hosszú emlékezetűek.

A FARIMA folyamatok autokovariancia- és autokorreláció-függvényét bonyolult meghatározni. De $\xi(t)$ frakcionális fehérzaj előáll így: $\xi(t) = (1-B)^{-d}\varepsilon(t)$, valamint $\underline{\alpha}(B)X(t) = \underline{\beta}(B)\xi(t)$, így ezt felhasználva $X(t)$ FARIMA folyamat autokovariancia-függvényét úgy határozzuk meg, hogy először az ARIMA kovarianciáját meghatározzuk és ezt "kombináljuk" a frakcionális fehérzaj autokovarianciájával.

**2.20. Megjegyzés.** *$X(t)$ FARIMA folyamat $R(k)$ k-adrendű autokovarianciája: $R(k) = \sum_{j,l=0}^{\infty} \delta_j \delta_l R^*(k-j-l)$, ahol $R^*()$ $\xi(t)$ frakcionális fehérzaj autokovariancia-függvénye.*

**2.3. Állítás.** *FARIMA spektrálsűrűség-függvénye:*
$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{|\underline{\beta}(e^{-i\lambda})|^2}{|\underline{\alpha}(e^{-i\lambda})|^2} |1 - e^{-i\lambda}|^{-2d}$.
*Kis frekvenciákra, ha $\lambda \to 0$, akkor*
$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{\underline{\beta}^2(1)}{\underline{\alpha}^2(1)} \lambda^{-2d}$.

**2.21. Megjegyzés.** *A spektrálsűrűség-függvénynek a fenti képlet szerint 0-ban pólusa van. FARIMA(p,d,q), $d = H - 1/2$, hosszú memóriájú folyamat.*

**2.22. Megjegyzés.** *FARIMA folyamatoknál d határozza meg a hosszú távú viselkedést, p és q paraméter pedig a rövid távú viselkedést.*

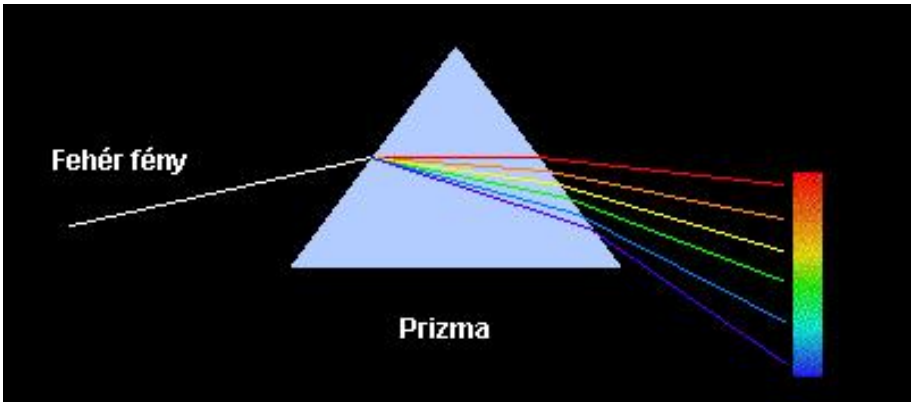
# 3. fejezet

# Spektrum és periodogram

## 3.1. Bevezetés

### 3.1.1. Bevezető fogalmak

A spektrum fogalma először a fizikában bukkant fel, ahol a fény felbontásánál jelent meg. A spektrum "színképet", "imázst", "jelenést" jelent. Klasszikus példa a prizmán áthaladó (fehér) fény, ami különböző (színű: piros, kék, zöld,...) frekvenciájú hullámokra bomlik.




3.1. ábra. Prizmára eső fény különböző színekre történő felbomlása

Tehát a függvények felbonthatóak különböző frekvenciájú szinusz és koszinusz függvények kombinációjára. A spektrum felfogható egyfajta sűrűségfüggvényként, ami azt mutatja meg, hogy az egyes frekvenciájú komponensek milyen intenzitással, gyakorisággal jelennek meg (különböző frekvenciák előfordulásának gyakorisága). Vagyis azt mutatja meg, hogy egy adott frekvenciájú periodikus komponens mekkora súllyal járul hozzá a megfigyelt folyamathoz, idősorhoz.
Tehát a periodikus függvények felbonthatóak különböző frekvenciájú szinusz és koszinusz függvények kombinációjára.

### Alapfogalmak

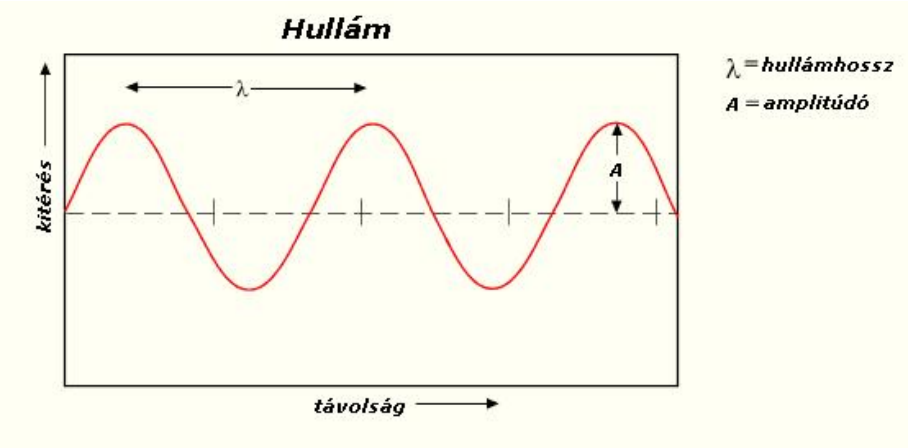
**3.1.1. Definíció.** *Jelöljük $T$-vel a periódusidőt, ami egy periódus lefutásának ideje.*

**3.1.2. Definíció.** *Frekvencia, jele: $f$, azt mutatja meg, hogy adott időegység (például 1 másodperc) alatt hány periódus fut le, ismétlődés gyakorisága.*
*$f = \frac{1}{T}$, ahol $T$ jelöli a periódusidőt.*

A szinusz-függvény periódusa: $T = 2\pi$. A frekvencia a hullámhosszal fordítottan arányos, így $f$ a sebesség és a hullámhossz hányadosával is definiálható.

**3.1. Megjegyzés.** *$f = \frac{v}{\lambda}$, ahol $v$ jelöli a sebességet, $\lambda$ pedig a hullámhosszt.*

**3.1.3. Definíció.** *A hullámhossz, $\lambda$, azonos fázisú pontok távolságát jelenti.*




3.2. ábra. Hullámhossz és Amplitúdó

**3.1.4. Definíció.** *Jelöljük A-val az amplitúdót. Ez a nyugalmi helyzethez képest mért legnagyobb kitérés,* $A > 0$.

**3.1.5. Definíció.** *Körfrekvencia jele:* $\omega$*,és* $\omega = 2\pi \cdot f$*, ahol* $f$ *a frekvencia.*

**3.2. Megjegyzés.** *Fázis a folyamat perióduson belüli pillanatnyi állapota.*

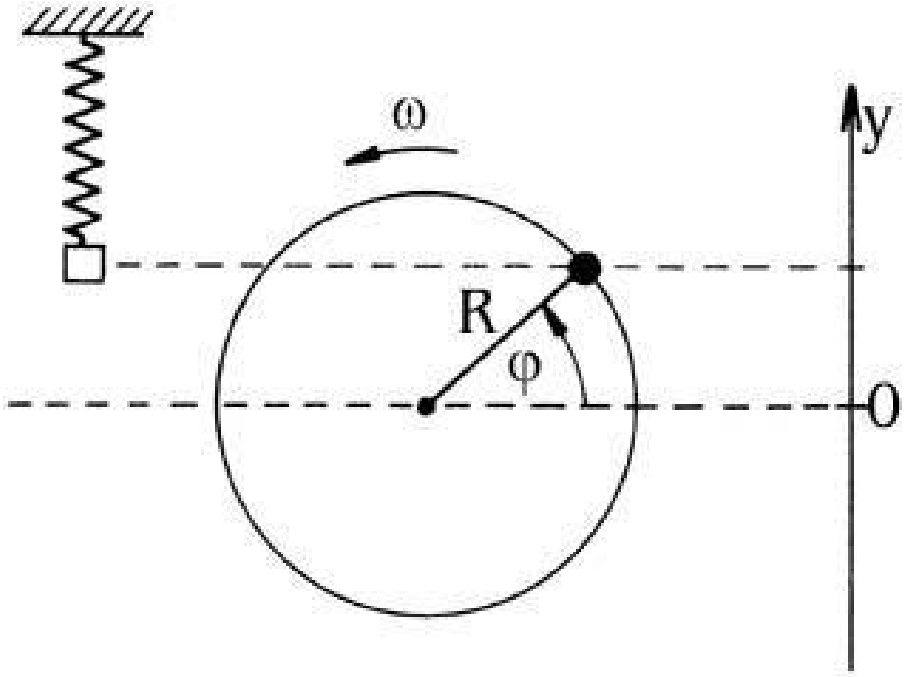
A szögsebesség jele, $\omega$, ami a sebesség $= \frac{\text{út}}{\text{idő}}$ definíció szerint: $\omega = \frac{\text{szögelfordulás}}{\text{idő}}$.

**3.3. Megjegyzés.** *A szögsebesség másképp szögfrekvencia, amit az azonos jelölés mutat, vagyis a körfrekvencia jele is:* $\omega$.

**3.1.6. Definíció.** *Körfrekvencia:* $\omega = \frac{2\pi}{T}$*, ahol* $T$ *a periódusidő.*

**3.1.7. Definíció.** *Fázisszög, jele:* $\varphi$*, egy adott időpontban* $\varphi = \varphi_0 + \omega \cdot t$*, ahol* $\varphi_0$ *a kezdő szög,* $t$ *pedig az eltelt idő.*

**3.4. Megjegyzés.** *A szögelfordulás az út, vagyis szögelfordulás = sebesség · idő.*




3.3. ábra. Fázis és a fázisszög.

### Általános bevezetés

A szabályos, ciklikus folyamatokat szinusz-, koszinusz függvényekkel le lehet írni:

$y_t = A \cdot \cos(\lambda t - \varphi)$,

ahol az előző részben már definiált fogalmakkal találkozunk,

vagyis $A$ az amplitúdó, $\varphi$ a fázisszög.

Legyen a frekvencia jele most $\lambda$.

**3.5. Megjegyzés.** *A frekvencia fordítottan arányos a hullámhosszal.*

A szögfüggvények argumentuma különféleképpen megadható, tehát a szinusz-, koszinusz függvények argumentuma, azaz $\sin(x)$, $\cos(x)$, $x \in [0, 1]$ vagy $x \in [0°, 360°]$ vagy $x \in [0, 2\pi]$ (radián).

Valamint $\cos(x) = \sin(x + \frac{\pi}{2})$.

Tehát:

$y_t = A \cdot \cos(\lambda t - \varphi) = a\cos(\lambda t) + b\sin(\lambda t)$

**3.6. Megjegyzés.** *Az amplitúdó,* $A = \sqrt{a^2 + b^2}$.

A trigonometrikus rendszer ortogonális rendszert alkot a folytonos periodikus függvények terén, vagyis:

$1, \cos x, \sin x, \cos 2x, \sin 2x, \ldots \cos nx, \sin nx \ldots$ ortogonális rendszer alkot.

Ortogonális (két dimenzióban merőleges: skaláris szorzat=0), vagyis páronkénti szorzat integrálja nullával egyenlő.

Legyen $f$ egy periodikus, vagyis $f(t) = f(t + k \cdot T)$, ekkor $f$ függvény előállítható így is:

$f(t) = \sum_{n=0}^{\infty}(a_n \cos(t\lambda_n) + ib_n \sin(t\lambda_n))$.

**3.7. Megjegyzés.** $e^{i \cdot x} = \cos(x) + i \cdot \sin(x)$

## 3.2. Spektrum

**3.2.1. Definíció.** $X(\lambda) = |\int\limits_{-\infty}^{\infty} x(t)e^{-i\lambda t}dt|^2$

*$x(t)$ folyamat Fourier transzformáltjának abszulút értékének négyzetét nevezzük spektrumnak.*

**3.2.2. Definíció.** $f_1(\lambda) = \sum_{l=-\infty}^{\infty} R(l)e^{-2\pi i\lambda l}$,

*$R(l)$ $X_t$ idősor l-edik autokovarianciája, azaz $R(l) = cov(X_t, X_{t+l})$, $\lambda$ a frekvencia jele és $\lambda \in [0, 1]$.*

**3.2.3. Definíció.** $f_2(\lambda) = \frac{1}{2\pi}\sum_{l=-\infty}^{\infty} R(l)e^{-i\lambda l}$,

*$R(l)$ $X_t$ idősor l-edik autokovarianciája, azaz $R(l) = cov(X_t, X_{t+l})$, $\lambda$ pedig a frekvencia és $\lambda \in [0, 2\pi]$.*

A két definíció ekvivalens egymással.

**3.8. Megjegyzés.** $f_1(\lambda) = \sum_{l=-\infty}^{\infty} R(l)e^{-2\pi i\lambda l}$

$f_1(\lambda)$ *:* $[0, 1] \mapsto \mathbb{R}$ *függvény,*

$f_2(\lambda) = \frac{1}{2\pi}\sum_{l=-\infty}^{\infty} R(l)e^{-i\lambda l}$

$f_2(\lambda)$ *:* $[0, 2\pi] \mapsto \mathbb{R}$ *függvény. A két függvény integrálja egyenlő, azaz:*

$\int\limits_0^1 f_1(\lambda)d\lambda = \int\limits_0^{2\pi} f_2(\lambda)d\lambda$

$\int\limits_0^1 (\sum_{l=-\infty}^{\infty} R(l)e^{-2\pi i\lambda l})d\lambda = \int\limits_0^{2\pi} (\frac{1}{2\pi}\sum_{l=-\infty}^{\infty} R(l)e^{-i\lambda l})d\lambda.$

## 3.3. Spektrum becslése, a tapasztalati autokovarianciából

A spektrum becslésénél a spektrum definícióját $f(\lambda) = \sum_{l=-\infty}^{\infty} R(l)e^{-2\pi i\lambda l}$ használjuk fel úgy, hogy $R(l)$ autokovariancia helyére a tapasztalati autokovarianciát, $\hat{R}(l)$ írjuk be.

**3.1. Állítás.** *A periodogram jele: $I(\lambda)$, és a Fourier-transzformáció négyzetes alakjaként írható fel.*

**3.9. Megjegyzés.** *Esetünkben a $I(\lambda)$ periodogram a diszkrét Fourier-transzformáció négyzetes alakja $\lambda = \frac{k}{n}$ frekvenciával.*

**3.10. Megjegyzés.** $\lambda = \frac{k}{n}$ *Fourier frekvenciának nevezzük.*

**3.2. Állítás.** *$E(I(\lambda)) \to f(\lambda)$, ha a megfigyelés szám nő, $n \to \infty$. $I(\lambda)$ aszimptotikusan torzítatlan becslése $f(\lambda)$ spektrumnak.*

**3.1. Tétel.** *Ha $X_t$ idősor autokovarianciájára igaz az, hogy $\sum_{l=-\infty}^{\infty} |R(l)| < \infty$, ekkor a spektrálsűrűség-függvény vagy spektrum így definiálható:*
*$f(\lambda) = \sum_{l=-\infty}^{\infty} R(l)e^{-2\pi i\lambda l}$, minden $\lambda \in \mathbb{R}$.*

Legyen a tapasztalati autokovariancia a következőképpen definiálva:
$\hat{R}(l) = \frac{1}{n}\sum_{t=1}^{n-|l|}(x_{t+|l|} - \bar{x})(x_t - \bar{x})$, minden $-n < l < n$-re.
Az így meghatározott tapasztalati autokovarianciát beírva spektrum definíciójába ezt kapjuk:
$\hat{f}(\lambda) = \sum_{l=-n+1}^{n-1} \hat{R}(l)e^{-2\pi i\lambda l}$, minden $\lambda \in [-1/2, 1/2] - re$.

**Diszkrét Fourier-transzformáció**

Definiáljuk $(x_1, \ldots, x_n)$ sorozat diszkrét Fourier-transzformációját $(X(\lambda_0), \ldots, X(\lambda_{n-1}))$-ként, azaz:

$X(\lambda_k) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-2\pi\lambda_k t}$,

$\lambda_k = k/n$ diszkrét Fourier frekvencia, $k = 0, 1, 2, \ldots, n-1$-re.

**3.11. Megjegyzés.** $\{\lambda_k : k = 0, 1, 2, \ldots, n-1\}$, $\lambda \in [0, 1]$ *diszkrét frekvencia.*

**3.12. Megjegyzés.** *Fourier-transzformáció lineáris transzformáció.*

Vegyük a következő halmazt $\mathbb{C}^n$ komplex téren:

$\{\phi_j : j = 0, 1, 2, \ldots, n-1\}$, ahol $\phi_j$ és $\phi_k$ ortonormált n-dimenziós vektorok $\mathbb{C}^n$-en, azaz:

$$\langle \phi_j, \phi_k \rangle = \begin{cases} 1, & \text{ha } j = k \\ 0, & \text{különben} \end{cases}$$

Ennek alapján $x = (x_1, x_2, \ldots, x_n)'$ vektor előállítható ebben az új ortonormált bázisban:

$x = \sum_{j=0}^{n-1} \langle \phi_j, x \rangle \phi_j$.

Legyen most a

$\phi_j = e_j = \frac{1}{\sqrt{n}}(e^{2\pi i\lambda_j}, e^{2\pi 2i\lambda_j}, e^{2\pi 3i\lambda_j}, \ldots, e^{2\pi ni\lambda_j})'$ n-dimenziós vektor.

Tehát vegyük a következő halmazt:

$\{e_j = \frac{1}{\sqrt{n}}(e^{2\pi i\lambda_j}, e^{2\pi 2i\lambda_j}, \ldots, e^{2\pi ni\lambda_j})' : j = 0, 1, 2, \ldots, n-1\}$, ennek elemei ortonormáltak:

$\langle e_j, e_k \rangle = \frac{1}{n} \sum_{t=1}^{n} e^{2\pi it(\lambda_k - \lambda_j)}$

$= \frac{1}{n} \sum_{t=1}^{n} (e^{2\pi i(\frac{k-j}{n})})^t$

$$= \begin{cases} 1, & \text{ha } j = k \\ \frac{1}{n} e^{2\pi i \frac{(k-j)}{n} \frac{1-(e^{2\pi i \frac{(k-j)}{n}})^n}{1-e^{2\pi i \frac{(k-j)}{n}}}}, & \text{különben} \end{cases}$$

$$= \begin{cases} 1, & \text{ha } j = k \\ 0, & \text{különben} \end{cases}$$

(A mértani sor összeg képlete alapján.)

Ezért $x \in \mathbb{C}^n$ vektor így áll elő:

$x = \sum_{j=0}^{n-1} \langle \phi_j, x \rangle \phi_j$

$= \sum_{j=0}^{n-1} \langle e_j, x \rangle e_j$

$= \sum_{j=0}^{n-1} X(\lambda_j) e_j$, ahol $X(\lambda_j) = \langle e_j, x \rangle$ a Fourier-bázis együtthatója.

$X(\lambda_j)$ felbontható a következőképpen:

$X(\lambda_j) = \langle e_j, x \rangle$

$= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} x_t e^{-2\pi\lambda_j t}$

$= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \cos(2\pi t \lambda_j) x_t - i \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sin(2\pi t \lambda_j) x_t$

$= X_{cos}(\lambda_j) - i X_{sin}(\lambda_j)$,

ahol $X_{cos}$ és $X_{sin}$ $x$ vektor koszinusz- és szinusz "transzformáltja".

## 3.4. Periodogram

**3.4.1. Definíció.** *$I(\lambda_j)$ periodogram meghatározása a következő:*

$I(\lambda_j) = |X(\lambda_j)|^2$.

Tehát kifejtve:

$I(\lambda_j) = |X(\lambda_j)|^2 = \frac{1}{n}|\sum_{t=1}^{n} e^{-2\pi i t \lambda_j} x_t|^2 =$

$= X_{cos}^2(\lambda_j) + X_{sin}^2(\lambda_j)$.

Írjuk fel most $x$ komplex értékű vektor önmagával vett skaláris szorzatát, azaz:

$\langle x, x\rangle = x^* x$, ahol $x^*$ az $x$ vektor konjugáltjának transzponáltja.

Tehát:

$\langle x, x\rangle = x^* x = (\sum_{j=0}^{n-1} X(\lambda_j) e_j)^* (\sum_{j=0}^{n-1} X(\lambda_j) e_j) =$

$= \sum_{j=0}^{n-1} |X(\lambda_j)|^2 = \sum_{j=0}^{n-1} I(\lambda_j)$.

Diszkrét eset. Ha $x$ átlaga 0, akkor $x$ tapasztalati varianciája ezzel egyenlő:

$\sigma_x^2 = \frac{1}{n}\sum_{t=1}^{n} x_t^2 = \frac{1}{n}\sum_{j=0}^{n-1} I(\lambda_j)$.

Ehhez hasonlóan $x$ varianciája folytonos esetben így néz ki:

$\sigma_x^2 = R_x(0) = \int\limits_{-1/2}^{1/2} f_x(\lambda) d\lambda$.

**3.13. Megjegyzés.** *$I(\lambda_j)$ periodogram az $f(\lambda)$ spektrálsűrűség-függvény diszkrét változata $\lambda_j = \frac{j}{n}$ Fourier frekvenciára.*

(A fenti megjegyzés szerint a $\lambda$ szerinti integrál "diszkrét változata" az "$\frac{1}{n}\sum$-ának.)

### 3.4.1. Spektrum becslése periodogrammal

**3.3. Állítás.** *A periodogram, $I(\lambda_j)$, $\lambda_j = \lambda$, megegyezik a tapasztalati autokovariancia Fourier-transzformáltjával, azaz $f(\lambda)$ spektrummal.*

**3.14. Megjegyzés.** *Kivétel, ha $\lambda_0 = 0$ és $x$ várható értéke nem egyenlő 0.*

*A többi esetben mind igaz, azaz $\bar{x} = 0$-ra és $\lambda_j$-re minden $j = 0, 1, 2, 3, \ldots, n-1$.*

**3.1. Bizonyítás.** $I(\lambda_j) = |X(\lambda_j)|^2 = \frac{1}{n}|\sum_{t=1}^{n} e^{-2\pi i t\lambda_j} x_t|^2$ *definíció szerint*

*Tegyük fel, hogy $x$ átlaga 0, azaz $\bar{x} = 0$, ekkor $\bar{x}$ beírható:*

$I(\lambda_j) = \frac{1}{n}|\sum_{t=1}^{n} e^{-2\pi i t\lambda_j}(x_t - \bar{x})|^2$

$= \frac{1}{n}(\sum_{t=1}^{n} e^{-2\pi i t\lambda_j}(x_t - \bar{x}))(\sum_{t=1}^{n} e^{-2\pi i t\lambda_j}(x_t - \bar{x}))$

$= \frac{1}{n}\sum_{t,s} e^{-2\pi i(t-s)\lambda_j}(x_t - \bar{x})(x_s - \bar{x})$

*$\hat{R}(l)$ tapasztalati kovariancia definíciója a következő volt:*

$\hat{R}(l) = \frac{1}{n}\sum_{t=1}^{n-|l|}(x_{t+|l|} - \bar{x})(x_t - \bar{x})$, *ahol* $-n < l < n$

*Ezért:*

$I(\lambda_j) = \frac{1}{n}\sum_{t,s} e^{-2\pi i(t-s)\lambda_j}(x_t - \bar{x})(x_s - \bar{x})$

*Ha $\lambda_j \neq 0$, akkor $\sum_{t=1}^{n} e^{-2\pi i t\lambda_j} = 0$,*

*amit akkor láttunk be, amikor a Fourier-bázis ortonormalitását igazoltuk, így:*

$I(\lambda_j) = \sum_{l=-n+1}^{n-1} \hat{R}(l) e^{-2\pi i l\lambda_j} = \hat{f}(\lambda_j)$

### 3.4.2. A periodogram aszimptotikus viselkedése

$I(\lambda)$ periodogram $n$-től függően (a megfigyelésszám növekedésével) $E(I(\lambda)) \to f(\lambda)$, ha $n \to \infty$.

Legyen $\{X_1, X_2, \ldots, X_n\}$ Gauss fehérzaj, azaz $\{X_1, X_2, \ldots, X_n\}$ elemei független azonos eloszlású normálisak 0 várható értékkel és $\sigma^2$ varianciával.

A periodogramnál definiált $X_{cos}(\lambda_j)$ és $X_{sin}(\lambda_j)$:

$X_{cos}(\lambda_j) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} \cos(2\pi t\lambda_j) x_t$ és

$X_{sin}(\lambda_j) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sin(2\pi t\lambda_j)x_t$.

$X_{cos}(\lambda_j)$ és $X_{sin}(\lambda_j)$ normálisak 0 várható értékkel.

Nézzük meg a varianciájukat:

Amit eddig tudunk:

$X(\lambda_j) = X_{cos}(\lambda_j) - iX_{sin}(\lambda_j)$, valamint

definíció szerint: $I(\lambda_j) = |X(\lambda_j)|^2 = X_{cos}^2(\lambda_j) + X_{sin}^2(\lambda_j)$.

Ezeket felhasználva:

$Var(X_{cos}(\lambda_j)) = cov(X_{cos}(\lambda_j), X_{cos}(\lambda_j))$

Ide beírva definíció szerint: $X_{cos}(\lambda_j) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \cos(2\pi t\lambda_j)x_t$

$= cov(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \cos(2\pi t\lambda_j)x_t, \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \cos(2\pi t\lambda_j)x_t)$

$= \frac{\sigma^2}{n} \sum_{t=1}^{n} \cos^2(2\pi t\lambda_j)$

Megjegyzés: $2\cos^2(\alpha) = \cos(2\alpha) + 1$ ezt használjuk fel, így tehát:

$= \frac{\sigma^2}{2n} \sum_{t=1}^{n} (\cos(4\pi t\lambda_j) + 1) = \frac{\sigma^2}{2}$.

Hasonlóan:

$Var(X_{sin}(\lambda_j)) = \frac{\sigma^2}{2}$.

A kovarianciájuk:

$cov(X_{cos}(\lambda_j), X_{sin}(\lambda_j)) = \frac{\sigma^2}{n} \sum_{t=1}^{n} \cos(2\pi t\lambda_j)\sin(2\pi t\lambda_j)$

$= \frac{\sigma^2}{2n} \sum_{t=1}^{n} \sin(4\pi t\lambda_j) = 0$.

Ehhez hasonlóan a többi is 0:

$cov(X_{cos}(\lambda_j), X_{cos}(\lambda_k)) = 0$,

$cov(X_{sin}(\lambda_j), X_{sin}(\lambda_k)) = 0$ minden $j \neq k$-ra.

Tehát összefoglalva:

ha $\{X_1, X_2, \ldots, X_n\} \sim N(0, \sigma^2)$ és $f(\lambda) = \sigma^2$, akkor

$X_{cos}(\lambda_j)$ és $X_{sin}(\lambda_j) \sim N(0, \sigma^2/2)$. Így:

$\frac{2}{\sigma^2} I(\lambda_j) = \frac{2}{\sigma^2}(X_{cos}^2(\lambda_j) + X_{sin}^2(\lambda_j)) \sim \chi_2^2$.

Vagyis Gauss fehérzajra a periodogram $\chi_2^2$-eloszlású.

Általánosabban:

$\{X_t\}$ normális eloszlású folyamat vagy $\{X_t\}$ lineáris folyamat gyorsan lecsengő autokovarianciával, valamint $X_{cos}(\lambda_j)$, $X_{sin}(\lambda_j)$ aszimptotikusan függetlenek és normális eloszlásúak 0 várható értékkel és $\sigma^2/2 = f(\lambda_j)/2$ varianciával. A frekvencia legyen most $\lambda$, és $n$-re legyen $\hat{\lambda}^{(n)} = j/n$ úgy, hogy $\min_j |\lambda - j/n|$, hasonlóan a Fourier frekvenciához.

Ekkor $\hat{\lambda}^{(n)} \to \lambda$.

Hasonló feltételekkel: aszimptotikusan normálisak és függetlenek a "szinusz, koszinusz transzformáltak", $Xcos(\lambda_j)$ és $Xsin(\lambda_j)$, így $f(\hat{\lambda}^{(n)}) \to f(\lambda)$.

Tehát esetünkben, visszatérve a periodogramra:

$\frac{2}{f(\lambda)} I(\hat{\lambda}^{(n)}) = \frac{2}{f(\lambda)}(X_{cos}^2(\hat{\lambda}^{(n)}) + X_{sin}^2(\hat{\lambda}^{(n)})) \xrightarrow{d} \chi_2^2$, azaz

$\frac{2}{f(\lambda)} I(\hat{\lambda}^{(n)})$ eloszlásban $\chi_2^2$-hez.

Utolsó lépés következik $I()$ periodogram és spektrálsűrűség-függvény kapcsolatának elemzésében, mégpedig, hogy $E(I(\hat{\lambda}^{(n)})) \to f(\lambda)$.

Tehát:

$E(I(\hat{\lambda}^{(n)})) = E(X_{cos}^2(\hat{\lambda}^{(n)}) + X_{sin}^2(\hat{\lambda}^{(n)}))$

$= E(\frac{2}{f(\lambda)} \frac{f(\lambda)}{2}(X_{cos}^2(\hat{\lambda}^{(n)}) + X_{sin}^2(\hat{\lambda}^{(n)})))$

$= \frac{f(\lambda)}{2} E(\frac{2}{f(\lambda)}(X_{cos}^2(\hat{\lambda}^{(n)}) + X_{sin}^2(\hat{\lambda}^{(n)})))$

$= \frac{f(\lambda)}{2} E(\frac{2}{f(\lambda)} I(\hat{\lambda}^{(n)})) \to \frac{f(\lambda)}{2} E(Z_1^2 + Z_2^2) = f(\lambda)$,

ahol, $Z_1$ és $Z_2 \sim N(0,1)$ független normális eloszlású valószínűségi változók.

Tehát $E(I(\hat{\lambda}^{(n)})) \to f(\lambda)$, $n \to \infty$.

## 3.5. Fehérzaj, AR(1) és MA(1) folyamatok spektruma

### 3.5.1. A fehérzaj spektruma

Legyen $\{\varepsilon_t\}$ fehérzaj folyamat, azaz azonos eloszlású korrelálatlan valószínűségi változó.

Legyen $E(\varepsilon_t) = 0$ minden $t$-re.

$R(0) = \sigma_\varepsilon^2$ és $R(t) = 0$ minden $t > 0$-ra.

A levezetéseknél a spektrum 2. a korábban felírt definícióját fogom használni, vagyis ezt:

**3.5.1. Definíció.** $f(\lambda) = \frac{1}{2\pi}\sum_{t=-\infty}^{\infty} R(t)e^{-it\lambda}$, $\lambda \in [0, 2\pi]$

Ekkor $\varepsilon_t$ spektruma:

$f_\varepsilon(\lambda) = \frac{1}{2\pi}\sum_{t=-\infty}^{\infty} R(t)e^{-it\lambda} = \frac{\sigma_\varepsilon^2}{2\pi}$.

### 3.5.2. AR(1) spektruma

`AR(1)` spektruma:

$f_X(\lambda) = \frac{1}{2\pi}\sum_{t=-\infty}^{\infty} R(t)e^{-it\lambda}$

$= \frac{1}{2\pi}[\sum_{t=-\infty}^{0} R(t)e^{-it\lambda} + \sum_{t=0}^{\infty} R(t)e^{-it\lambda} - R(0)]$

$= \frac{1}{2\pi}R(0)[\sum_{t=0}^{\infty} \alpha^t e^{-it\lambda} + \sum_{t=0}^{\infty} \alpha^t e^{it\lambda} - 1]$

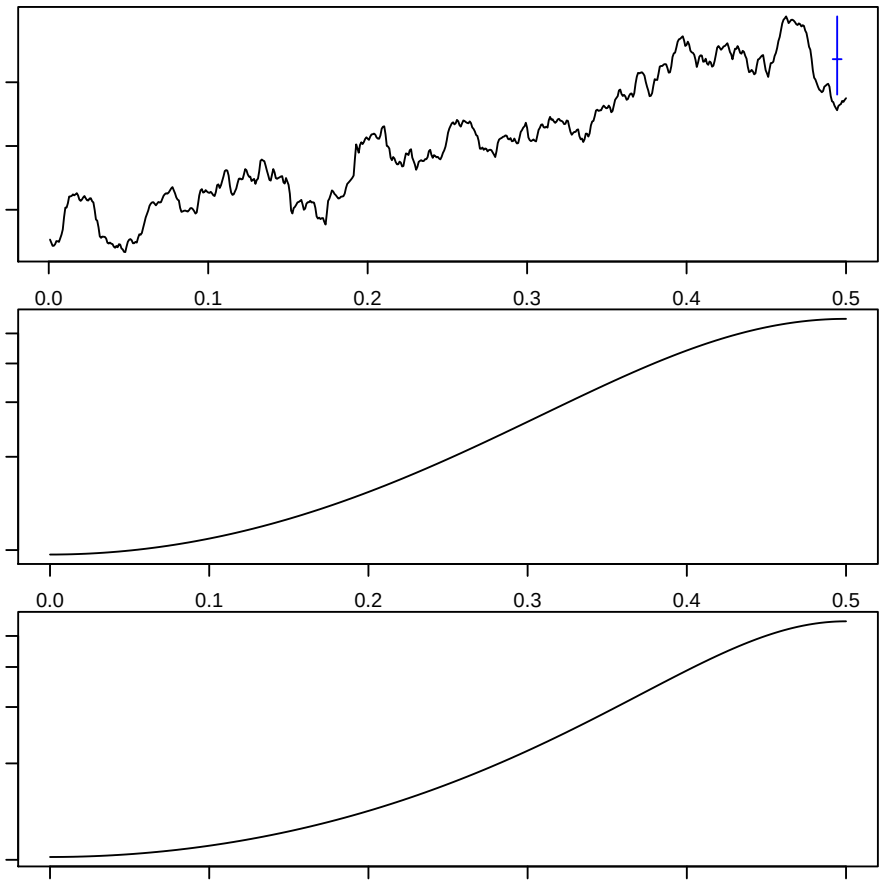
megjegyzés: $R(t) = R(0)\alpha^t = \sigma_X^2\alpha^t$ és $\sigma_X^2 = \frac{\sigma_\varepsilon^2}{1-\alpha^2}$

$= \frac{\sigma_X^2}{2\pi}[\frac{1}{1-\alpha e^{-i\lambda}} + \frac{1}{1-\alpha e^{i\lambda}} - 1]$

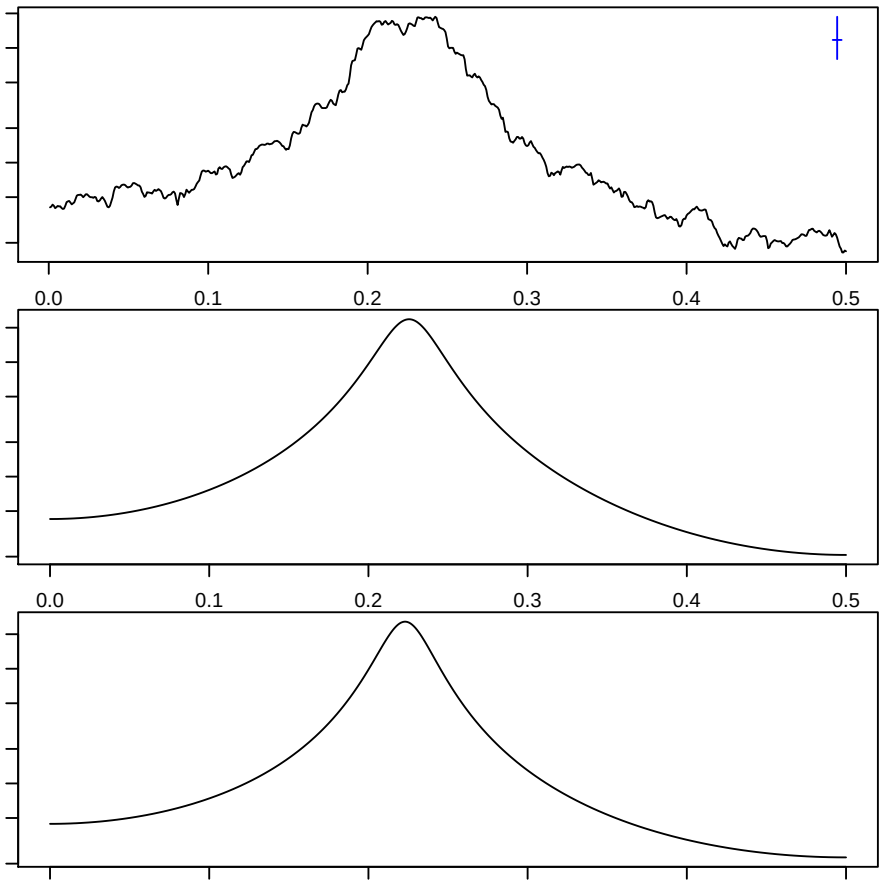
közös nevezőre hozunk, $\sigma_X^2$ helyére beírjuk $\frac{\sigma_\varepsilon^2}{1-\alpha^2}$-t és egyszerűsítünk:

$= \frac{\sigma_\varepsilon^2}{2\pi(1-\alpha^2)}[\frac{1-\alpha^2}{|1-\alpha e^{i\lambda}|^2}]$

$= \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{|1-\alpha e^{i\alpha}|^2}$.




3.4. ábra. AR(1), fentről lefelé: periodogram, illesztett folyamat spektruma, tényleges spektrum




3.5. ábra. AR(2), fentről lefelé: periodogram, illesztett folyamat spektruma, tényleges spektrum

### 3.5.3. AR(p) spektruma

`AR(p)` spektruma:

$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{|P(e^{i\lambda})|^2}$

$= \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{|1-\alpha e^{i\lambda}|^2}$

$P(e^{i\lambda})$ $AR(p)$ karakterisztikus polinomja.

**3.15. Megjegyzés.** $P(X) = \sum_{k=0}^{p} \widetilde{\alpha_k} x^{p-k} = \sum_{k=0}^{p} \widetilde{\alpha_k} e^{i\lambda(p-k)}$, $x = e^{i\lambda}$,

$\widetilde{\alpha_0} = 1$ *és* $\widetilde{\alpha_k} = -\alpha_k$

### 3.5.4. MA(q) spektruma

`MA(q)` spektruma:

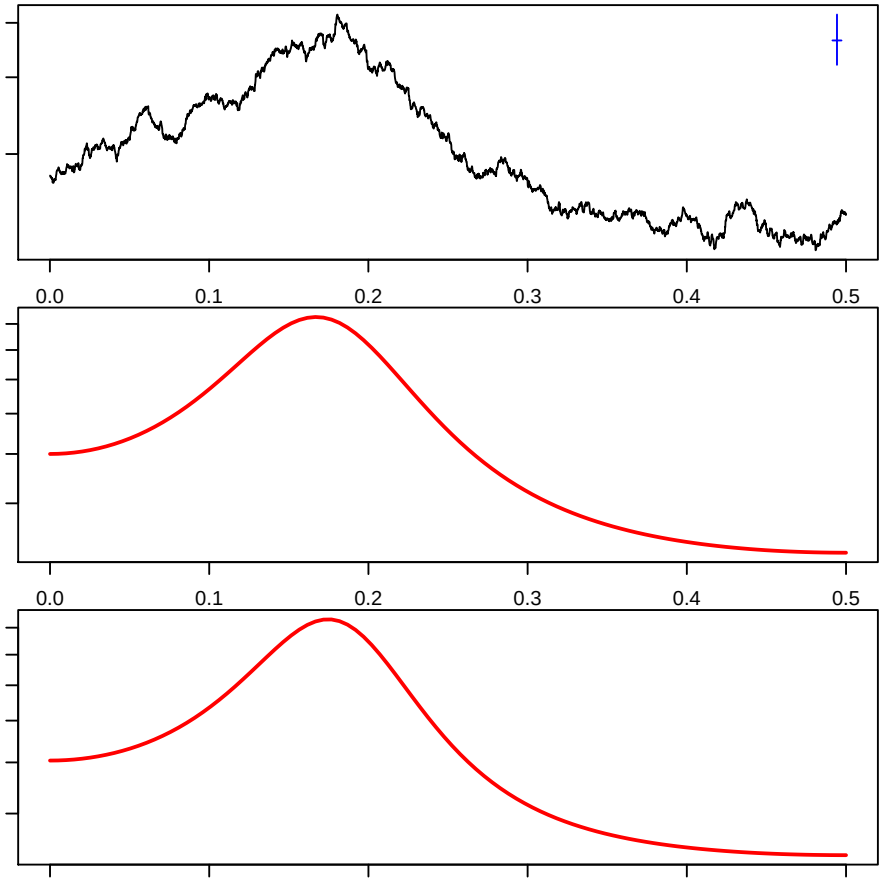
$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}|Q(e^{i\lambda})|^2$,

ahol $Q(e^{i\lambda})$ mozgóátlag folyamat karakterisztikus polinomja.

### 3.5.5. ARMA(p,q) spektruma

`ARMA(p,q)` spektruma:

$f_X(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi}\frac{|Q(e^{i\lambda})|^2}{|P(e^{i\lambda})|^2}$.




3.6. ábra. ARMA, fentről lefelé: periodogram, illesztett folyamat spektruma, tényleges spektrum

# 4. fejezet

# Zajos ARIMA, FARIMA

## 4.1. Zajos ARIMA és FARIMA folyamatok

### 4.1.1. Bevezetés

Ez a részt C.W.J. Granger: Long Memory Relationships and The Aggregation of Dynamic models című 1980-as cikke alapján tárgyalom.

Legelőször tegyük fel, hogy $X(t)$ idősort nulla várható értékű és $\sigma^2$ varianciájú $\varepsilon(t)$ fehérzaj sorozat generálja, a következő lineáris szűrőn keresztül:

$X(t) = a(B)\varepsilon(t)$, ahol $B$ operátor az előzőekben bevezetett késleltetési operátort jelenti, vagyis

$B\varepsilon(t) = \varepsilon(t-1)$ és $B^k\varepsilon(t) = \varepsilon(t-k)$,

valamint tegyük fe, hogy $a(B)$ lineáris filtert felírhatjuk így is:

$a(B) = (1-B)^{-d}a'(B)$, illetve $a'(B) = (1-b)^d \cdot a(B)$,

ahol $a'(B)$ véges polinom vagy végtelen polinom.

**4.1. Megjegyzés.** *A fent bevezetett $X(t)$ idősor várható értéke szintén 0 az őt generáló fehérzaj 0 várható értéke miatt.*

Kezdeti modell:

$$X(t) = a(B)\varepsilon(t), \text{ és } B^k\varepsilon(t) = \varepsilon(t-k) \tag{4.1}$$

$E(X(t)) = 0$, $E(\varepsilon(t)) = 0$ minden $t$-re és $D^2(\varepsilon(t) = \sigma^2$

$$a(B) = (1-B)^{-d}a'(B), \text{ ahol } a'(z)\text{-nek nincsen gyöke, ha } z = 0. \tag{4.2}$$

Ekkor $X(t)$-re azt mondjuk, hogy 'd-edrendben integrált', jelölése: $X(t) \sim I(d)$.

**4.2. Megjegyzés.** *$X(t) \sim I(d)$-nél a $d$ nem feltétlenül egész számot jelöl.*

**4.1.1. Definíció.** $X'(t) = (1-B)^d X(t) = a'(B)\varepsilon(t)$.

*Hiszen: $a'(B) = (1-b)^d \cdot a(B)$, ekkor $X'(t) \sim I(0)$, a már említett $a'(B)$ tulajdonságai miatt.*

Az $a(B)$ lineáris filtert 'd-edrendű integráló szűrűnek' nevezzük.

Ha $a'(B)$ két véges polinom hányadosa, amelyeknek rendjei legyenek $p$ és $q$, valamint ha $d$ egész szám, akkor $X(t)$ idősorunk ARIMA(p,d,q) lesz.

A további modelleknél nem követeljük meg, hogy $d$ egész legyen.

Következzenek a frakcionálisan (tört) differenciált folyamatok.

Az előzőekben használt 'szokásos' differenciáló operátor a $(1-B)$ volt. Most tegyük fel, hogy a $\alpha(B)$ operátor olyan, hogy ha kétszer használjuk, akkor a 'szokásos' differenciáltat adja, vagyis: $\alpha^2(B) = (1-B)$ (ekkor azt mondhatjuk, hogy $d = 1$). Ilyen operátor egyértelműen létezhet, és ha $\alpha(B)$-t csak egyszer használjuk, akkor erre úgy gondolhatunk, mint egy 'félszer differenciáló' operátor, vagyis olyan mint $d = 1/2$-es frakcionális differenciált.

Integrált sorozat egyik 'kívánatos' tulajdonsága, hogy a frakcionális differenciálás után stacionárius ARMA-t kapjunk.

$X(t)$ folyamat, amit (5.1) és (5.2) egyenlet ír le, tulajdonságai következnek:

$X(t)$ spektruma:

$f_x(\omega) = \frac{1}{|1-z|^{2d}}|a'(z)|^2\frac{\sigma^2}{2\pi}$, ahol $z = e^{i\omega}$.

Ebből következik, hogy elég kicsi $\omega$-ra:

$$f_x(\omega) \simeq c \cdot \omega^{-2d},\ c = \frac{\sigma^2}{2\pi}(a'(1))^2 \tag{4.3}$$

**4.3. Megjegyzés.** *Stacionárius ARMA folyamatokra a spektrum így alakul $f_x(\omega) \simeq c$, elég kicsi $\omega$-ra.*

**4.4. Megjegyzés.** *Viszont ARIMA(p,1,q)-ra a spektrum* (4.3) *egyenlet szerint alakul $d = 1$-gyel.*

Így viszont az látszik (az előző két megjegyzésből), hogy ez a modell nem jó megközelítése a frakcionálisan integrált modellhez.

**4.5. Megjegyzés.** *Meg kell még azt is jegyezni ennél a résznél, hogy a hosszabb-futású előrejelzési céloknál a spektrum alacsony frekvenciatartománya a legfontosabb.*

Tekintsük akkor most a következő új esetet, ahol az $a(B)$ integráló filter $d$-ed rendben, tehát a modell:

$$X(t) = (1-B)^{-d}\varepsilon(t) \tag{4.4}$$

Granger és Joyeux megmutatta, hogy $X(t)$ és $X(t-1)$ közti autokovarianciafüggvény a következőképpen alakul:

$cov(X(t), X(t-1)) = \frac{\sigma_\varepsilon^2}{2\pi}\sin(\pi d)\frac{\Gamma(k+d)}{\Gamma(k+1-d)}\Gamma(1-2d)$,

ahol $d < 1/2$.

$X(t)$ varianciája $d$ paraméterrel együtt növekszik, és ha $d \geq 1/2$, akkor végtelen lesz. $X(t)$ és $X(t-k)$ közti (k-ad rendű) autokorreláció-függvény így alakul:

$$\rho_k = corr(X(t), X(t-k)) = \frac{\Gamma(1-d)}{\Gamma(d)}\frac{\Gamma(k+d)}{\Gamma(k+1-d)}\ ,\ d < 1/2\text{-re és } d \neq 0. \tag{4.5}$$

Ha $d = 0$, akkor $k > 0$-ra $\rho_k = 0$ (fehérzaj esete).

Írjuk fel $X(t)$ $MA(\infty)$ és $AR(\infty)$ előállítását.

$MA(\infty)$: $X(t) = \sum_{j=0}^{\infty} b_j \varepsilon(t-j)$

$AR(\infty)$: $\varepsilon(t) = \sum_{j=0}^{\infty} a_j X(t-j)$

$X(t)$ folyamat (4.4) egyenlet szerint lett generálva, $a_j$ és $b_j$ együtthatók pedig így néznek ki:

$$b_j = \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} \text{ ,} j \geq 1 \text{ , } j \neq 0 \text{ és } a_j = \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)} \text{ ,} j \geq 1. \tag{4.6}$$

$\frac{\Gamma(j+a)}{\Gamma(j+b)}$ jól közelíthető $j^{a-b}$-vel elég nagy $j$-re, ezt felhasználva az előző 2 egyenlet, (4.5), (4.6), jól közelíthető a következő módon:

$$\rho_j \simeq A_1 j^{2d-1} \tag{4.7}$$

$$b_j \simeq A_2 j^{d-1} \tag{4.8}$$

$$|a_j| \simeq A_3 j^{d+1} \tag{4.9}$$

$A_1$, $A_2$ és $A_3$ konstansok. Ha $X(t)$ folyamatot a (5.1) és (5.2) egyenletek generálják, a (4.7), (4.8), (4.9) egyenletek elég nagy $j$-re megtartják az autokorreláció, a mozgó átlag és az autoregresszív paramétereit, de $A_1$, $A_2$, $A_3$ konstansok változhatnak. Az, hogy $\rho_j$ és $b_j$ lassan lecsengőek tetszőleges $ARMA(p,q)$-re, azt jelenti, hogy hosszú emlékezetű a folyamat megfelelő $d > 0$-ra, ez látszik (4.3) egyenletből is, ami alacsony frekvenciára határozta meg a spektrumot. Így az alacsony frekvenciák vannak a központban a hosszútávú előrejelzésnél, cél megtalálni a helyes modellt, ami megfeleltethető a (4.3) egyenlet spektrum közelítésével. Valamint meg kell még jegyezni, hogy $d \to \infty$ és $b_j \to 0$ minden $j > 0$-ra, így kapjuk meg a fehérzajt.

Ha $X(t) \sim I(d)$, ekkor $Y(t)$ új folyamat legyen ez:

$Y(t) = (1-B)^{-d'} X(t)$.

Így tehát, ha $X(t) \sim I(d)$,akkor $Y(t) \sim I(d + d')$.

Ha $X_1(t) \sim I(d_1)$ és $X_2(t) \sim I(d_2)$, ahol $X_1(t)$ és $X_2(t)$ egymástól függetlenek, akkor $X_1(t)$ és $X_2(t)$ összege legyen $\bar{X}$, és $\bar{X} \sim I(max(d_1, d_2))$, vagyis:
$\bar{X} = X_1(t) + X_2(t) \sim I(max(d_1, d_2))$.

Ha $X(t)$-re és $Y(t)$-re a következő igaz:
$X(t) = a(B)Y(t) + e_t$, ahol $X(t) \sim I(d_x)$ és $Y(t) \sim I(d_y)$, $a(B)$ d-ed rendű integráló filter, akkor $e_t \sim I(d_e)$ és $d_x = max(d_e, d + d_y)$ és ha $d_y > d_x$, akkor $d < 0$.

### 4.1.2. Független sorok aggregációja

Két független folyamat összege:
Legyen $X_1(t)$ és $X_2(t)$ független folyamatok:

$$X_j(t) = \alpha_j X_j(t-1) + \varepsilon_j(t),\ j = 1, 2 \tag{4.10}$$

, ahol $\varepsilon_1(t)$ és $\varepsilon_2(t)$ páronként független, nulla várható értékű fehérzaj.
Valamint $X_1(t)$ és $X_2(t)$ összege $\bar{X} \sim ARMA(2, 1)$. Az autoregresszív része a modellnek az $(1 - \alpha_1 B)(1 - \alpha_2 B)$.

N darab független folyamat összege:
Ha $N$ darab független különböző $\alpha$ paraméterű $AR(1)$-et adunk össze, akkor az összeg-folyamat, $\bar{X} \sim ARMA(N, N-1)$.

A legtöbb mikroökonómiai változó aggergált (összetett) és számos mikro-változóból tevődnek össze, mint például: 1 főre jutó bruttó jövedelem, munkanélküliség, nem tartós jószágok fogyasztása, megtakarítások és a profit. Annak ellenére, hogy ezek a makroökonómiai változók komponensei nem függetlenek, a fenti eredmények kiterjeszthető rájuk kellően nagy számú paraméter esetén.

Az összeg-folyamat spektruma, jele: $f(\bar{\omega})$ a következő módon adható meg:
$X(\bar{t}) = \sum_{j=1}^{N} X_j(t)$, ahol $X_j(t)$ minden $j$-re $AR(1)$ modell.
$X_j(t)$ spektruma:

$$f_j(\omega) = \frac{1}{|1-\alpha_j z|^2} \cdot \frac{var(\varepsilon_j(t)}{2\pi},\ z = e^{-i\omega}, \tag{4.11}$$

Ekkor $\bar{X}$ spektruma:
$f(\bar{\omega}) = \sum_{j=1}^{N} f_j(\omega)$, $f_j(\omega)$-k függetlenek.

Ha $\alpha_j$-k véletlen változók $F(\alpha)$ eloszlásfüggvénnyel, és $var(\varepsilon_j(t))$ függetlenek $\alpha$-ától.
Az $\bar{x}$ spektruma nem csak $f_j(\omega)$-k összegeként írható fel, hanem így is megközelíthető:

$$f(\bar{\omega}) = \frac{N}{2\pi} E[var(\varepsilon_j(t))] \cdot \int \frac{1}{|1-\alpha z|^2} dF(\alpha). \tag{4.12}$$

Ha $F(\alpha)$ -1 és 1 közötti diszkrét valószínűségi változók eloszlásfüggvénye, és ha $\alpha$ pontosan $m$ darab értéket vesz fel, akkor $f(\bar{\omega})$ egy $ARMA(m, m-1)$ folyamat spektruma lesz. Azonban ha $\alpha$ más értékeket is felvevő, folytonos változó, ez nem teljesül.

Továbbiakban $\alpha$ $\beta$-eloszlású változó (0,1)-en.

A$\beta$-eloszlás egy speciális alakját használjuk a következő egyenletben:

$$\begin{aligned} dF(\alpha) &= \frac{2}{B(p,q)}\alpha^{2(p-1)}(1-\alpha^2)^{q-1}d\alpha, \text{ ha } 0 \leq \alpha \leq 1 \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{különben} \end{aligned} \tag{4.13}$$

A fenti egyenletben szereplő $dF(\alpha)$ a béta-eloszlás sűrűségfüggvénye, $B(p,q)$ béta-függvény.

**4.1.2. Definíció.** *$B(p,q)$ béta-függvény, ha $B(p,q) = \int\limits_0^1 x^{p-1}(1-x)^{q-1}$, $p > 0$ és $q > 0$ konstansok.*

**4.1. Állítás.** *$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$, ahol $\Gamma()$ gamma-függvény.*

**4.1.3. Definíció.** *$\Gamma(x) = \int\limits_0^\infty t^{x-1}e^{-t}dt$, ha $x > 0$,*
*$\Gamma(x) = (x-1)\Gamma(x-1)$, ha $x > 1$,*
*$\Gamma(x) = (x-1)!$, ha $x \in \mathbb{Z}_+$, és speciálisan $\Gamma(1/2) = \sqrt{\pi}$*

**4.6. Megjegyzés.** $\frac{1}{|1-\alpha z|^2} = \frac{1}{(1-\alpha)^2}[\frac{1+\alpha z}{1-\alpha z} + \frac{1+\alpha \bar{z}}{1-\alpha \bar{z}}]$
*valamint*
$\frac{1+\alpha z}{1-\alpha z} = 1 + 2\sum_{j=1}^{\infty}(\alpha z)^j$
*így $z^k$ együtthatója $\bar{f}(\omega)$-ban:*
$\frac{2}{B(p,q)}\int\limits_0^1 \alpha^{2p+k-1}(1-\alpha^2)^{q-2}d\alpha$
*ez az együttható $\bar{\mu_k}$ $\bar{X(t)}$ k-adik autokovarianciája.*

*Ha $q > 1$, akkor:*
$\bar{\mu_k} = \frac{B(p+k/2,q-1)}{B(p,q)} = \frac{\Gamma(q-1)}{B(p,q)} \cdot \frac{\Gamma(p+k/2)}{\Gamma(p+k/2+q-1)}$,
*ha $k$ elég nagy, akkor $\bar{\mu_k}$ előáll így: $\bar{\mu_k} = A_4 k^{1-q}$.*

Ebből az következik, hogy:

$\bar{X(t)} \sim I(1-q/2)$, vagyis az integráció rendje: $d = 1 - q/2$.

**4.7. Megjegyzés.** *Ha $q > 1$, akkor $1 - q/2 < 1/2$ és $\bar{X(t)}$ varianciája véges. Viszont, ha $0 < q \leq 1$, akkor $\bar{X(t)}$ varianciája nem véges.*

Az integrálás rendje, $d = 1 - q/2$, nem függ $p$-től, és $dF(\alpha)$ alakja kevésbé fontos kivéve $\alpha = 1$ körül. Ha $\alpha$ értelmezési tartományát megváltoztatjuk $(a, 1)$re, ahol $a > 0$, akkor ez az eredményen nem változtat. A felső határ megváltoztatása okoz eltérést. Ha $b$-re változtatjuk, ahol $b > 1$, akkor $\bar{X}$ szétrobbanó lesz. Ha $b < 1$, akkor béta-eloszlás $(0, b)$-n ez lesz:

$\frac{2}{B(p,q)} \frac{\alpha^{2(p-1)}(b^2-\alpha^2)^{q-1}}{b^{p+q-1}}$, ekkor

$\bar{\mu_k} = A_5 b^k k^{1-q}$, elég nagy $k$-ra.

A fenti egyenlet nem felel meg szigorúan az autokovarianciának bármely véges paraméterű $ARMA$ folyamat esetén, mégis jól közelíthető megfelelően megválasztott $b$-vel:

1. eset, ha $\bar{X}_t \sim I(0)$ folyamat és $\alpha_j$-k független $AR(1)$-es együtthatók,

2. eset, ha $b = 1$ és $q \to \infty$.

Legyen: $x_{jt} = \alpha_j x_{j,t-1} + y_{jt}$,

$y_{jt}$ spektruma: $f_y(\omega, \theta_j)$, $y_{jt}$-k függetlenek.

Ekkor $\bar{f}(\omega) \cong NE_\theta[f_y(\omega, \theta)] \int \frac{1}{|1-\alpha z|^2} dF(\alpha)$.

Ha $\bar{y_t} = \sum_{j=1}^{N} y_{jt}$, akkor $\bar{y_t} \sim I(d_y)$, ekkor $\bar{x_t} = \sum_{j=1}^{N} x_{jt} \sim I(d_y + 1 - q)$.

Kereszt-spektrum, jele $cr_j(\omega)$, $x_{jt}$ és $y_j t$ között ezzel egyenlő:

$cr_j(\omega) = \frac{f_y(\omega,\theta_j}{1-\alpha_j z}$ és $z = e^{-i\omega}$.

Kereszt-spektrum $\bar{x_t}$ és $\bar{y_t}$ között:

$\bar{cr}(\omega) = \sum_{j=1}^{N} cr_j(\omega)$.

A $\bar{cr}(\omega)$ jól közelíthető így:

$\bar{cr}(\omega) \cong NE_\theta[f_y(\omega, \theta)] \int \frac{1}{1-\alpha z} dF(\alpha)$.

Ugyanez átírva:

$\bar{cr}(\omega) \cong NE_\theta[f_y(\omega, \theta)] \sum_{k=0}^{\infty} z^k \frac{B(p+k/2,q)}{B(p,q)}$.

Tehát: $x_t = a(B)y_t + e_t$ alapján $\bar{x_t} = a(B)\bar{y_t} + e_t$

$a(z) = \frac{\bar{cr}(\omega)}{f_y(\omega)}$

$a(B)$ integrálási szűrő , rendje $d = 1 - q$ ((4.8) alapján), $q \neq 1$

**Nem független sorok aggregációja**

Tegyük fel, hogy $x_{jt}$ teljesen összefüggő halmazt alkotnak.

$x_{jt} = \alpha_j x_{j,t-1} + \beta_j W_t$

és legyen

$\bar{x}_t = (\sum_{j=1}^{N} \frac{\beta_j}{1-\alpha_j B})W_t$

valamint ugyanez az $\bar{x}_t$ megközelíthető így:

$\bar{x}_t = NE[\beta](\int \frac{1}{1-\alpha B} dF(\alpha))W_t,$

$\alpha$ és $\beta$ függetlenek.

Az előző rész eredményéből:

$\bar{x}_t \cong NE[\beta](\sum_{k=0}^{\infty} B^k \frac{B(p+k/2,q)}{B(p,q)})W_t.$

A (4.8) egyenlet alapján $W_t \cong I(d_w)$, ekkor:

$\bar{x}_t \sim I(1 - q + d_w)$ és

$\bar{x}_t = a(B)W_t$, ahol $a(B)$ a már jól ismert integrálási szűrő $d = 1 - q$ renddel.

A következő még általánosabb modellben már $x_{jt}$ és $y_{jt}$ közti összefüggést is megjeleníti:

$x_{jt} = \alpha_j x_{j,t-1} + y_{jt} + \beta_j W_j + \varepsilon_{jt}$, ahol

$y_{jt}$, $W_t$, $\varepsilon_{jt}$ egymástól függetlenek minden $j$-re.

$\varepsilon_{jt}$ a fehérzaj, $D(\varepsilon_{jt}) = \sigma^2_{\varepsilon_{jt}}$

$y_{jt}$ spektruma $f_y(\omega, \theta_j)$

ennél a felírásnál nincsen visszautalás, és $x_{jt}$ nem következik $x_{jt}$-ből vagy $W_t$-ből.

Így a fenti eredményekből azt kapjuk, hogy:

1. $\bar{x}_t \sim I(d_x)$, ahol $d_x = \min(1 - q/2 + d_y, 1 - q + d_w, 1 - q/2)$ és $\bar{y}_t \sim I(d_y)$, $W_t \sim I(d_w)$,

továbbá $\alpha$ béta-eloszlású valószínűségi változó (4.11) szerint $P(\alpha = 1) = 0$, ha $q > 0$, és

ha $d_y = d_w = 0$, akkor $d_x < 1$ és $q > 0$.

Viszont ha $d_y = d_w = 0$ mellett $d_x = 1$, akkor $P(\alpha = 1) \neq 0$.

2. transzformáló-függvény modell: $\bar{x}_t$ felírása:

$\bar{x}_t = a_1(B)\bar{y}_t + a_2(B)W_t + e_t$,

$a_1(B)$ és $a_2(B)$ $d = 1 - q$ rendű integrálási szűrők.

Az $\bar{y}_t$ $1 - q + d_y$-nal járul hozzá $d_x$-hez.

Ha $q > 0$, akkor $\bar{y}$ kevésbé járul hozzá $d_x$-hez, így információt vesztünk, ha $\bar{y}$-t használunk $\bar{x}$ megmagyarázásához, így jobb ha $y_{jt}$-ket használunk.

Ha $e_t \sim I(d_x)$ és $d_x > 1 - q + d_w$, akkor $d_e = d_x$, de ha $d_x = 1 - q + d_w$, akkor $d_e \leq d_x$, feltéve ha $\bar{y}_t$ és $e_t$ függetlenek.

Forrás: C. W. J. Granger, 1980, Long memory relationships and the aggregation of dynamic models, Journal of Econometrics 14 (1980), 227-238.

# 5. fejezet

# Távolságok

## 5.1. Idősorok klaszterezése

### 5.1.1. Bevezetés

Idősorok közti távolság definíálásakor számos probléma merülhet fel. Ezek közül a legjellegzetesebb, hogy nem egyenlő a különböző idősorok megfigyelésszáma. Azaz, hogy általában különböző hosszúságú az a két megfigyeléssor, ami arra a két objektumra vonatkozik, ami közt a távolságot mérni akarjuk. Pontosabban, ha a megfigyelt idősorokat, mint $\mathbb{R}^T$-beli vektorokat tekintjük, akkor olyan vektorok közt kell távolságszerű függvényt meghatározni, amik különböző $(T_1, T_2, ...)$ dimenzióju tereknek az elemei.

Az a probléma, amit a különböző hossz, a különböző hosszúságú vektorok közti távolság meghatározása jelent, többféle technikával oldható meg.

A régebbi módszerek szinte kizárólag a megfigyelések azonos hosszra való csonkolásán alapultak, esetleg a hosszabb megfigyelésekből vett bootstrap-szerű minták

segitségével. Ugyanis az azonos hosszuságú megfigyeléssorokra formálisan már alkalmazhatóak a klasszikus (például az Euklideszi, Minkovszki stb) távolság mértékek. Csakhogy ha ezt a megoldást választjuk, akkor probléma még megmarad az így nyert eredmények értelmezésekor, hogy a feldolgozott megfigyelés vektorok különböző időponthoz tartozó elemei nem függetlenek egymástól. Azaz ezt a technikát egzaktul, ismert tulajdonságok mellett csak például a folyamat becsült innovációira alkalmazhatjuk.

Problémásnak tekinthető a másik egyszerűnek látszó módszer, az idősorok paraméteres összehasonlítása is. Ugyanis tipikusan olyan idősorok összehasonlítására van szükség, amelyeknek a paraméterszáma eltérő. Ez persze azonos, például az ARIMA modellcsaládba tartozó folyamatok esetén megoldható úgy, hogy a hiányzó paramétereket 0-nak vesszük. De az utóbbi időben egyre gyakrabban bukkannak fel olyan feladatok, amikor a legvegyesebb modellcsoportok szerinti folyamatokat kell egymáshoz hasonlítani, osztályozni. Egyszerre fordulnak elő lineáris és kvadratikus, várható érték és szórás stb modellek.

### 5.1.2. Technika

Jelenleg két olyan technika látszik kikristályosodni, ami az idősorok gyakorlati osztályozására általánosan alkalmas. Mindkettő azonos dimenziójú tapasztalati vektorok klasszikus távolság statisztikáira vezeti vissza a különböző hosszúságú idősorok távolságának problémáját. E két technika egyfelől a folyamat spektrumának a vizsgálata, másfelől az előrejelzési sűrűségfüggvény. A dolgozatban csak az utóbbival foglalkozunk.

A folyamat-távolságok definiálására az irodalomban rengeteg példa található. Az alábbiakban csak 3 alaptechnikát mutatunk be.

### 5.1.3. Alaptechnikák

#### Távolságok a folyamatok korrelációs statisztikái alapján

Viszonylag széles azon módszerek köre, amik a folyamatok autokorrelációin, parciális autokorrelációin, esetleg ezek inverz változatán alapulnak. Ennek az eljárásnak lényeges korlátja, hogy a korreláció mint mérték alapvetően a folyamatok lineáris függősének mértéke, és így alkalmatlan például az olyan folyamatok mint az ARCH kiértékelésére. E probléma kezelésének egy lehetősége a kvantilis korrelogram és az arra épülő kvantilis spektrum alkalmazása.

Felírhatjuk egy folyamatok korrelációin alapuló távolságot. Ha $\rho_{X,k}$ illetve $\rho_{Y,k}$ jelöli a két folyamat $k$-ad rendű tapasztalati autokorrelációját, és ha $\rho_X$ illetve $\rho_Y$ az ezen korrelációkból képzett vektor, akkor

$d(X,Y) = \sqrt{(\hat{\rho}_X - \hat{\rho}_Y)'\Omega(\hat{\rho}_X - \hat{\rho}_Y)}$,

ahol az $\Omega$ a tapasztalati autokovarianciák kovarianciájának inverze.

Forrás: J. Caiado, N. Crato D. Pena, 2006, A periodogram-based metric for time series classification Computational Statistics & Data Analysis 50, pp 2668-2684.

#### Távolságok a modellek kanonikus paraméterezése alapján

Ennek a módszernek a legismertebb, legtöbbet vizsgált változata a Piccolo-féle távolság. Ez a távolság feltételezi, hogy rendelkezésre áll az a két $\pi_{X,k}$ és $\pi_{Y,k}$, együtthatósorozat, ami az $X$ és az $Y$ idősort $AR(\infty)$ folyamatként közelíti meg. Ekkor a $d(X,Y) = \sqrt{\sum_{k=1}^{\infty}(\pi_{X,k} - \pi_{Y,k})^2}$ mennyiséget nevezik a két idősor Piccolo

távolságának. Forrás: Piccolo D., 1990, A distance measure for classifying ARIMA models. Journal of Time Series Analysis, 11, 153-164.

### Távolságok a spektrum alapján

Mint ahogyan az előző fejezetben is bemutattuk a standardizált spektrum kezelhető úgy, mint egy sűrűsségfüggvény. Ennek megfelelően, két folyamat távolsága származtatható a megfelelő spektrálssűrűség-függvények Kullback-Leibler távolságaként.

Ha a két spektrálsűrűség-függvény $f_X(\omega)$ illetve $f_Y(\omega)$, akkor

$d_{\mathtt{KL}}(X,Y) = \frac{1}{2N}\sum_{k=1}^{N}\left(\frac{f_X(\omega_k)}{f_Y(\omega_k)} - \log\frac{f_X(\omega_k)}{f_Y(\omega_k)} - 1\right)$.

A megfelelő $J$-divergencia:

$d_{\mathtt{J}}(X,Y) = \frac{1}{2N}\sum_{k=1}^{N}\left(\frac{f_X(\omega_k)}{f_Y(\omega_k)} - \frac{f_Y(\omega_k)}{f_X(\omega_k)} - 2\right)$.

Továbbá megfelelő $\alpha \in [0,1]$ érték mellett a szimmetrikus Chernoff-féle távolság:

$d_{\mathtt{JCh}}(X,Y) =$

$= \frac{1}{2N}\sum_{k=1}^{N}\left(\log\frac{|\alpha f_X(\omega_k)-(1-\alpha)g_Y(\omega_k)|}{|f_Y(\omega_k)|} + \log\frac{|\alpha g_X(\omega_k)-(1-\alpha)f_Y(\omega_k)|}{|g_Y(\omega_k)|}\right)$.

Itt $\omega_k$, $k = 1, ..., n$ a figyelembe vett frekvenciák, általában $\omega_k = k2\pi/T$.

Forrás: G. R. Dargahi-Nourbary, 1992, Discrimination between Gaussian time series based ontheir spectral differences. Commun. Statist. Theory Meth, 21. (9) 2439-2458, TW. Liao, 2005, Clustering of time series data-a survey, Pattern Recognition 38, pp 1857-1874.

## 5.2. Távolságok többváltozós idősorok esetén

### 5.2.1. Bevezetés

Legyen $\{X_t\}$ idősor 0 várható értékű stacionárius idősor, valamint

$\underline{X}_t = (X_1(t), X_2(t), \ldots, X_m(t))'$ $m$-dimenziós vektor $t = 1, 2, \ldots, T$-re.

Legyen $\underline{x} = (X(1)', X(2)', \ldots, X(T)')'$ $mT \times 1$ dimenziós vektor.

Az $\underline{x}$ vektor sűrűségfüggvénye: $p(x)$ és $q(x)$. Ez a két sűrűségfüggvény két különböző $\underline{x}$-re vonatkozó hipotézis alapján van meghatározva. Stacionárius esetben spektrálsűrűség-függvényeket használunk, amiket $f(\lambda)$-val és $g(\lambda)$-val jelölünk. Ezek is mátrix alakúak, és így $\underline{f}(\lambda)$, $\underline{g}(\lambda)$ spektrálsűrűség-mátrixok megfeleltethetőek $\underline{R}_p(s-t)$ és $\underline{R}_q(s-t)$ $m \times m$-es autokovariancia-mátrixoknak.

### 5.2.2. Távolságok

Az egyik legismertebb mutató vagy mérőszám a Kullback-Leibler divergencia, ami a két (többváltozós) sűrűségfüggvény közti eltérést méri.

**5.2.1. Definíció.** *Kullback-Leibler divergencia:* $I(p,q) = E_p(\log(\frac{p(x)}{q(x)}))$.

**5.1. Megjegyzés.** *Az* $E_p()$ *a* $p()$ *szerinti várható értéket jelöli.*

Valamint:

$$I(p,q) = 1/2(tr\{R_p R_q^{-1}\} - \log\frac{|R_p|}{|R_q|} - mT), \tag{5.1}$$

ha $p(x)$ és $q(x)$ nulla várható értékű normális eloszlásúak.

A két $mT \times mT$-és kovariancia-mátrix, $R_p$ és $R_q$, tartalmazza az $m \times m$-es $R_p(s-t)$ és $R_q(s-t)$ mátrixok $s, t = 1, \ldots, T$ blokkjait.

**5.2.2. Definíció.** *J-divergencia a Kullback-Leibler I divergencia alapján:* $J(p,q) = I(p,q) + I(q,p)$, $J(.,.)$ *szimmetrikus.*

Ez már tartalmazza a távolság fogalom összes tulajdonságát,kivéve a háromszög egyenlőtlenség tulajdonságát, ezért "kvázi-távolságnak" hívjuk.

**5.2.3. Definíció.** *Jelöljük d-vel az* $\mathbb{R}$*-be képző függvényt, ha d függvény a következő tulajdonságokat teljesíti:*

1. $d(x,y) \geq 0$ *és* $d(x,y) = 0 \Leftrightarrow$ *ha* $x = y$,
2. *szimmetrikus, vagyis* $d(x,y) = d(y,x)$,
3. *háromszög-tulajdonság:* $d(x,z) \leq d(x,y) + d(y,z)$,

*akkor d függvényt távolságnak hívjuk.*

A klaszteranalízisben gyakran használják $J(.,.)$-t.

**5.2.4. Definíció.** *Chernoff információs mérték:* $B_\alpha(p,q) = -\log E_p(\frac{q(x)}{p(x)})^\alpha$, $p()$ *és* $q()$ *sűrűségfüggvények közti távolságot méri, az indexben szereplő* $\alpha \in (0,1)$.

Ha $\alpha = 0,5$, akkor $B_\alpha(p,q)$ szimmetrikus divergenciával fog megegyezni.

Két normális eloszlású vektorváltozó közti összefüggést, eltérését, divergenciáját a kovariancia írja le. Így, ha a Chernoff-féle információs mértékbe beírjuk a kovarianciamátrixokat, akkor ezt kapjuk:

$$B_\alpha(p,q) = 1/2(\log \frac{|\alpha R_p + (1-\alpha)R_q|}{|R_q|} - \alpha \log \frac{|R_p|}{|R_q|}) \tag{5.2}$$

**5.2. Megjegyzés.** $B_\alpha(p,q) = B_{1-\alpha}(q,p)$

**5.3. Megjegyzés.** *$B_\alpha(p,q)$ $I(p,q)$-hoz konvergál, ha $\alpha \to 0$*
*$B_\alpha(p,q)$ $I(q,p)$-hez tart, ha $\alpha \to 1$*

Tehát 0 és 1 körül $B_\alpha(p,q)$ hasonlóan viselkedik, mint $I(p,q)$.

**5.1. Állítás.** *Az előző megjegyzés alapján a $J(p,q)$ "kvázi-távolságot" átírhatjuk így:*
$JB_\alpha(p,q) = B_\alpha(p,q) + B_\alpha(q,p)$.

A gyakorlatban $I(p,q)$ és $B_\alpha(p,q)$ túl nagy és nehezen kezelhető, ezért $X(t)$ folyamat $f(\lambda)$ és $g(\lambda)$ spektrál-mátrixait használjuk fel.
$\lim_{T\to\infty} \frac{1}{T} I(p,q) = I(f,g) = 1/2 \int\limits_{-\pi}^{\pi} (tr\{fg^{-1}\} - \log \frac{|f|}{|g|} - p) \frac{d\lambda}{2\pi}$
$\lim_{T\to\infty} \frac{1}{T} B_\alpha(p,q) = B_\alpha(f,g) = 1/2 \int\limits_{-\pi}^{\pi} (\log \frac{|\alpha f + (1-\alpha) g|}{|g|} - \alpha \log \frac{|f|}{|g|}) \frac{d\lambda}{2\pi}$

**5.4. Megjegyzés.** *$f(\lambda)$ és $g(\lambda)$ spektrál-mátrixok $p(x)$ és $q(x)$ sűrűségfüggvényeknek felelnek meg.*

A két határérték elég nagy $T$-re, nem veszít el semmit az információ-tartalomból, és a kovariancia-mátrixokat a spektrál-mátrixok reprezentálják. Cél a dimenziócsökkentés volt és a $Tm \times Tm$-es mátrixokból $m \times m$-es lesz.

**5.2. Állítás.** *Legyen $f$ és $g$ távolsága következő:*
$D_H(f,g) = 1/2 \int\limits_{-\pi}^{\pi} H(fg^{-1}) \frac{d\lambda}{2\pi}$,
*ahol $H()$ mátrixértékű függvény.*

$D_H()$ függvényre igaz, hogy nem negatív és akkor és csak akkor nulla, ha az argumentumában levő két szereplő egyenlő egymással. $D_H(.,.)$ kielégíti a "kvázi-távolság" fogalmát (háromszög-tulajdonságot nem teljesíti csak).

**5.3. Állítás.** *$I(.,.)$-re: $H_I(Z) = tr\{Z\} - \log|Z| - p$.*

*$B_\alpha(.,.)$-ra: $H_B(Z) = \log|\alpha Z + (1-\alpha)I_p| - \alpha \log|Z|$.*

*Tetszőleges kvadratikus függvényre: $H_Q(Z) = 1/2tr(z - I_p)^2$.*

**5.5. Megjegyzés.** *$D_H(.,.)$ nem szimmetrikus, de könnyen azzá tehető: $\tilde{H}(Z) = H(Z) + H(Z^{-1})$.*

Így $H(.)$ bevezetésével egyaránt megkapható J-divergencia és Chernoff-féle kvázi távolság is.

Legyen a frekvencia $\lambda_s = \frac{2\pi s}{T}$, $s = 1, 2, \ldots, T$. Ekkor $D_H(f, g)$ a következőképpen közelíthető:

$D_H(f, g) \approx 1/2T^{-1} \sum_s H(f(\lambda_s)g^{-1}(\lambda_s))$.

A gyakorlatban tulajdonképpen a magas frekvenciák nem játszanak szerepet. A magas frekvenciák közelében $g(\lambda)$ spektrál-mátrix a nullához tart. Ezért a képletünkbe a frekvenciák súlyozására szolgáló $\phi(.)$ függvény beírásával elkerüljük azt, hogy $g(\lambda)$ nulla legyen:

$D_{\phi H}(f, g) = 1/2 \int\limits_{-\pi}^{\pi} \phi(\lambda) H(f(\lambda)g^{-1}(\lambda)) \frac{d\lambda}{2\pi}$,

ahol $\phi(.)$ nemnegatív szimmetrikus súly-függvény.

Forrás: Yoshihide Kakizawa, Robert H. Shunway, and Masanobu Taniguchi, Discrimination and Clustering for Multivariate Time Series, Journal of the American Statistical Association, Vol. 93, No. 441 (Mar., 1998), pp. 328-340.

# 6. fejezet

# Gyakorlat és eredmények

## 6.1. Adatok

Az adatokat az Eurostat honlapjáról töltöttem le.

Forrás: http://ec.europa.eu/economy finance/db indicators/surveys/time series, http://ec.europa.eu/economy finance/db indicators/surveys/documents/userguide en.pdf

Az Eurostat hitvallása a következő: cél, hogy vezető szerepet töltsön be magas minőségű statisztikai szolgáltatás terén egész Európában. Székhelye Luxemburgban található.

A letöltött dokumentumban 29 adatsor van, ami az Európai Unió 27 tagországából, az egész Európára vonatkozó és az Euró Zóna idősorából áll össze.

Az adathalmaz az úgynevezett "Economic Sentiment Indicator" , röviden (ESI), magyarul gazdasági bizalmi index. Ez egy havi adatsor, ami a fent említett 27 országra és 2 régióra vonatkozóan tartalmaz adatokat. A rendelkezésre álló havi adatsorok 1985 januárjától 2013 áprilisáig tartanak.

Az ESI egy összetett mutató, ami 5 szektor bizalmi indexéből tevődik össze, és az 5 szektor indexe más-más súllyal szerepel az ESI-ben. Az 5 bizalmi index a következő:

1. ipari bizalmi index,
2. üzleti (szolgáltatás) bizalmi index,
3. fogyasztói bizalmi index,
4. konjunktúraindex,
5. kereskedelmi bizalmi index.

Minden egyes bizalmi indexekhez tartozik egy-egy kérdőív, amikben a feltett kérdések olyanok, hogy az egyes területek állapotát mérik fel. A kérdésekre adott válaszok sorrendbe rendezettek, minősítő jellegűek, viszont nem számszerűek. A lehetséges válaszok száma kérdésenként 3, illetve 5. Az átlagaikat szezonálisan kiigazították, megadható a fő irányvonal.
A felmérést a Joint Harmonised EU Program of business and Consumer Surveys program szabályozza.
Az ESI standardizált változó mégpedig 100 minta átlaga és 10 minta szórása alapján (standardizálás: $x' = \frac{x-\bar{x}}{\sigma_x}$). Komponensenként standarzilálják, 0 átlag, 1 tapasztalati szórás. Súlyozva adják őket össze.
Adatsor neve: NACE Rev. 2.
Adatsor forrása:
http://ec.europa.eu/economy finance/db indicators/surveys/documents/series/nace2 ecfin 1304b/esi nace2.zip.

Az 5 bizalmi index ESI-ben szereplő súlyai:

1. ipar: 40%,

2. szolgáltatás: 30%,
3. fogyasztás: 20%,
4. konjunktúra: 5%,
5. kereskedelem: 5%.

Az adathalmazból felhasznált adatoszlopok:
EU.ESI EA.ESI BE.ESI BG.ESI CZ.ESI DK.ESI DE.ESI EE.ESI IE.ESI EL.ESI ES.ESI FR.ESI IT.ESI CY.ESI LV.ESI LT.ESI LU.ESI HU.ESI MT.ESI NL.ESI AT.ESI PL.ESI PT.ESI RO.ESI SI.ESI SK.ESI FI.ESI SE.ESI UK.ESI
Időtartam: 1985.01.-2013.04.
Összesen 240 megfigyelés adatoszloponként.
Országok és kódjaik:

- BE Belgium
- BG Bulgaria
- CZ Czech Republic
- DK Denmark
- DE Germany
- EE Estonia
- IE Ireland
- EL Greece
- ES Spain
- FR France
- IT Italy

- CY Cyprus
- LV Latvia
- LT Lithuania
- LU Luxembourg
- HU Hungary
- MT Malta
- NL Netherlands
- AT Austria
- PL Poland
- PT Portugal
- RO Romania
- SI Slovenia
- SK Slovakia
- FI Finland
- SE Sweden
- UK United Kingdom

Plusz a két régió:

- EU European Union
- EA Euro Area

## 6.2. Matematikai módszerek

A következő három módszert használtunk fel az adatok elemzésénél:

- Skálázás,
- Hierarchikus klaszterelemzés,
- K-közép-módszer.

**A három módszer rövid leírása**

A skálázás: n objektumra ráillesztünk egy $T$ távolság-mátrixot, és úgy keresünk $\mathbb{R}^2$-ben $n$ darab pontot, hogy a pontpárok távolsága jól közelíti $T$-ben szereplő értékeket.

Hierarchikus klaszterelemzés: Adott $n$ objektum. Az első lépésben minden egyes objektum önálló klaszter. Ez után vonjunk össze a klasztereket. Mindig azt a kettőt, ami a legközelebb van. A klaszterezési eljárást az objektumok közti távolság, és az ezek alapján számolt (aggregált) klasztertávolság határozza meg. A Ward eljárást alkalmaztuk. Ennél a módszernél az aggregált távolság a klaszterközepek közti távolsággal egyenlő.

K-közép-módszer (Voronoi): K közepű osztályozás esetén az osztály közepeket iteratív módon határozzuk meg. Mindegyik lépésben az egyes osztályokba az osztályközepeknek megfelelő Voronoi cellabeli megfigyeléseket soroljuk, és az új osztályközép az adott osztályba sorolt megfigyelések átlaga.

## 6.3. Eredmények

Az idősorok osztályozásánál alapvető probléma, hogy az idősorok eltérő hosszúságúak, ezért olyan mértéket kell találni, ami az idősorok hosszától független [T.Warren Liao-Clustering of time series data survey, 2005].

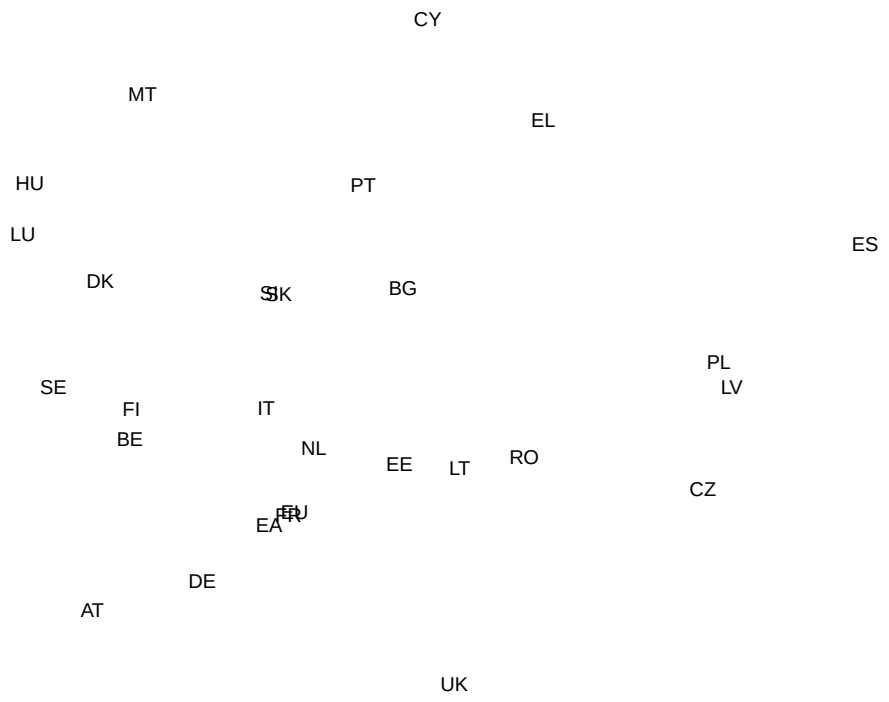
A tipikus osztályozási módszerek azonos hosszúságú megfigyelési sorok alapján működnek, ezért szükséges, hogy az idősoroknak a fix hosszúságú interpretációját vegyük, és azokat osztályozzuk.

Ha a fix hosszúságú leírását keressük az előző részben tárgyalt folyamatoknak, akkor erre a spektrum egy célszerű megoldás. Egyéb megoldás lehet például az előrejelzési sűrűségfüggvény [José A. Vilar, Juan M. Vilar-Time series clustering based on nonparametric multidimensional forecast densities].

Ennek a gondolatmenetnek megfelelően a következő eljárást alkalmaztuk. Vettük a kiválasztott idősoroknak és a lineáris modell szerinti maradékoknak a spektrumát. A spektrumot simítottuk. Ez torzít, de szórást csökkenti, és a megfelelő választás mellett a becslés konzisztens lesz.

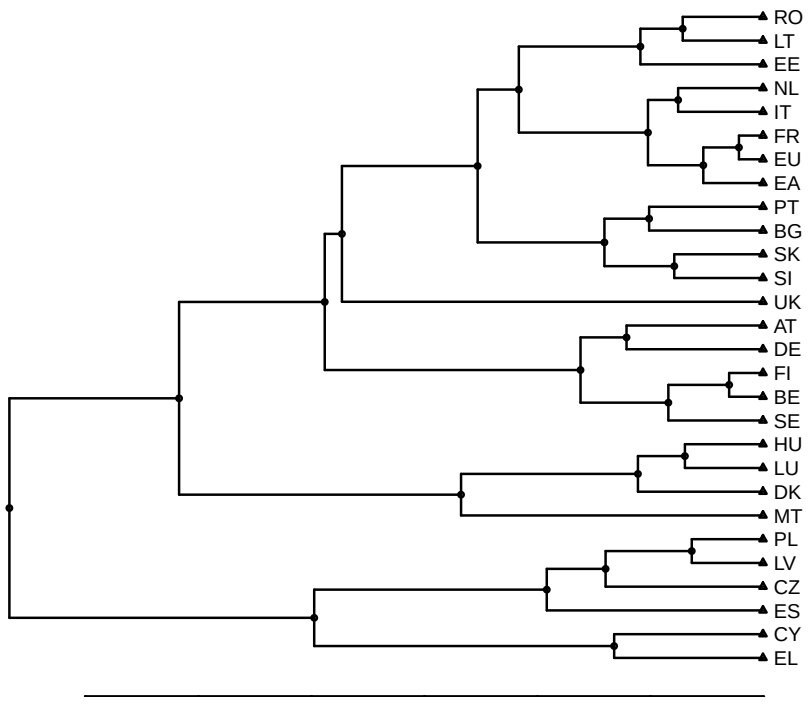
Megbizsgáltuk az 5. fejezetben szereplő távolságokat, és a [10]-es, [19]-es cikkek modelljeire következőket fogom bemutatni a 6.1.-es fejezet adataira:

1. A skálázás során kapott $T$ távolságmátrix a spektrum párokból képzett távolságmátrix, szimmetrikus és a diagonálisban 0-ák vannak. A spektrálsűrűségfüggvény ekvidisztáns felosztásban 100 pontban lett kiszámítva.




6.1. ábra. Skálázott EU-27

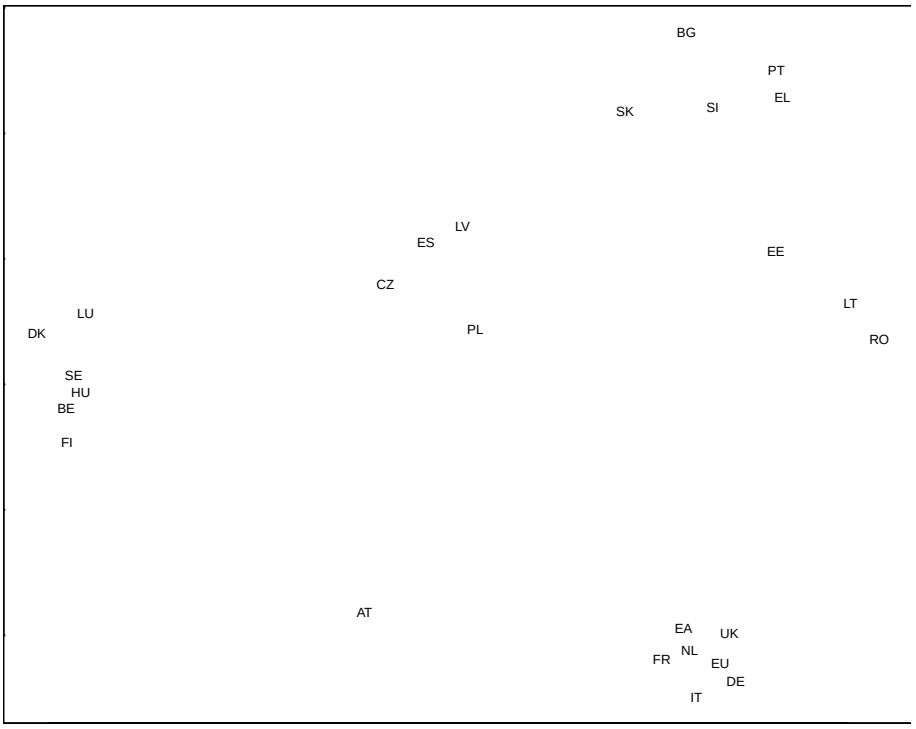
2. A klaszterezésnél két legközelebbi pont kerül egy csoportba, és az új elem csoportba sorolásánál az új távolságot a már meglévő klaszter közepétől vesszük.




6.2. ÁBRA. Európai országok klaszterezve a bizalmi indexek alapján

Ez után megvizsgáltuk, hogy a klaszterezés mennyire stabil.

Hierarchikus klaszterezéssel megállapítottuk azt, hogy lényegében 4 osztály keletkezik. Ezt követően K-közép módszert használtunk. Ez után a "jackknife", illetve a "bootstrap" gondolatmenete alapján eggyel kevesebb elemű részhalmazokra vonatkozóan megismételtük az osztályozást. Az egyes osztályozásokkor vizsgáltuk, hogy melyek azok az idősorok, amelyek a részhalmazban különböző csoportba kerülnek. Az így nyert különbség számot, mint spektrumok (idősorok) közti távolságot értelmeztük, azaz egy spektrumok közti távolságot határoztunk meg, annak alapján, hogy milyen gyakran kerülnek 4 osztályon belül különböző osztályba a spektrum-párok.




6.3. ÁBRA. Mennyire stabil a klaszterezés

A periodogram az autokovarianciából számolt spektrum-becslés, ami a folyamat lineáris szerkezetét képzi le.

Számos tapasztalat alapján a közgazdasági folyamatok körében gyakran előfordul, hogy a modellek nemlineárisak (ARCH, GARCH) [Junbum Lee, Suhasini Subba Rao-The quantile spectral density and comparison based tests for nonlinear time series].

Az előbb hivatkozott cikk többek között azt vizsgálja, hogy mi lehet a haszna annak, ha a nemlinearitás miatt a "hagyományos" spektrum helyett kvantilis spektrumot használjuk.

Két tetszőleges esemény kovarianciája a következő:

**6.1. Megjegyzés.** *$cov(A, B) = P(AB) - P(A)P(B)$, ahol $A$ és $B$ a két esemény.*

A kvantilis spektrum a cikkben szereplő definícióra épül.

Két változó kvantilis kovarianciája egy kétváltozós függvény:

**6.3.1. Definíció.** $C_r(x,y) = cov(I_{X_0 \leq x}, I_{X_r \leq y}) = P(X_0 \leq x, X_r \leq y) - P(X_0 \leq x)P(X_r \leq y)$,
*minden* $x, y \in \mathbb{R}$ *és* $I$ *az indikátorfüggvény.*

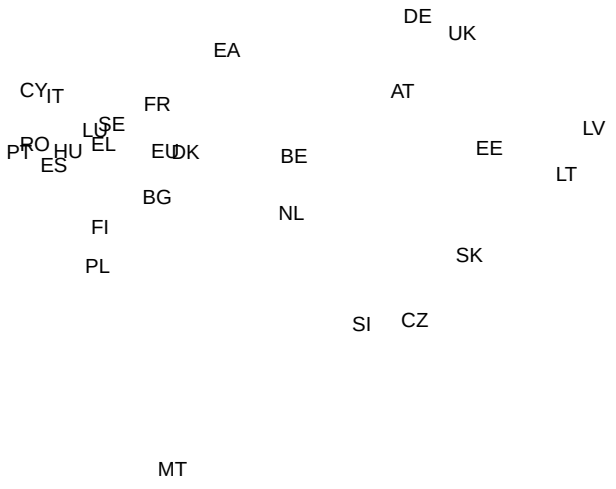
**6.2. Megjegyzés.** *A kvantilis kovariancia fontos tulajdonsága, hogy* $cov(I_{X_0 \leq x}, I_{X_r \leq y}) = cov(I_{X_0 \geq x}, I_{X_r \geq y}) = -cov(I_{X_0 \leq x}, I_{X_r > y})$.

A kvantilis spektrum a kvantilis korrelációra épül.

**6.3.2. Definíció.** $G(x,y,\lambda) = \frac{1}{2\pi} \sum_r C_r(x,y) \cdot e^{ir\lambda}$.
*A kvantilis spektrum egy háromváltozós függvény. A* $G$ *a két kvantilis szint és frekvencia függvénye.*

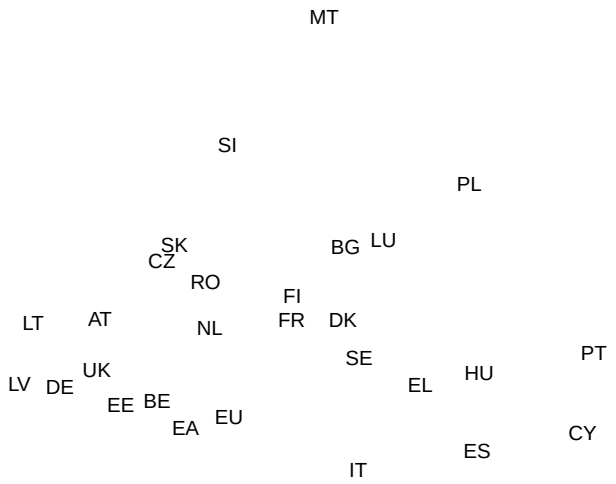
A [Junbum Lee, Suhasini Subba Rao] cikke alapján megpróbáltuk a kvantilis korreláció és a kvantilis spektrum alapján klaszterezni a folyamatainkat. A kérdés, mennyire jól lehet így osztályozni. A kvantilis spektrum közelítésére 10, 25, 75, 90 kvantilist választottuk. Tehát a kvantilis spektrumot 3 pontban vettük: $(x,y) = (10,75)$, $(x,y) = (10,90)$ és $(x,y) = (25,90)$. A kvantilisek autokovarianciák alapján számolt spektrumok távolságát skáláztuk. Ezek alapján az eredmény a következő lett:

1. 10-90: A $cov(I_{X_0<0.1}, I_{X_r>0.9})$ kvantilis kovariogram alapján számolt spektrumok távolságainak skálázott képe.



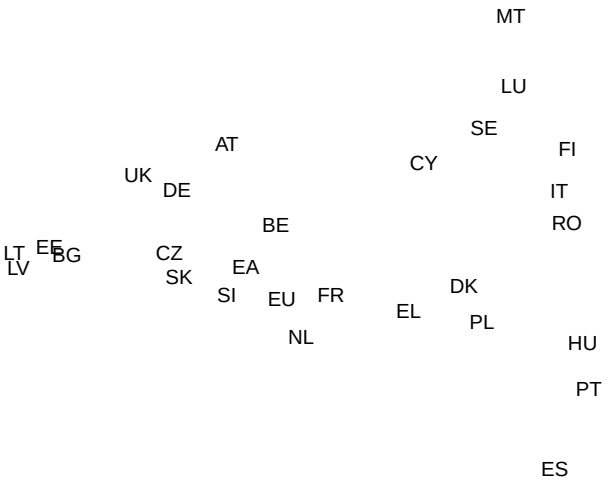

6.4. ábra. 10%-90%-es kvantilisek alapján

2. 10-75: A $cov(I_{X_0<0.1}, I_{X_r>0.75})$ kvantilis kovariogram alapján számolt spektrumok távolságainak skálázott képe.




6.5. ábra. 10%-75%-ös kvantilisek alapján

3. 25-90: A $cov(I_{X_0<0.25}, I_{X_r>0.9})$ kvantilis kovariogram alapján számolt spektrumok távolságainak skálázott képe.




6.6. ábra. 25%-90%-es kvantilisek alapján

Ezeket a csoportosításokat vettük.

az eredmények értékeléséhez az alapadatok és a fenti rajzok összevetése szükséges. Nem kell, hogy a különböző ábrák azonos képet mutassanak, éppen a különbözősség jelent fontos információt. Érdekes megfigyelni, hogy például Magyarország (HU), míg a szélsőséges 10-90-es kvantilisek szerint átlagos, addig az átlagból való kilépést 25-90, illetve az átlagba való belépést mérő 10-75 kvantilis vonatkozásában extremális helyezetű.

### 6.3.1. Összefoglalás

Összefoglalva, a dolgozat témája az idősorok osztályozása, ennek nyomán áttekintettem a zajos idősorok modellezésének kérdését. A spektrumot, a spektrum-becslés témakörében a spektrum általánosításának lehetőségeit, a kvantilis spektrumot. Majd ezeket alkalmaztam az európai ESI 1985.01-től 2013.04-ig tartó havi adatsorára. Eredményül arra jutottam, hogy a felhasznált módszerek alapján stabil

osztályozás lehetséges elsődleges értékelés szerint a szemlélettel azonos eredményt ad.

## 6.4. Klasszikus idősor modellek klaszterezése

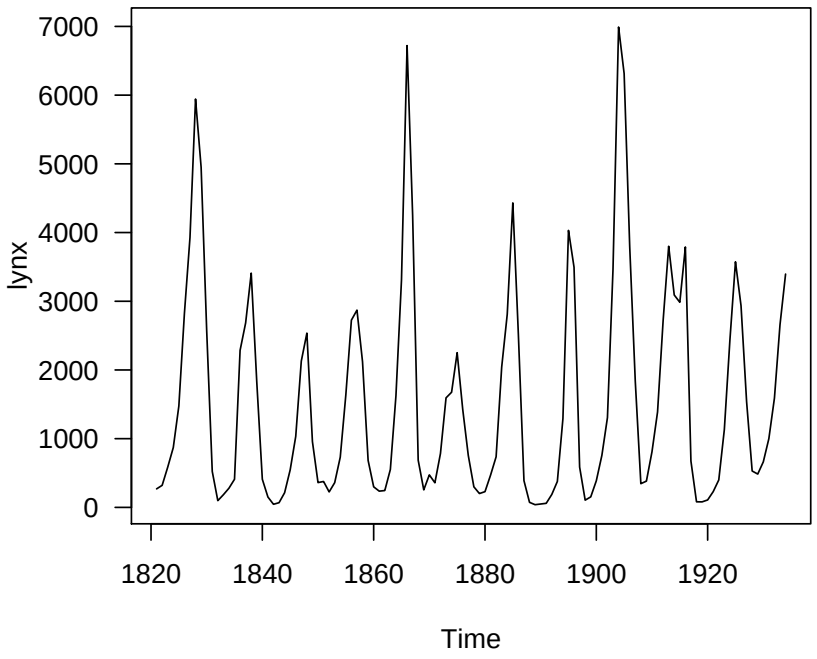
Ebben a részben megnéztük mennyire lehet spektrumok alapjan folyamatokat csoportosítani. Vettünk két irodalmi esetet és egy saját példát.
Ez a két irodalmi eset:

1. Az első esetben a klasszikus "lynx"-re (hiúz) adatsorra illesztett nem feltétlen lineáris folyamatokat vizsgáljuk. az eredeti cikk főkomponens módszerrel hasonlította össze a modelleket, mi viszont spektrum-távolságokon alapuló összehasonlító módszer hatékonyságát vizsgáltuk. Megnézzük a spektrum hatékonyságát.

2. A második esetben a spektrum-távolságokkal foglalkozó Peña cikk közel szinguláris modelljeit vizsgáltuk. A cikkben szereplő modellek lineárisak. Az eredeti cikk ezeken a modelleken különböző távolságokat tesztelt. Mi a saját módszerünk hatékonyságát ellenőriztük. Azt tapasztaljuk, hogy jól működik.

### 6.4.1. Kanadai hiúz adatsor (lynx)

Az idősorelemzésben klasszikus adatsor a "lynx" adatsor, ami 1820-tól 1934-ig éves hiúz-szám, az adatsor hossza 114.




6.7. ábra. Lynx adatsora

A `lynx` adatsorra illesztett modellek képletei a következők voltak (12 darab modell volt):

Gömbölyű zárójelben azoknak szerzőknek a neve szerepel, akik az adott modellt illesztették, a szögletes zárójelben pedig a modellek rövid emlékeztető tipus leírásai szerepelnek.

1. Model A (Moran, 1953) [AR(2)]

$$X_t = \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t$$

ahol $\mu = 2.90$, $\alpha_1 = 1.41$, $\alpha_2 = -0.77$ $\sigma_\varepsilon^2 = 0.04591$

2. Model B (Ozaki, 1982) [AR(2), nemlineáris]

$$X_t = (\mu_{1,1} + (\mu_{1,2} + \alpha_1 X_{t-1})K_t) * X_{t-1} - (\mu_{2,1} + (\mu_{2,2} + \alpha_2 X_{t-1})K_t) * X_{t-2} + \varepsilon_t$$

itt $K_t = exp(-\kappa X_{t-1}^2)$ a $\mu_{1,1} = (1.167$, $\mu_{1,2} = -.316$, $\alpha_1 = .982$ $\mu_{2,1} = .437$, $\mu_{2,2} = .659$, $\alpha_2 = 1.260$, $\sigma_\varepsilon^2 = 0.04327$ és $\kappa = -3.89$ paraméterekkel.

3. Model C (Tong, 1983) [AR(2), szintváltó]

$$X_t = \begin{cases} \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \gamma\varepsilon_t & \text{ha} \quad X_{t-2} \leq \lambda \\ \mu' + \alpha_1' X_{t-1} + \alpha_2' X_{t-2} + \gamma'\varepsilon_t & \text{ha} \quad X_{t-2} \geq \lambda \end{cases}$$

a $\lambda = 3.25$ továbbá a $\mu = .62$, $\alpha_1 = 1.25$, $\alpha_2 = -.43$, $\gamma = .195$, $\mu' = 2.25$, $\alpha_1' = 1.52$, $\alpha_2' = -1.24$, $\gamma' = .250$, $\sigma_\varepsilon^2 = 1$ konstansokkal.

4. Model D (Subba Rao, 1980) [ARMA(11), nemlineáris]

$$X_t - \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_4 X_{t-4} - \alpha_{10} X_{t-10} + \alpha_{11} X_{t-11} =$$

$$= \mu - \beta_8 X_{t-8}\varepsilon_{t-10} + \beta_5 X_{t-5}\varepsilon_{t-8} + \beta_1 X_{t-1}\varepsilon_{t-1} + \varepsilon_t$$

ahol $\alpha_1 = .8845$, $\alpha_2 = .1699$, $\alpha_4 = .1271$, $\alpha_{10} = .5514$, $\alpha_{11} = .5280$ továbbá $\mu = 1.117$, $\beta_8 = .1653$, $\beta_5 = .0970$, $\beta_1 = .0922$ és $\sigma_\varepsilon^2 = .0329$.

5. Model E (Tong, 1977) [AR(11)]

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \alpha_4 X_{t-4} + \alpha_5 X_{t-5}$$

$$+\alpha_6 X_{t-6} + \alpha_7 X_{t-7} + \alpha_8 X_{t-8} + \alpha_9 X_{t-9} + \alpha_{10} X_{t-10} + \alpha_{11} X_{t-11} + \varepsilon_t$$

ahol $\sigma_\varepsilon^2 = 0.0437$ és $\alpha_1 = 1.13$, $\alpha_2 = .51$, $\alpha_3 = .23$, $\alpha_4 = .29$, $\alpha_5 = .14$ $\alpha_6 = .14$, $\alpha_7 = .08$, $\alpha_8 = .04$, $\alpha_9 = .13$, $\alpha_{10} = .19$, $\alpha_{11} = .31$ .

6. Model F (Tong, 1977) [AR(11), szelektált]

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_4 X_{t-4} + \alpha_{10} X_{t-10} + \alpha_{11} X_{t-11} + \varepsilon_t$$

ahol $\alpha_1 = 1.0938$, $\alpha_2 = -.3571$, $\alpha_4 = -.1265$, $\alpha_{10} = .3244$, $\alpha_{11} = -.3622$ ĂŠs $\sigma_\varepsilon^2 = 0.04405$ .

7. Model G (Gabr&Subba Rao, 1981) [ARMA(12,9), nemlineáris]

$$X_t + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \alpha_4 X_{t-4} + \alpha_9 X_{t-9} + \alpha_{12} X_{t-12} =$$

$$= \mu + \beta_3 X_{t-3}\varepsilon_{t-9} + \beta_9 X_{t-9}\varepsilon_{t-9} + \beta_6 X_{t-6}\varepsilon_{t-2} +$$

$$+\beta_1 X_{t-1}\varepsilon_{t-1} + \beta_2 X_{t-2}\varepsilon_{t-7} + \beta_4 X_{t-4}\varepsilon_{t-2} + \varepsilon_t$$

ahol $\alpha_1 = .77227$, $\alpha_2 = .091572$, $\alpha_3 = .083073$, $\alpha_4 = .261493$, $\alpha_9 = -.22558$, $\alpha_{12} = .245841$ továbbá $\mu = 1.486292$, $\beta_3 = -0.7893$, $\beta_9 = .4798$, $\beta_6 = .3902$, $\beta_1 = .1326$, $\beta_2 = .07944$, $\beta_4 = -.3212$ és $\sigma_\varepsilon^2 = .0223$.

8. Model H (Gabr&Subba Rao, 1981) [AR(12)]

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \alpha_4 X_{t-4} + \alpha_5 X_{t-5} + \alpha_6 X_{t-6} +$$

$$+\alpha_7 X_{t-7} + \alpha_8 X_{t-8} + \alpha_9 X_{t-9} + \alpha_{10} X_{t-10} + \alpha_{11} X_{t-11} + \alpha_{12} X_{t-12} + \varepsilon_t$$

ahol $\sigma_\varepsilon^2 = 0.0437$ és $\alpha_1 = -1.0541$, $\alpha_2 = .4538$, $\alpha_3 = -.32597$, $\alpha_4 = .37912$, $\alpha_5 = -.123452$ $\alpha_6 = .17570$, $\alpha_7 = .09598$, $\alpha_8 = .12843$, $\alpha_9 = -.27435$, $\alpha_{10} = -.11427$, $\alpha_{11} = .18534$, $\alpha_{11} = .17218$ .

9. Model I (Gabr&Subba Rao, 1981) [AR(12), szelektált]

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \alpha_4 X_{t-4} + \alpha_9 X_{t-9} + \alpha_{12} X_{t-12} + \varepsilon_t$$

ahol $\sigma_\varepsilon^2 = .0389$ ĂŠs $\alpha_1 = -1.01705$, $\alpha_2 = .39997$, $\alpha_3 = -.25851$, $\alpha_4 = .22037$, $\alpha_9 = -.21099$, $\alpha_{12} = .25343$, $\alpha_{11} = .17218$ .

10. Model J (Tong, 1983) [AR(5), 2-alternáló]

$$X_t = \begin{cases} \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \alpha_4 X_{t-4} + \alpha_5 X_{t-5} + \gamma \varepsilon_t \\ \mu' + \alpha_1' X_{t-1} + \alpha_2' X_{t-2} + \gamma' \varepsilon_t \end{cases}$$

annak függvényében, hogy $X_{t-2} \leq \lambda$ teljesĂźl-e vagy a $X_{t-2} > \lambda$ a $\lambda = 3.05$ továbbá a $\mu = .768$, $\alpha_1 = 1.064$,$\alpha_2 = -.200$, $\alpha_3 = .164$,$\alpha_4 = -.428$, $\alpha_5 = -181$, $\gamma = .174$, $\mu' = 2.254$, $\alpha_1' = 1.474$, $\alpha_2' = -1.202$, $\gamma' = .238$, $\sigma_\varepsilon^2 = 1$ konstansokkal.

11. Model K (Tong, 1983) [AR(7), 2-alternáló]

$$X_t = \begin{cases} \mu + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} + \alpha_4 X_{t-4} + ... + \alpha_7 X_{t-7} + \gamma \varepsilon_t \\ \mu' + \alpha_1' X_{t-1} + \alpha_2' X_{t-2} + \gamma' \varepsilon_t \end{cases}$$

annak függvényében, hogy $X_{t-2} \leq \lambda$ teljesül-e vagy a $X_{t-2} > \lambda$ az $\lambda = 3.116$ továbbá a $\mu = .546$, $\alpha_2 = 1.032$,$\alpha_1 = -.173$, $\alpha_2 = .171$,$\alpha_1 = -.431$, $\alpha_5 = .332$,$\alpha_6 = -.284$,$\alpha_7 = -.210$, $\gamma = .161$, $\mu' = 2.632$, $\alpha_1' = 1.492$, $\alpha_2' = -1.324$, $\gamma' = .225$, $\sigma_\varepsilon^2 = 1$ konstansokkal.

12. Model L (Ozaki, 1982) [Nemlineáris, állapotteres modell, Y folyamat: AR(2) nemlineáris zaj, X folyamat: AR(9)]

$$Y_t = (\mu_{11} + (\mu_{12} + \beta_{11} y_{t-1} + \beta_{12} y_{t-1}^2 + \beta_{13} y_{t-1}^3) * \exp(-g * y_{t-1}^2)) * y_{t-1}] +$$

$$+(\mu_{21} + (\mu_{22} + \beta_{21} y_{t-1} + \beta_{22} y_{t-1}^2 + \beta_{23} y_{t-1}^3) * \exp(-g * y_{t-1}^2)) * y_{t-2})] + e_t$$

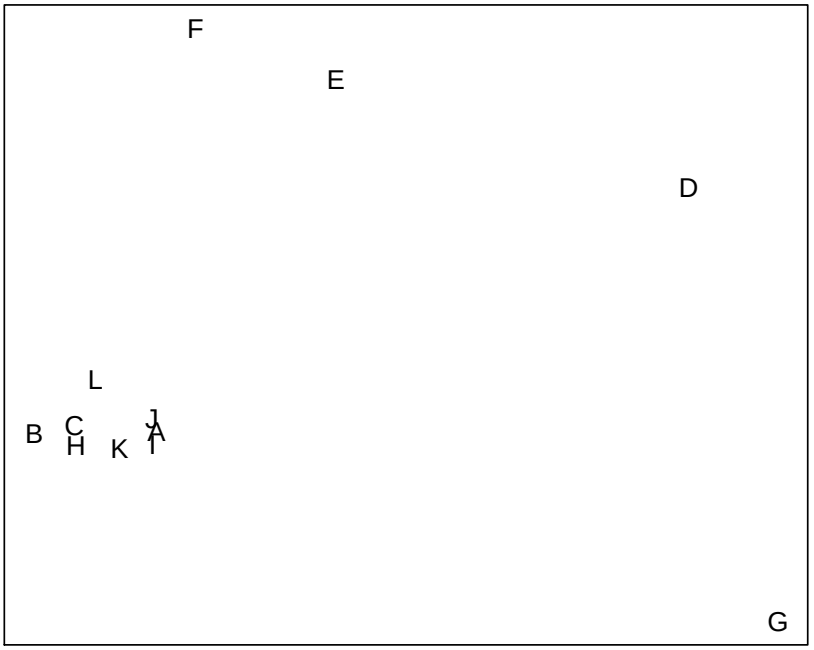
$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + ... + \alpha_8 x_{t-8} + \alpha_9 x_{t-9} + y_t$$

ahol $\alpha_1 = .481$, $\alpha_2 = -.247$, $\alpha_3 = .318$, $\alpha_4 = .230$, $\alpha_5 = .352$, $\alpha_6 = .096$, $\alpha_7 = -.085$, $\alpha_8 = -.289$, $\alpha_9 = -.181$, $\mu_{11} = 1.514$, $\mu_{12} = .480$, $\beta_{11} = -3.332$, $\beta_{12} = - - .610$, $\beta_{13} = 8.906$, $\mu_{21} = -.902$, $\mu_{22} = -.228$, $\beta_{21} = .923$, $\beta_{22} = .193$, $\beta_{23} = -4.216$.

**Összefoglalva a 12 modellt**

Táblázat az illesztett modellekről:

| Kód | Forrás | Modell | Trajektória jellege |
|---|---|---|---|
| A | Moran, 1953 | AR(2) | stacionárius |
| B | Ozaki, 1982 | AR(nemlin 2) | stacionárius |
| C | Tong, 1983 | altAR(2) | stacionárius |
| D | Subba Rao, 1980 | ARMA(11,nemlin 8) | erősen kitörő |
| E | Tong, 1977 | AR(11) | kitörő |
| F | Tong, 1977 | AR(11,szel) | kitörő |
| G | Gabr& Subba Rao, 1981 | ARMA(12,nemlin 9) | igen erősen kitörő |
| H | Gabr& Subba Rao, 1981 | AR(12) | gyengén kitörő |
| I | Gabr& Subba Rao, 1981 | AR(12, szel) | gyengén kitörő |
| J | Tong, 1983 | AR(5)//AR(2) | stacionárius |
| K | Tong, 1983 | AR(7)//AR(2) | stacionárius |
| L | Ozaki, 1982 | SS(all:2,mfi:9) | stacionárius |




6.8. ÁBRA. Lynx adatsorra illesztett 12 modell spektrumának skálázott képe

**Értékelés**

A spektrum távolságok skálázott képe helyesen tükrözi vissza a generált trajektóriák hasonlóságát és különbözőségét.

A 12 spektrumot reprezentáló pontok közül 8 lényegében egy csoportban helyezkedik el. A két, trajektóriáját tekintve "erősen kitörő" jellegű nemlineáris modell [a D és a G] egyenként elkülönülve a kép jobb oldalán látható. A két "gyengén kitörő" trajektóriájú `AR(11)` modell [az E és az F] a képen bal oldalon fent látható egy közös csoportban. Érdekes, hogy a két további enyhén kitörő jellegű `AR(12)` modell [a H és az I] a nagy közös 8-as csoport tagja a kép bal alsó sarkában.

## 6.4.2. Közel szinguláris lineáris modellek

Ebben a részben 12 lineáris modell spektrumát, illetve spektrumaiknak klasztereződését vizsgáltuk.

A 12 modell R-program kódu leírása az alábbi:

```
ModelA<-list(ar=.9)                          # AR(1)
ModelB<-list(ar=c(.95,-.1))                  # AR(2)
ModelC<-list(ar=.95,ma=.1)                   # ARMA(1,1)
ModelD<-list(ar=-.1,ma=-.95)                 # ARMA(1,1)
ModelE<-list(ma=-.9)                         # MA(1)
ModelF<-list(ma=c(-.95,-.1))                 # MA(2)
ModelG<-list(order=c(1,1,0),ar=-.1)          # ARIMA(1,1,0)
ModelH<-list(order=c(0,1,0))                 # ARIMA(0,1,0)
ModelI<-list(order=c(0,1,1),ma=.1)           # ARIMA(0,1,1) +
ModelJ<-list(order=c(0,1,1),ma=-.1)          # ARIMA(0,1,1) -
ModelK<-list(order=c(1,1,1),ar=.1,ma=-.1)    # ARIMA(1,1,1)
ModelL<-list(order=c(1,1,1),ar=.05,ma=-.05) # ARIMA(1,1,1) fél
```

A 12 modell alapján szimulált folyamatoknak kiszámítottuk a tapasztalati spektrumát. Vettük a spektrum-párok távolságait, és ezt skáláztuk. Az így kapott a kép:

6.9. ábra. 12 modell spektrális távolságának skálázott képe

Ezen az ábrán a skálázott kép látható. A skálázott képen szimbolikusan van jelölve, hogy milyen paraméterű ARIMA folyamatról van szó. A jelölésrendszer logikája a következő:
Az ARIMA kódszónak csak az a része szerepel, ami a modellben ténylegesen benne van pozitív paraméterrel. A számok fokszámokat jelentenek. Ha a név nagybetüvel írt, akkor az 1 körüli paraméter értékeket jelent, ha kicsivel akkor az 0 körülit.

Tehát a képen a következőt láthatjuk.
A skálázás alapján három csoport különböztethető meg.

1. Bal felső sarok az AR (lényegében AR) folyamatok (A, B, C),
2. bal alsó sarok a MA (lényegében MA) folyamatok (D, E, F),
3. jobbra az integrált folyamatok (G, H, I, J, K, L) középen az I1 (Wiener), a + jel az egyetlen olyan vizsgált aminek a MA része pozitív, az aláhuzás jel még kisebb MA és AR részt jelent az arIma-val jelölthöz képest.

**Konklúzió**

A vizsgált folyamatok a spektrum alapján jól megkülönböztethetőek: a paraméterezés szerint hasonlóak közelieknek, a különbözők távolinak adódnak. Tehát a módszer használható.

# Nyilatkozat

**Név:** Budai Fruzsina Mária

**ELTE Természettudományi Kar, szak:** alkalmazott matematikus MSc

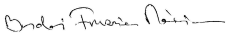
**Szakirány:** sztochasztika

**Neptun azonosító:** GP22YE

**Szakdolgozat címe:** Idősorok osztályozása

Én, Budai Fruzsina Mária, a felelősségem teljes tudatában kijelentem, hogy a dolgozatom önálló munkám eredménye, saját szellemi termékem, abban a hivatkozások és idézetek standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2013. május 31.



# Irodalomjegyzék


[1] C. W. J. Granger, 1980, Long memory relationships and the aggregation of dynamic models, Journal of Econometrics 14 (1980), 227-238

[2] T. Warren Liao, 2005, Clustering of time series data-a survey, Pattern Recognition 38 (2005), 1857-1874

[3] Jorge Caiado, Nuno Crato, Daniel Pena, 2005, A periodogram-based metric for time series classification, Computational Statistics and Data Analysis 50 (2006), 2668-2684

[4] Terence Tai-leung Chong, Gilbert Chiu-sing Lui, Estimating the fractonally integrated process in the presence of measurement errors, 1999, Economics Letters 63 (1999), 285-294

[5] Martin J. Bailey, Prediction of an Autoregressive Variable Subject Both to Disturbances and to Errors of Observation, Journal of the american association, Vol. 60, No. 309 (Mar., 1965), pp. 164-181

[6] Takeshi Amemiya and Roland Y. Wu, The Effect of Aggregation on Prediction in the Autoregressive Model, Journal of the American Statistical association, Vol. 67, No. 339 (Sep., 1972), pp 628-632

[7] C. W. J. Granger and M. J. Morris, , Time Series Modelling and Interpretation, Journal of the Royal Statistical Society, Series a (General), Vol. 139, No. 2 (1976), pp. 246-257

[8] Niels Haldrup, Morten Orregaard Nielsen, 2006, Estimation of fractional integration in the presence of data noise, Computational Statistics and Data Analysis 51 (2007), 3100-3114

[9] Junbun Lee and Suhasini Subba Rao, The quantile spectral density and comparison based tests for nonlinear time series, Department of Statistics, Texas A&M University College Station, U.S.A., arXiv:1112.2759v2 (math.ST) 10 Mar 2012

[10] Jorge Caiado, Nuno Crato, Daniel Pena, 2009, Comparison of Times Series with Unequal Length in the Frequency Domain, Communications in Statistics-simulation and Computation, http://www.tandfonline.com/loi/lssp20

[11] Philip Preu and Thimo Hildebrandt, July 25, 2012, Comparing spectral densities of stationary time series with unequal sample sizes, Fakultat für Mathematik Ruhr Universitat Bochum, Germany

[12] Adam Zagdanski, Rafal Kustra, Exploration of high-dimenzional time series using regularized reduced rank approach: application in time-course microarray data analysis

[13] Laney Kuenzel, June 6, 2010, Gene clustering methods for time series microarray data, Biochemistry 218

[14] Andrew J Barbour and Robert L. Parker, March 12, 2013, Normalization of Power Spectral Density estimates

[15] Zazli Chik Pusat Asasi Sains,2002, A Periodogram-Based Test Method for Comparing Stationary Stochastic Signal, University of Malaya, pp. 208-2012

[16] Emma Sarno, Further results on the asymptotic distribution of the Euclidean distance between MA models, Quaderni di statistica Vol.3, 2001,

[17] Henning Rust, Spectral Analysis of Stochastic Processes, september 2007

[18] Marcella Corduas and Domenico Piccolo, an Application of the AR Metric to Seasonal Adjustment, Dip. di Scienze Economiche e Statistiche, Universita' di Napoli Federico II

[19] H. Tong and P. Dabas, Cluster of time series models: an example, 05 June 2011, Institute of Mathematics, University of Kent

[20] R. H. Shumway and A. N., Linear Discriminant functions for Stationary Time Series, Journal of American Statistical Association, Vol. 69, No. 348 (Dec., 1974), pp. 948-956

[21] Yoshihide Kakizawa, Robert H. Shumway and Masanobu Taniguchi, Discrimination and Clustering for Multivariate Time Series, Journal of the American Statistical Association, Vol. 93, No. 441 (Mar., 1998), pp. 328-340

[22] Umberto Triacca, A note on distance and parallelism between two ARIMA processes, Facolta di Economia Universita di L'Aquila

[23] B. M. Potscher and E. Reschenhgfer, 27 Jun 2007, Distribution of the coates-diggle test statistic in case of replicated observations, Univerity of Technology Vienna

[24] Barry Quinn, Testing for Descrete and Continuous Spectral Diferences, Macquarie University, Statistics Department, Australia

[25] Dr. Márkus László egyetemi jegyzete, kézirat, forrás: http://www.math.elte.hu/probability/markus/AlkmatIdosor1.html, http://www.math.elte.hu/probability/markus/AlkmatIdosor2.html